Knowledge Distillation of Transformer-based Language Models Revisited

Chengqiang Lu \(^1\), Jianwei Zhang \(^1\), Yunfei Chu \(^1\), Zhengyu Chen \(^2\), Jingren Zhou \(^1\), Fei Wu \(^2\), Haiqing Chen \(^1\), and Hongxia Yang \(^1\)

\(^1\)Alibaba Group
\(^2\)Zhejiang University

Abstract
In the past few years, transformer-based pre-trained language models have achieved astounding success in both industry and academia. However, the large model size and high run-time latency are serious impediments to applying them in practice, especially on mobile phones and Internet of Things (IoT) devices. To compress the model, considerable literature has grown up around the theme of knowledge distillation (KD) recently. Nevertheless, how KD works in transformer-based models is still unclear. We tease apart the components of KD and propose a unified KD framework. Through the framework, systematic and extensive experiments that spent over 23,000 GPU hours render a comprehensive analysis from the perspectives of knowledge types, matching strategies, width-depth trade-off, initialization, model size, etc. Our empirical results shed light on the distillation in the pre-train language model and with relative significant improvement over previous state-of-the-arts (SOTA). Finally, we provide a best-practice guideline for the KD in transformer-based models.

1 Introduction
Recently, the emergence of pre-trained language models, especially the transformer-based model such as BERT \cite{Devlin2019BERTPO}, and GPT-3 \cite{Brown2020LanguageMA}, has revolutionized the research on various natural language processing (NLP), compute vision (CV), and multimodal tasks \cite{Dosovitskiy2021AnIC, Liu2021Swin:AN, Lin2021Detectron2:AC, Wang2022MMA} and achieve stunning success. These researches follow the pretrain-then-finetune paradigm: the models are first pre-trained on a large unlabeled corpus and then fine-tuned for specific downstream tasks. Even though these models are effective and prevalent, the heavy model size and high latency limit their application in real-world scenarios, particularly on resource-constrained devices, e.g. mobile phones, IoT devices, and autonomous cars \cite{Zualkernan2022IC, Li2021IC}.

Many model compression techniques have been proposed to obtain a much smaller and eco-friendly model with comparable performance to alleviate the former shortcomings. Among all these methods, knowledge distillation (KD) \cite{Hinton2015DistillingTO} is simple yet effective and has been frequently used \cite{Wang2020Fine-GrainedLM, Jiao2020NeMo:AM}. KD often trains a large and elaborate model as the teacher model to guide the training of a smaller model, named the student model. During the learning procedure, the student model is forced to mimic the behavior of the teacher so that the knowledge from the teacher model will be transferred to the student model.

Despite considerable previous literature having grown up to apply knowledge distillation to transformer-based models for model compression \cite{Wang2020Fine-GrainedLM, Jiao2020NeMo:AM, Sanh2019DistilBERTAD, Sun2020DistilBERTa}, there are still too many unexplored areas in the mechanism of KD. In this work, we attempt to provide a comprehensive overview of KD for transformer-based models. The main contributions of our work are summarized as follows.
We present a generic distillation framework that contains three main components: initialization, knowledge type, and matching strategy. Any existing method could be identified and incorporated into the framework. To tease apart, we categorize common initialization schemes, knowledge types, and matching strategies and propose a unified formulation of distillation.

We conduct systematic and extensive experiments which consist of about 30,000 experimental results and cost over 23,000 GPU hours to investigate the effects of different parts of the proposed framework. We provide exhaustive analyses about the initialization, temperature and hard label weight, layer match, width-depth trade-off, and teacher model size.

Based on the empirical results, we establish a best-practice guideline on the knowledge distillation of transformer-based models. The model following the guideline achieves better scores with a smaller size compared to previous compact models.

2 Preliminary

2.1 Distillation

Knowledge Distillation (KD) is a wide-used technique in deep learning due to its plug-and-play feasibility. It shares many core concepts with transfer learning, Ahn et al. (2019) label smoothing Yuan et al. (2020), ensemble learning, Hinton et al. (2015) and contrastive learning Tian et al. (2020). Although KD could achieve the purpose of model compression, inference acceleration, and generalization improvement Gou et al. (2021), we focus on model compression in this paper. The key idea of KD is to drive a large model (the teacher model $T$) to guide the learning of a small model (the student model $S$). Let $\Omega$ denote the function to extract part of “dark knowledge” from the model $S/T$ and the input $x$. Aim to train the student model $S$ to mimic the behaviors of the teacher model $S$, KD minimizes the following objective function:

$$L_{KD} = \sum_{x \in X} L(\Omega(T, x), \Omega(S, x))$$

where $X$ is the dataset and $L$ is the loss function. The choice of loss function $L$ and the design of knowledge extractor $\Omega$ will significantly influence the effectiveness of knowledge distillation and we discuss them later in the Section 3.2 respectively.

2.2 Transformer

In this paper, our goal is to explore the distillation framework of language models which fit strict memory and computation constraints. Since Transformer-based language models have achieved much progress in a wide range of NLP tasks Vaswani et al. (2017); Devlin et al. (2019), we select the most popular Transformer as the backbone network and review its architecture first. The vanilla Transformer model follows the encoder-decoder architecture based on a multi-head attention mechanism. Therefore, Transformer consists of two types of building blocks: a self-attention module and a feed-forward network.

Self-attention Module The self-attention module utilizes the multi-head attention mechanism to generate outputs with a query and a set of key-value pairs. The output of each head is a weighted sum of values according to the attention distribution. The independent attention heads are concatenated and multiplied by a linear layer to match the desired output dimension:

$$\text{MultiHead}(Q, K, V) = \bigoplus (\text{head}_1, \cdots, \text{head}_H) W^O$$

$$\text{head}_i = \text{Attention}(QW^Q_i, KW^K_i, VW^V_i) = A^i VW^V_i = \text{softmax} \left( \frac{QW^Q_i(KW^K_i)^T}{\sqrt{d_k}} \right) VW^V_i$$

where $\bigoplus$ denotes concatenation operation. $W^Q, W^K, W^V$ and $W^O$ are weight matrices for queries, keys, values, and outputs separately. $A$ is the attention score of $i$-th head. $d_k$ is the dimension of each head and $d_k \times H$ is equal to the hidden dimension $h_b$ in Transformer.

Feed-forward Network The feed-forward network (FFN) is a two-layer network with two linear projection and an activation function (e.g. ReLU):

$$\text{FFN}(x) = \max(0, xW^{f_1} + b_1)W^{f_2} + b_2$$
3 The framework of Distillation

For the transformer-based model, as aforementioned in Section 2.2, it is convenient to regard the teacher-student architectures as homogeneous. Therefore, we choose the BERT as the backbone model without loss of generality in this paper. Given the teacher model, there are two main stages in the progress of distillation: the initialization of the student model and the distillation in the downstream task. We will discuss them in this section.

3.1 Initialization

Since the initialization is crucial (Zhang et al., 2021; Sutskever et al., 2013) in the distillation, a bunch of initialization schemes were proposed to speed up the training progress and improve the final performance (Jiao et al., 2020; Wang et al., 2020; Turc et al., 2019; Sun et al., 2020; Sanh et al., 2019). Generally speaking, there are four kinds of initialization schemes:

- Random initialization: train the student model from scratch.
- Pre-train: pre-train the student model on an unlabeled dataset with a masked LM objective.
- General distillation: pre-train the student model with the aid of the teacher model by introducing the distillation loss to the masked LM objective.
- Pre-load: load part of the weight of the teacher model directly.

Random initialization is the simplest way but usually suffers from the shortage of data in the downstream tasks. Pre-train has been shown to be effective (Devlin et al., 2019; Liu et al., 2019) recently. General distillation, also known as pre-train distillation, utilizes the power of the teacher model when pre-train the student model (Jiao et al., 2020; Wang et al., 2020; Sanh et al.). initialized the student from the selected layers of the teacher. We perform controlled experiments on these schemes to test their effect in Section 4.2.

3.2 Knowledge

In this subsection, we discuss the different categories of knowledge that transfer from the teacher model to the student model. Furthermore, how to calculate the distillation loss for different types of knowledge is also vital and worth investigating in knowledge distillation. Basically, the knowledge could be split into the following three categories: response-based knowledge, feature-based knowledge, and relation-based knowledge.

3.2.1 Response-Based Knowledge

A vanilla knowledge distillation utilizes the output logits of the teacher model as knowledge (Hinton et al., 2015; Ba and Caruana, 2014). The simple but effective method is widely used in model compression. Let \( z_t \) and \( z_s \) denote the logits of the teacher model and student model respectively, the response-based knowledge loss can be formulated as

\[
\mathcal{L}_{res}(z_t, z_s) = D(\varphi(z_t), \varphi(z_s))
\]  

(5)

where \( D \) indicates the computation of the cost function. \( \varphi \) is the transformation function of logits and the simplest transformation function is \( \varphi(z) = z \). However, directly matching logits could be ineffective because the output logits of the cumbersome teacher model could be very noisy. A much more powerful and popular transformation is converting logits to soft targets (Hinton et al., 2015)

\[
\varphi(z_i) = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}
\]  

(6)

where \( T \) is the temperature factor, \( z_i \) is the logit for the \( i \)-th class. The temperature \( T \) controls the "hardness" of soft targets and plays a vital role in knowledge distillation which will be discussed later in Section 4.3. Analogous to label smoothing and regularization (Yuan et al., 2020; Ding et al., 2019; Müller et al., 2019), the utilization of soft targets prevents the student model from overfitting and improves its performance significantly. However, merely using the output of the last layer as auxiliary information limits the competency of KD, especially when the teacher model is very deep or the data amount is small. Consequently, some techniques were proposed to exploit the intermediate-level supervision of the teacher model besides the response-based knowledge.
3.2.2 Feature-Based Knowledge
To provide auxiliary information for mimicking the behavior of the teacher model in intermediate layers rather than simply matching the output logits of the last layer, a considerable amount of literature has been worked on feature-based knowledge distillation (Romero et al., 2015; Zagoruyko and Komodakis, 2017; Kim et al., 2018; Passban et al., 2021). The inspiration of feature-based distillation is simple: directly match the intermediate feature between the teacher model and the student model. It could be formulated as

$$L_{\text{feat}}(f_t(x), f_s(x)) = D(\phi(f_t(x)), \phi(f_s(x))).$$

(7)

Here $D$ is the similarity function to compute the feature loss. $f_t$ and $f_s$ indicate the function used to generate a feature map with input $x$ in the teacher model and the student model respectively. As some similarity functions require the elements to share the same dimension, $\phi$ denotes the mapping function that transforms the features to a proper shape.

In practice of distilling transformer-based models (Jiao et al., 2020; Sun et al., 2020; Wang et al., 2020), the feature map $f(x)$ could be embeddings in the embedding layer, attention matrices $A$, and hidden states $H$. With regard to the similarity function $D$, cross-entropy loss, $L_2$-norm loss, and cosine similarity loss are common choices. Due to the dimension of the teacher model and the student model usually being different, $\phi$ is necessary for feature-based knowledge. The simplest way is to use some dimensionality reduction techniques (e.g. PCA, LDA). However, these methods are not flexible to achieve excellent performance. The most common way to address the problem is to introduce a trainable linear projection layer between the feature map of the teacher model and the student model.

3.2.3 Relation-Based Knowledge
Different from the previous two types of knowledge, which are the output of different layers, relation-based knowledge focus on the relationship of the representations of samples (Tung and Mori, 2019; Park et al., 2019). The core tenet is that the relations of the learned representations contain more and better knowledge than individual ones. The objective of relation-based knowledge loss is expressed as

$$L_{\text{rel}}(f_t(x), f_s(x)) = D(\psi(f_t(x)), \psi(f_s(x)))$$

(8)

where $\psi$ denotes the relational potential function that measures a relationship of given inputs $x$. Here we only consider pair-wise relationship, $\hat{f}_t(x)$, $\hat{f}_s(x)$ are the feature map generator of the teacher model and the student model.

For example, neuron selectivity transfer (Huang and Wang, 2017) computes the similarity matrix of hidden states using Maximum Mean Discrepancy (MMD) in two models then compute the MSE loss between two similarity matrices. In this case, $\hat{f}_t(x) = H_t$, $\hat{f}_s(x) = H_t^T$ indicate the generation of hidden states in the $i$-th and $j$-th layer. $\psi(\cdot)$ here is simply matrix multiplication. Therefore, the objective function could be rewritten as $L_{\text{rel: mmd}}(z_t, z_s) = D_{\text{MSE}}(H_S \cdot H_T^T, H_T \cdot H_T^T)$. Other types relationship-based knowledge of transformer-based model include gram matrices (Yim et al., 2017), value relation (Wang et al., 2020), query and key relation (Wang et al., 2021). $D$ could be mean square error, cross entropy loss, Frobenius norm, and KL divergence.

3.3 Matching Strategy
Section 3.2 addresses the problem of how to distill knowledge. In this section, we explore the problem of how to match the student model $S$ and the teacher model $T$. If the depth of $T$ is equal to the depth of $S$ ($L_T = L_S$), it is easy to solve the problem by matching $T$ and $S$ layer by layer. However, in the most application of distillation, $L_S$ is smaller than $L_T$ in order to compress the student model. Since the representations learned in different layers and different trained models vary a lot (Kornblith et al., 2019; Li et al., 2015), it is vital to select the proper pair of layers to match between $S$ and $T$. Generally, the matching strategy includes three types: 1) First-$k$: select the first $k$ layers to match. 2) Last-$k$: select the last $k$ layers to match. 3) Dilatation: evenly select the matching layers. Figure 1 demonstrates the three strategies when $L_T = 4$, $L_S = 2$, $k = 2$.

3.4 Objective Function
The overall objective function could be formulated as

$$\mathcal{L} = \mathcal{L}_{\text{res}} + \alpha \mathcal{L}_{\text{hard}} + \sum_k \sum_i \beta_{kl} \mathcal{L}_{\text{kl}}(f_t^i(x), f_s^i(x))$$

(9)
where $L_{\text{res}}$ is the response-based knowledge loss (soft label loss). We add the hard label loss $L_{\text{hard}}$ that is used in common supervised learning with the ground-truth label as a previous study (Hinton et al., 2015) found it could significantly improve the performance of the student model. $L_{kl}$ denotes the $k$-th feature-based or relation-based knowledge loss which is applied in the $l$-th pairs of layers between $T$ and $S$. $\alpha$ and $\beta_{kl}$ are all hyper-parameters to balance these loss terms.

4 Empirical Results And Analyses

In this section, we conduct extensive and systematic experiments to investigate the effects of the different parts of knowledge distillation in the transformer-based model. We upload the source code to supplementary material.

4.1 Dataset & Settings

To evaluate different aspects of the distillation of the transformer-based language model, we select the commonly used GLUE benchmark (Wang et al., 2018). Especially, we conduct experiments on Paraphrase Similarity Matching on MPRC (Dolan and Brockett, 2005), QQP, and STS-B (Conneau and Kiela, 2018). For Sentiment Classification, we test on SST-2 (Socher et al., 2013); for Natural Language Inference, we test on QNLI (Rajpurkar et al., 2016) and RTE (Wang et al., 2018); for linguistic Acceptability, we test on CoLA (Warstadt et al., 2019).

We use the BERT$_{\text{base}}$ ($L = 12, d = 768$) as the structure of the teacher model unless otherwise specified. For the optimizer, AdamW (Loshchilov and Hutter, 2017, 2019) is used. For the evaluation metrics in most tasks, we use accuracy for the convenience of comparison. However, for the STS-B task, we select the Pearson correlation coefficient as the metric. For more details about the dataset and related experimental setting and hyperparameters, please refer to Appendix A.1.1.

4.2 Initialization

In this subsection, we test aforementioned four initialization schemes (see Section 3.1). In the setting of pre-train and general distillation, we train the model on the corpus that contains the English Wikipedia and the Toronto Book Corpus (Zhu et al., 2015) following the suggestion of original BERT. We select three structures of the student models: BERT$_{\text{tiny}}$ ($L = 2, d = 128$), BERT$_{\text{mini}}$ ($L = 4, d = 256$), and BERT$_{\text{small}}$ ($L = 4, d = 512$). As the pre-load scheme requires the same dimension between $T$ and $S$, we train a student model with $d = 768$ in this setting.

Table 1 shows the results of different initialization schemes. The figures indicate that random initialization is the worst choice among all four methods. Besides, the pre-load technique shows little advantage in practice. The score of pre-load in the QQP and SST-2 task is relatively high because the width (768) here is much bigger than in others (128), which makes an unfair comparison. Generally speaking, the general distillation and pre-train are better initialization methods because the unsupervised representation of the student model is significant. As a rough guideline, for a comparatively small model size of $S$, just pre-train the student model is the best way to initialize it. If the model size of $S$ increases, it is better to consider general distillation because the student model is able to take more advantage of complementary information provided by the teacher model (Turc et al., 2019).
Table 1: Experimental Results of Different Initialization Schemes

| Initialization | Random | Pre-load | General Distillation | Pre-train | Random | Pre-load | General Distillation | Pre-train | Random | Pre-load | General Distillation | Pre-train |
|---------------|--------|----------|----------------------|----------|--------|----------|----------------------|----------|--------|----------|----------------------|----------|
| QNLI          | 0.6158 | 0.6711   | 0.6266               | 0.7943   | 0.6074 | 0.6711   | 0.8411               | 0.8428   | 0.6149 | 0.7439   | 0.8561               | 0.8673   |
| MRPC          | 0.6838 | 0.7233   | 0.7034               | 0.7647   | 0.7010 | 0.7132   | 0.7843               | 0.7917   | 0.7313 | 0.7206   | 0.8015               | 0.7941   |
| RTE           | 0.5307 | 0.5487   | 0.5487               | 0.6209   | 0.5487 | 0.5451   | 0.5776               | 0.6751   | 0.5596 | 0.5343   | 0.5704               | 0.657    |
| STSB          | 0.0229 | 0.4081   | 0.0907               | 0.6289   | 0.0639 | 0.2448   | 0.7503               | 0.8523   | 0.158  | 0.217    | 0.8256               | 0.8654   |
| QQP           | 0.7853 | 0.8826   | 0.8484               | 0.8563   | 0.8342 | 0.8649   | 0.8884               | 0.8914   | 0.8378 | 0.8813   | 0.9995               | 0.901    |
| MNLI          | 0.5704 | 0.7270   | 0.6216               | 0.7016   | 0.6208 | 0.7479   | 0.7574               | 0.7664   | 0.6287 | 0.7613   | 0.7908               | 0.7891   |
| SST2          | 0.6030 | 0.6329   | 0.6329               | 0.7046   | 0.6120 | 0.7386   | 0.7607               | 0.7695   | 0.6303 | 0.7695   | 0.7965               | 0.7893   |
| CoLA          | 0.6913 | 0.6913   | 0.6922               | 0.6913   | 0.6913 | 0.6913   | 0.7450               | 0.6913   | 0.6989 | 0.7833   | 0.767                |          |

4.3 Temperature and Hard Label

The temperature in the distillation plays an important role in controlling the communication between $T$ and $S$. Higher temperature softens the distribution generated by the teacher model and works in a way that is similar to the label smoothing (Yuan et al., 2020). Hinton et al. found that a weighted average of soft logits loss and hard label loss helps the knowledge transfer from the cumbersome teacher model to the student model. Therefore, the weight of the hard label is also crucial.

To test the effect of two main hyper-parameters and tune them for experiments afterwards, we search from a grid of parameter values (temperature : {1, 2, 4, 8}, hard label weight : {0.1, 0.2, 0.5, 1.0, 2.0, 5.0}). Here, we use BERT$_{min}$ as the student model. Table 1 in Appendix A.3.1 illustrates some interesting facts about these two hyperparameters. First, when the data amount of the downstream task is small, the model distilled with a higher temperature (above 2) achieves better performance. On bigger datasets, lower temperatures result in higher scores. Secondly, although recent studies claim that the hard label is not necessary as the soft logits are sufficiently informative (Shen et al., 2019; Shen and Savvides, 2020), we found a slight hard label weight (e.g. 0.1 or 0.2) is always helpful. In the following experiments, we will use the best hyperparameter setting in Table 1 as the default setting.

4.4 Layer Match

As mentioned in Section 3.2, apart from the response-based knowledge (e.g. soft target) in original knowledge distillation, feature-based knowledge, and relation-based knowledge could provide more nuanced information to help the distillation of knowledge. In this subsection, we select several types of knowledge that are widely used. The core idea of KD is to let the student model learn the behavior of the teacher model. The soft target enables the imitation of the result and other knowledge strives to mimic the intermediate layers. Therefore, we name this part of the experiments as layer match experiments.

We select ten kinds of knowledge that widely used in previous studies (Wang et al., 2020, 2021; Sanh et al., 2019; Jiao et al., 2020; Sun et al., 2019; Huang and Wang, 2017; Yim et al., 2017), including five types of feature-based knowledge: attention mse, attention ce, hidden mse, cos, pkd; and five types of relation-based knowledge: mmd, gram, query relation, key relation, and value relation. See the Appendix A.2 for their definitions and formulas. Three student models are used in this group of experiments: BERT$_{tiny}$, BERT$_{min}$, and BERT$_{small}$. Not only the knowledge types, but we also conduct extensive experiments to test the effect of the three matching strategies mentioned in Section 3.3.

Knowledge Type For the knowledge types, we consider the situation of using only one layer match (single-match) firstly. Table 1 in Appendix A.3.2 shows the result of distilling different knowledge. Compared with solely using soft targets, almost adding any feature-based knowledge or relation-based knowledge improves the performance. When the size of $T$ is smaller or the amount of data in the task is smaller, the model aided by relation-based knowledge tends to achieve a better score than feature-based ones. One reason is the inequality of the dimension of $T$ and $S$ necessitate a learnable projection matrix. However, for some tasks with data shortage, the labeled data is insufficient to train these matrices. Another reason is to preserve the relationship in the representation space of $S$ is easier than mimicking the representation space of $T$ directly. Besides, among the feature-based knowledge, the knowledge about the attention score is more tractable than hidden states as the attention itself.
could be regarded as a self-relation knowledge. In previous experiments, we set the hyperparameters of loss weight $\beta_d$ to be 1. Nevertheless, the magnitudes of different types of knowledge vary a lot. Therefore, we designed an experiment to see if the loss of weight affects the final results. We tuned the weight so that the loss term value of a single-match reaches about 1/10 of the soft label loss. Table 3 in Appendix A.3.3 illustrates that even the roughly selected loss weight improves the performance of over 80% of the student models in different tasks.

To study the effect of the combination of different knowledge types, the second group of experiments tests the models that are distilled with two types of knowledge. We divide all the knowledge types into three categories by the region they take effect: attention (attention mse, attention ce), hidden state (hidden mse, mmd, gram, cos, pkd), query/key/value (query relation, key relation, value relation). Then we test the binary combinations of these 3 tuples. All the 31 double-match settings are applied in three kinds of student models and trained on 8 downstream tasks. The result in Appendix A.3.3 shows that not all double-match settings are better than single-match due to the conflict between different knowledge. However, some double-match could improve the performance significantly, especially the combination of attention ce and relation-based knowledge. It reveals a compound effect as they both respond to the self-attention module.

Matching Strategy  

In the absence of theoretical underpinnings, the choice of matching strategy is really tricky. We conduct extensive controlled experiments to explore this area. Based on three matching strategies mentioned in Section 3.3, we design five settings: (1) match the first $L_S$ layers (First), (2) match the first one layer (First-1), (3) match the last $L_S$ layers (Last), (4) match the last one layer (Last-1), and (5) match the layers evenly (Dilatation).

In the single-match setting, the average variance of different matching strategies in different tasks and models is about only 0.00045. However, it does not reveal that the matching strategy is not important. In fact, among all experimental conditions in the single-match setting (3 model size × 9 downstream tasks), the best configuration in 25 out of 27 is Last-1 or First-1. Similarly, the ratio in the double-match setting is 22 out of 27 (see Appendix A.3.3). It is not a coincidence. Some previous studies point out that, from lower layers to higher layers, the function of each layer varies and function of the first or the last layer are comparably similar. Accordingly, the discrepancy of these layers between the teacher model and the student model is slighter. Therefore, a superior way to select a matching strategy is to use Last-1 or First-1 as the initial trial in the application.

4.5 Deeper or Wider

In the application of small pre-train language models, the limited computing power of mobile devices necessitates the compression of the student model. Given a typical BERT model, the space complexity is $O(L(hdn + h\eta^2))$ and the time complexity is $O(Lhdn^2)$. $L$ is the number of transformer layers and $d$ is the embedding dimension. $\eta$ denotes the length of the input sequence and $h$ is the number of heads in the multi-head attention layer. As the sequence length is usually determined by the input of the downstream task, the depth $L$ and the width $d$ are the main hyper-parameters to reduce the model size and speed up inference time. Along this line, one crucial problem is the trade-off between the depth and the width. The width not only influences the number of parameters in transformer layers but also affects the embedding layer. The space complexity of the embedding layer is $O(|V|d)$ where $|V|$ is the fixed vocabulary size (set to be 30,522 in BERT). Therefore, the smaller a model is, the larger the proportion of the embedding layers to the total model. For instance, the embedding layer in BERT$_{tiny}$ makes up 71% of all parameters and embedding layer parameters account for a over 90% proportion in BERT$_{tiny}$ ($L = 2, d = 128$).

Levine et al. proved that for models with $L > L_{th}(d) \sim \log(d)$, the ability to model input dependencies increases similarity with depth and width. For small models, the network with the depth of $\log(d)$ is too shallower for good performance. Therefore the theoretical findings are not helpful in this situation. We design a bunch of experiments to probe into the matter. We construct several student models with 1) fixed model size of about 6 million parameters, 2) fixed flops (floating-point of
operations) of 2G, and uncover how student models perform vary with width and depth. These models were firstly general distilled with the aid of the same teacher model and then distilled in downstream tasks of GLUE. In the setting of fixed model size, the experimental results in Table 2 illustrates that depth-efficiency takes place in transformer-based models. Under the same hyperparameters except for the width and depth, the deeper models in different tasks usually outperform the other models. In the tasks with small datasets (MRPC, CoLA, STS-B, and RTE), relatively shallower (than the deepest) models achieve the best score. Besides, the results are similar to the conclusion of Kaplan et al.. However, the conclusion is contrary in the setting of fixed flops. The results in the bottom half of Table 2 reveals depth inefficiency. Another perspective is the time-space trade-off. In the first experiment, fixing the model size, the models take more time (bigger flops and higher latency) to perform better; in the second experiment, with similar time consumption, bigger models achieve better scores.

| Teacher Model | Student Model |
|---------------|---------------|
| bert-base     | bert-mini     |
| bert-larg     | bert-larg     |

In previous experiments, we fix the teacher model to study the behavior of the student model. Another crucial part to be explored is the teacher model. In this experiment, we mainly answer the research question: does the larger teacher model teach better? Two teacher models are tested here: BERT\textsubscript{base} (L = 12, d = 1024) and BERT\textsubscript{large} (L = 12, d = 1024). The left side of the Table 3 is the performance of these two teacher models, in all tasks the larger teacher gets better scores (better scores are bolded). However, when teaching students models, the conclusion of "the larger the better" does not hold true. Table 3 indicates that the larger \( T \) teaches better students when the model size of \( T \) is relatively larger (BERT\textsubscript{small}). Conversely, when the capacity of \( S \) is lower, the smaller teacher teaches better because of the capacity gap (Mirzadeh et al., 2020).

## 4.7 Best Practices of Distilling Extremely Small Models for On-device Application

### Constraints

Recently, high-end mobile phones have strong computing power. For instance, the A15 Bionic chip in iPhone 13 performs up to 1500 GFLOPS (Giga Floating Point Of Per Second) and the GPU FP32 floating point in Qualcomm 8 Gen 1 is 1800 GFLOPS. However, most devices in the world including low-to-mid-end mobile phones and IoT devices are not so fast. Therefore, considering the required runtime latency in the common device, we follow the constraints in previous studies (Ge and Wei, 2022) and use the 2G flops (floating-point of operations) as the restrictions.
Besides, we limit the model size up to 14 million parameters including the embedding layer following previous work (Wu et al., 2020). Therefore, the $\text{BERT}_{\text{mini}}$ model that contains 11 million parameters is a proper structure for the on-device application.

Based on the empirical results above, we provide several rules of thumb. The first step is to tune the three hyperparameters: learning rate, temperature, and hard label weight (See Section 4.3 for the guideline for tuning temperature and hard label weight). The second step is to choose the initialization method. We recommend the pre-train method for $\text{BERT}_{\text{mini}}$ and the general-distillation method for larger student models. Then, for the matching strategy, we suggest the First-1 or Last-1 as mentioned in Section 4.4. With regard to the knowledge types, relation-based knowledge is preferred and for smaller models (e.g. $\text{BERT}_{\text{tiny}}$) combining attention-related knowledge could further improve the performance. Besides, several tricks are also exceedingly useful including data augmentation and label smoothing (Jiao et al., 2020; Yuan et al., 2020). Finally, the student model after distillation achieves the comparative score while reducing about 20% model size of the previous SOTA (see Table 7 in Appendix A.3.4).

5 Conclusion

In this paper, we propose a generic framework to distill the transformer-based models, which includes the initialization schemes, knowledge types, and matching strategies. We conduct extensive experiments to investigate the effect of different components in knowledge distillation. Moreover, we provide a best-practice guideline to distill the $\text{BERT}_{\text{mini}}$ for on-device applications.

References

S. Ahn, S. X. Hu, A. C. Damianou, N. D. Lawrence, and Z. Dai. Variational information distillation for knowledge transfer. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 9155–9163, 2019.

J. Ba and R. Caruana. Do deep nets really need to be deep? In NIPS, 2014.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. J. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. ArXiv, abs/2005.14165, 2020.

A. Conneau and D. Kiela. Senteval: An evaluation toolkit for universal sentence representations. ArXiv, abs/1803.05449, 2018.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. ArXiv, abs/1810.04805, 2019.

Q. Ding, S. Wu, H. Sun, J. Guo, and S. Xia. Adaptive regularization of labels. ArXiv, abs/1908.05474, 2019.

W. B. Dolan and C. Brockett. Automatically constructing a corpus of sentential paraphrases. In IJCNLP, 2005.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv, abs/2010.11929, 2021.

T. Ge and F. Wei. Edgeformer: A parameter-efficient transformer for on-device seq2seq generation. ArXiv, abs/2202.07959, 2022.

J. Gou, B. Yu, S. J. Maybank, and D. Tao. Knowledge distillation: A survey. ArXiv, abs/2006.05525, 2021.

G. E. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. ArXiv, abs/1503.02531, 2015.

Z. Huang and N. Wang. Like what you like: Knowledge distill via neuron selectivity transfer. ArXiv, abs/1707.01219, 2017.

G. Jawahar, B. Sagot, and D. Seddah. What does bert learn about the structure of language? In ACL 2019-57th Annual Meeting of the Association for Computational Linguistics, 2019.
A. Simoulin and B. Crabbé. How many layers and why? an analysis of the model depth in transformers. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 221–228, 2021.

R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In EMNLP, 2013.

S. Sun, Y. Cheng, Z. Gan, and J. Liu. Patient knowledge distillation for bert model compression. In EMNLP, 2019.

Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou. Mobilebert: a compact task-agnostic bert for resource-limited devices. ArXiv, abs/2004.02984, 2020.

I. Sutskever, J. Martens, G. E. Dahl, and G. E. Hinton. On the importance of initialization and momentum in deep learning. In ICML, 2013.

Y. Tian, D. Krishnan, and P. Isola. Contrastive representation distillation. ArXiv, abs/1910.10699, 2020.

F. Tung and G. Mori. Similarity-preserving knowledge distillation. 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pages 1365–1374, 2019.

I. Turc, M.-W. Chang, K. Lee, and K. Toutanova. Well-read students learn better: The impact of student initialization on knowledge distillation. ArXiv, abs/1908.08962, 2019.

A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In NIPS, 2017.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. ArXiv, abs/1804.07461, 2018.

P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. arXiv preprint arXiv:2202.03052, 2022.

W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. ArXiv, abs/2002.10957, 2020.

W. Wang, H. Bao, S. Huang, L. Dong, and F. Wei. Minilmv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In FINDINGS, 2021.

A. Warstadt, A. Singh, and S. R. Bowman. Neural network acceptability judgments. Transactions of the Association for Computational Linguistics, 7:625–641, 2019.

Z. Wu, Z. Liu, J. Lin, Y. Lin, and S. Han. Lite transformer with long-short range attention. ArXiv, abs/2004.11886, 2020.

Z. Yang, Y. Cui, Z. Chen, W. Che, T. Liu, S. Wang, and G. Hu. TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 9–16. Association for Computational Linguistics, 2020. URL https://www.aclweb.org/anthology/2020.acl-demos.2.

J. Yim, D. Joo, J.-H. Bae, and J. Kim. A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 7130–7138, 2017.

L. Yuan, F. E. H. Tay, G. Li, T. Wang, and J. Feng. Revisiting knowledge distillation via label smoothing regularization. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 3902–3910, 2020.

S. Zagoruyko and N. Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. ArXiv, abs/1612.03928, 2017.

T. Zhang, F. Wu, A. Katiyar, K. Q. Weinberger, and Y. Artzi. Revisiting few-sample bert fine-tuning. ArXiv, abs/2006.05987, 2021.

Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. 2015 IEEE International Conference on Computer Vision (ICCV), pages 19–27, 2015.

I. A. Zualkernan, S. Dhou, J. Judas, A. R. Sajun, B. R. Gomez, and L. A. Hussain. An iot system using deep learning to classify camera trap images on the edge. Comput., 11:13, 2022.
Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [Yes]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]

2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you’re using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowd sourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]
A Appendix

A.1 Reproducibility

A.1.1 Settings

In most experiments, we use the following default settings unless otherwise specified:

- The hyperparameters of loss weight $\beta_{kl}$ is set to be 1.
- The temperature and hard label weight are tuned by grid search and select the best one in other experiments (See Section 4.3).
- The initialization scheme is pre-train (See Section 4.2).

In the experiments about temperature and hard label weight in Section 4.3, no feature-based or relation-based knowledge distillation is used.

A.1.2 Code

We provide source code of this paper in the supplementary material. The main file is `main.py`. We modified the implementation of BERT in huggingface in `custom_bert.py` for the convenience of distillation. In the environment of distributed multi-GPU, we use the DistributedDataParallel (DDP) provided by PyTorch and the main file is `distributed_wrapper.py`. For part of the implementations of knowledge distillation, we use the TextBrewer ([Yang et al., 2020] under Apache 2.0 license).

A.1.3 Teacher Models

We download the fine-tuned BERT$_{\text{base}}$ teacher models of different GLUE task in huggingface:

- MNLI: https://huggingface.co/ishan/bert-base-uncased-mnli
- QQP: https://huggingface.co/textattack/bert-base-uncased-QQP
- QNLI: https://huggingface.co/textattack/bert-base-uncased-QNLI
- SST-2: https://huggingface.co/textattack/bert-base-uncased-SST-2
- CoLA: https://huggingface.co/textattack/bert-base-uncased-CoLA
- STS-B: https://huggingface.co/textattack/bert-base-uncased-STS-B
- MRPC: https://huggingface.co/textattack/bert-base-uncased-MRPC
- RTE: https://huggingface.co/textattack/bert-base-uncased-RTE

The BERT$_{\text{large}}$ models are downloaded from https://huggingface.co/yoshitomo-matsubara.

A.2 Knowledge Types

In this subsection, we introduce the definitions of the knowledge used in Section 4.4. $T$ and $S$ denote the teacher model and the student model here. $l$ and $r$ indicate the layer number of $T$ and $S$ respectively. $N_h$ is the number of attention heads. $A$ is the attention matrix and $H$ is the hidden state. $W$ indicates the learnable projection matrix.

- Attention mse: the mse loss of the sum of attention heads between $T$ and $S$

$$
\mathcal{L} = D_{\text{mse}}(\sum_{a=1}^{N^g_h} A^S_{a,l}, \sum_{a=1}^{N^g_h} A^T_{a,r})
$$

- Attention ce: the cross-entropy loss of the mean of attention heads between $T$ and $S$

$$
\mathcal{L} = D_{\text{ce}}\left(\frac{1}{N^g_h} \sum_{a=1}^{N^g_h} A^S_{a,l}, 1 \frac{1}{N^g_h} \sum_{a=1}^{N^g_h} A^T_{a,r}\right)
$$
• Hidden mse: the mse loss of the hidden states between $T$ and $S$

\[ \mathcal{L} = D_{\text{mse}}(H^S_T, H^T_r W_r) \]  
(12)

• Cos: the cosine similarity loss between the hidden states between $T$ and $S$

\[ \mathcal{L} = D_{\text{cos}}(H^S_T, H^T_r W_r) = 1 - \cos(H^S_T, H^T_r W_r) \]  
(13)

• Pkd: the normalized mse loss of the hidden states between $T$ and $S$

\[ \mathcal{L} = D_{\text{mse}} \left( \frac{H^S_T}{|H^S_T|}, \frac{H^T_r W_r}{|H^T_r W_r|} \right) \]  
(14)

• Mmd: the mse loss between the similarity matrices of hidden states. $H_1$ and $H_2$ are two hidden states in models. $\top$ indicates the matrix transpose.

\[ \mathcal{L} = D_{\text{mse}}(H^T_1 \cdot H^T_2 \top, H^S_1 \cdot H^S_2 \top) \]  
(15)

• Gram: the mse loss between the similarity matrices of hidden states. The difference between mmd and gram is the order of matrix multiplication.

\[ \mathcal{L} = D_{\text{mse}}(H^{1T} \cdot H^{2T} \top, H^{S1} \cdot H^{S2} \top) \]  
(16)

• Query relation: the KL-divergence loss of the query relation between $T$ and $S$

\[ \mathbf{VR}^T_{l,a} = \text{softmax} \left( \frac{V^T_{l,a} \cdot V^{TT}_{l,a}}{\sqrt{d}} \right) \]

\[ \mathbf{VR}^S_{r,a} = \text{softmax} \left( \frac{V^S_{r,a} \cdot V^{ST}_{r,a}}{\sqrt{d}} \right) \]

\[ \mathcal{L} = \frac{1}{N_h} \sum_{a=1}^{N_h} D_{\text{KL}}(\mathbf{VR}^T_{l,a}, \mathbf{VR}^S_{r,a}) \]

• Key relation: the KL-divergence loss of the key relation between $T$ and $S$. The definition is similar to the query relation above, just replace $Q$ with $K$.

• Value relation: the KL-divergence loss of the value relation between $T$ and $S$. The definition is similar to the query relation above, just replace $Q$ with $V$.

### A.3 Detailed Experimental Results

#### A.3.1 Temperature & Hard Label Weight

Table 4: Hyper-parameters experiments about temperature and hard label weight

| Temperature | 1      | 2      | 4      | 8      |
|-------------|--------|--------|--------|--------|
| Hard Label Weight |       |        |        |        |
| MRPC        |        |        |        |        |
| 0.1         | 0.701  | 0.7181 | **0.7206** | 0.701  |
| 0.2         | 0.701  | 0.7034 | 0.7083 | 0.7059 |
| 0.5         | 0.7034 | 0.7034 | 0.701  | 0.7059 |
| 1           | 0.7034 | 0.701  | 0.7059 | 0.7132 |
| 2           | 0.7034 | 0.7059 | 0.7059 | 0.7083 |
| 5           | 0.701  | 0.7034 | 0.7083 | 0.7059 |
| SST-2       |        |        |        |        |
| 0.1         | **0.8716** | 0.8693 | 0.8681 | 0.8647 |
| 0.2         | 0.8658 | 0.8670 | 0.8658 | 0.8624 |
| 0.5         | 0.8647 | 0.8658 | 0.8681 | 0.8647 |
| Hard Label Weight | QQP   | QNLI   | RTE   | CoLA   | STS-B   | MNLI-mm | MNLI-m   |
|-------------------|-------|--------|-------|--------|---------|---------|----------|
| 0.1               | 0.8926 | 0.84   | 0.6751| 0.7373 | 0.8552  | 0.7775  | 0.7664   |
| 0.2               | 0.8925 | 0.8433 | 0.657 | 0.7469 | 0.8591  | 0.7775  | 0.7664   |
| 0.5               | 0.889  | 0.8422 | 0.6643| 0.7421 | 0.8454  | 0.7783  | 0.7545   |
| 1                 | 0.8903 | 0.842  | 0.6643| 0.7402 | 0.8421  | 0.7783  | 0.7545   |
| 2                 | 0.8901 | 0.8424 | 0.6643| 0.7421 | 0.8451  | 0.7783  | 0.7545   |
| 5                 | 0.8864 | 0.8456 | 0.6643| 0.7229 | 0.8424  | 0.7783  | 0.7545   |
|                   | 0.8601 | 0.8647 | 0.8681| 0.8601 | 0.8635  | 0.8624  | 0.7533   |
|                   | 0.8635 | 0.8578 | 0.8635| 0.8624 | 0.8635  | 0.8624  | 0.7533   |
|                   | 0.8624 | 0.8624 | 0.8601| 0.8647 | 0.8624  | 0.8624  | 0.7533   |
|                   |        |        |       |        | 0.8925  | 0.8926  | 0.8926   |
|                   |        |        |       |        | 0.8901  | 0.8901  | 0.8901   |
|                   |        |        |       |        | 0.889   | 0.89     | 0.89     |
|                   |        |        |       |        | 0.8903  | 0.8903  | 0.8903   |
|                   |        |        |       |        | 0.8907  | 0.8907  | 0.8907   |
|                   |        |        |       |        | 0.8864  | 0.8864  | 0.8864   |
## A.3.2 Single-Match Experiments

| MRPC | bert-small | First | Last | Dilatation | First-1 | Last-1 |
|------|------------|-------|------|------------|---------|--------|
| attention_mse_sum | 0.7868 | 0.7672 | 0.799 | 0.7917 | 0.8137 |
| attention_ce_mean | 0.7843 | 0.7819 | 0.7819 | 0.7892 | 0.8015 |
| hidden_mse | 0.7598 | 0.7672 | 0.7623 | 0.7917 | 0.7966 |
| mmd | 0.8113 | 0.7843 | 0.799 | 0.7917 | 0.799 |
| gram | 0.8064 | 0.8162 | 0.8235 | 0.799 | 0.8064 |
| cos | 0.7721 | 0.7598 | 0.7721 | 0.8186 | 0.7868 |
| pkd | 0.7525 | 0.8456 | 0.8284 | 0.8284 | 0.8407 |
| query_relation | 0.8137 | 0.799 | 0.799 | 0.799 | 0.8134 |
| key_relation | 0.8088 | 0.8015 | 0.8039 | 0.8137 | 0.7892 |
| value_relation | 0.8015 | 0.8015 | 0.8039 | 0.799 | 0.8039 |
| bert-mini | First | Last | Dilatation | First-1 | Last-1 |
| attention_mse_sum | 0.7941 | 0.7353 | 0.7426 | 0.8088 | 0.826 |
| attention_ce_mean | 0.8088 | 0.8137 | 0.7917 | 0.8186 | 0.8186 |
| hidden_mse | 0.75 | 0.75 | 0.7598 | 0.8088 | 0.8186 |
| mmd | 0.8333 | 0.8186 | 0.8431 | 0.8186 | 0.8064 |
| gram | 0.8333 | 0.8064 | 0.8358 | 0.8064 | 0.8186 |
| cos | 0.7549 | 0.7451 | 0.7475 | 0.8186 | 0.7721 |
| pkd | 0.7426 | 0.826 | 0.7941 | 0.8113 | 0.8088 |
| query_relation | 0.8211 | 0.8211 | 0.8235 | 0.8333 | 0.8145 |
| key_relation | 0.8284 | 0.8211 | 0.8235 | 0.8235 | 0.777 |
| value_relation | 0.8186 | 0.826 | 0.8284 | 0.8162 | 0.8137 |
| bert-tiny | First | Last | Dilatation | First-1 | Last-1 |
| attention_mse_sum | 0.7647 | 0.7623 | 0.7623 | 0.7598 | 0.7377 |
| attention_ce_mean | 0.7672 | 0.7623 | 0.7647 | 0.7647 | 0.7279 |
| hidden_mse | 0.7475 | 0.7451 | 0.7475 | 0.7672 | 0.723 |
| mmd | 0.7549 | 0.7647 | 0.7672 | 0.7574 | 0.7304 |
| gram | 0.7721 | 0.7647 | 0.7721 | 0.7745 | 0.7206 |
| cos | 0.7328 | 0.723 | 0.7328 | 0.7598 | 0.723 |
| pkd | 0.6838 | 0.723 | 0.7255 | 0.7402 | 0.7328 |
| query_relation | 0.7304 | 0.7279 | 0.7279 | 0.7328 | 0.7347 |
| key_relation | 0.7377 | 0.7304 | 0.7328 | 0.7279 | 0.7206 |
| value_relation | 0.7328 | 0.7328 | 0.7304 | 0.7304 | 0.7402 |
| SST2 | bert-small | First | Last | Dilatation | First-1 | Last-1 |
| attention_mse_sum | 0.8922 | 0.8956 | 0.8922 | 0.8865 | 0.8922 |
| attention_ce_mean | 0.8968 | 0.8956 | 0.8956 | 0.8865 | 0.8922 |
| hidden_mse | 0.8819 | 0.8853 | 0.8842 | 0.8968 | 0.8991 |
| mmd | 0.8876 | 0.8979 | 0.8922 | 0.8933 | 0.8933 |
| gram | 0.8911 | 0.8911 | 0.8922 | 0.8911 | 0.8956 |
| cos | 0.8807 | 0.8796 | 0.8807 | 0.8899 | 0.8968 |
| pkd | 0.8704 | 0.8819 | 0.8796 | 0.8899 | 0.8968 |
| query_relation | 0.8899 | 0.8911 | 0.8899 | 0.8933 | 0.8934 |
| key_relation | 0.8865 | 0.8853 | 0.8876 | 0.8911 | 0.8968 |
| value_relation | 0.8933 | 0.8876 | 0.8911 | 0.8865 | 0.8968 |
| bert-mini | First | Last | Dilatation | First-1 | Last-1 |
| attention_mse_sum | 0.8612 | 0.8693 | 0.8589 | 0.8635 | 0.8704 |
| Table 5 continued from previous page |
|--------------------------------------|
| attention_ce_mean | 0.8612 | 0.8681 | 0.8727 | 0.8647 | 0.867 | 0.8578 | 0.8544 | 0.8544 | 0.8647 | 0.8739 | 0.867 | 0.867 | 0.8647 | 0.867 | 0.867 |
| hidden_mse | 0.8681 | 0.8716 | 0.8784 | 0.8681 | 0.8739 | 0.867 | 0.867 | 0.867 | 0.8739 |
| mmd | 0.867 | 0.8658 | 0.867 | 0.867 | 0.867 |
| gram | 0.8291 | 0.8349 | 0.8337 | 0.8647 | 0.867 |
| cos | 0.82 | 0.8429 | 0.8452 | 0.8578 | 0.8681 |
| pkd | 0.8647 | 0.8567 | 0.8693 | 0.8612 | 0.8646 |
| query_relation | 0.8624 | 0.8658 | 0.8647 | 0.8635 | 0.8727 |
| value_relation | 0.8624 | 0.8635 | 0.8578 | 0.8589 | 0.8693 |

| bert-tiny | First | Last | Dilatation | First-1 | Last-1 |
|-----------|-------|------|------------|---------|-------|
| attention_mse_sum | 0.8257 | 0.8257 | 0.8222 | 0.82 | 0.8245 |
| attention_ce_mean | 0.8234 | 0.8222 | 0.8211 | 0.8234 | 0.8234 |
| hidden_mse | 0.8234 | 0.8245 | 0.828 | 0.8234 | 0.8257 |
| mmd | 0.8291 | 0.8234 | 0.8222 | 0.8234 | 0.8257 |
| gram | 0.8234 | 0.8211 | 0.8245 | 0.8222 | 0.8257 |
| cos | 0.82 | 0.8222 | 0.8268 | 0.8211 | 0.8291 |
| pkd | 0.82 | 0.8257 | 0.8245 | 0.8245 | 0.8314 |
| query_relation | 0.8222 | 0.8314 | 0.8222 | 0.8234 | 0.8245 |
| key_relation | 0.8314 | 0.8222 | 0.8314 | 0.8245 | 0.8268 |
| value_relation | 0.8245 | 0.8245 | 0.8245 | 0.8245 | 0.8257 |

**QQP**

| bert-small | First | Last | Dilatation | First-1 | Last-1 |
|------------|-------|------|------------|---------|-------|
| attention_mse_sum | 0.9003 | 0.8999 | 0.9002 | 0.901 | 0.8934 |
| attention_ce_mean | 0.9005 | 0.9007 | 0.9008 | 0.9001 | 0.8955 |
| hidden_mse | 0.8983 | 0.9013 | 0.9009 | 0.8995 | 0.8962 |
| mmd | 0.9001 | 0.9002 | 0.9021 | 0.8995 | 0.895 |
| gram | 0.9 | 0.9028 | 0.9023 | 0.9011 | 0.8953 |
| cos | 0.8979 | 0.9 | 0.9017 | 0.9007 | 0.8977 |
| pkd | 0.8953 | 0.9014 | 0.9037 | 0.8987 | 0.8982 |
| query_relation | 0.8941 | 0.8938 | 0.8947 | 0.895 | 0.8939 |
| key_relation | 0.8945 | 0.8952 | 0.8938 | 0.8941 | 0.8955 |
| value_relation | 0.8943 | 0.8959 | 0.8948 | 0.8936 | 0.8943 |

| bert-mini | First | Last | Dilatation | First-1 | Last-1 |
|------------|-------|------|------------|---------|-------|
| attention_mse_sum | 0.8909 | 0.8887 | 0.889 | 0.8939 | 0.8899 |
| attention_ce_mean | 0.8914 | 0.894 | 0.893 | 0.8945 | 0.8895 |
| hidden_mse | 0.8896 | 0.8926 | 0.8937 | 0.8918 | 0.8907 |
| mmd | 0.8919 | 0.8916 | 0.8917 | 0.8921 | 0.8898 |
| gram | 0.8928 | 0.8933 | 0.8931 | 0.8928 | 0.8922 |
| cos | 0.8876 | 0.8874 | 0.8884 | 0.8916 | 0.8925 |
| pkd | 0.8857 | 0.8911 | 0.8903 | 0.8919 | 0.894 |
| query_relation | 0.8906 | 0.8895 | 0.8911 | 0.8897 | 0.8878 |
| key_relation | 0.8897 | 0.8896 | 0.8901 | 0.8908 | 0.8909 |
| value_relation | 0.8903 | 0.8904 | 0.8911 | 0.8908 | 0.8888 |

| bert-tiny | First | Last | Dilatation | First-1 | Last-1 |
|------------|-------|------|------------|---------|-------|
| attention_mse_sum | 0.8634 | 0.8644 | 0.866 | 0.8663 | 0.8719 |
| attention_ce_mean | 0.8657 | 0.8645 | 0.8659 | 0.8643 | 0.872 |
| hidden_mse | 0.8615 | 0.8613 | 0.864 | 0.8607 | 0.8707 |
| mmd | 0.8665 | 0.8628 | 0.8642 | 0.8655 | 0.8719 |
| gram | 0.8666 | 0.8657 | 0.867 | 0.8659 | 0.872 |
| cos | 0.8626 | 0.8651 | 0.8626 | 0.8606 | 0.8689 |
| pkd | 0.8619 | 0.8671 | 0.8685 | 0.8591 | 0.8734 |
| query_relation | 0.8708 | 0.8714 | 0.8714 | 0.8703 | 0.8715 |
| key_relation | 0.8714 | 0.8711 | 0.8716 | 0.8699 | 0.8716 |
Table 5 continued from previous page

|          | QNLI          | RTE          |
|----------|--------------|--------------|
|          | bert-small   | bert-mini    | bert-tiny    |
| value_relation | 0.8713       | 0.8718       |
|           | 0.8708       | 0.874        |
|           | 0.8716       |              |
|           | 0.8718       |              |
| attention_mse_sum | 0.8602       | 0.8378       |
| attention_ce_mean | 0.8706       | 0.8417       |
| hidden_mse | 0.842        | 0.8329       |
| mmd       | 0.8695       | 0.8418       |
| gram      | 0.8666       | 0.8422       |
| cos       | 0.8389       | 0.8272       |
| pkd       | 0.8221       | 0.806        |
| query_relation | 0.8695       | 0.8446       |
| key_relation | 0.8724       | 0.8475       |
| value_relation | 0.8726       | 0.8431       |
|            |              |              |
| attention_mse_sum | 0.8706       | 0.8378       |
| attention_ce_mean | 0.8706       | 0.8417       |
| hidden_mse | 0.842        | 0.8329       |
| mmd       | 0.8695       | 0.8418       |
| gram      | 0.8666       | 0.8422       |
| cos       | 0.8389       | 0.8272       |
| pkd       | 0.8221       | 0.806        |
| query_relation | 0.8695       | 0.8446       |
| key_relation | 0.8724       | 0.8475       |
| value_relation | 0.8726       | 0.8431       |
|            |              |              |
| attention_mse_sum | 0.8706       | 0.8378       |
| attention_ce_mean | 0.8706       | 0.8417       |
| hidden_mse | 0.842        | 0.8329       |
| mmd       | 0.8695       | 0.8418       |
| gram      | 0.8666       | 0.8422       |
| cos       | 0.8389       | 0.8272       |
| pkd       | 0.8221       | 0.806        |
| query_relation | 0.8695       | 0.8446       |
| key_relation | 0.8724       | 0.8475       |
| value_relation | 0.8726       | 0.8431       |
| Table 5 continued from previous page |  
|-------------------------------------|
| cos                                | 0.5812 0.5957 0.574 0.6715 0.6209 |
| pkd                                | 0.556 0.5776 0.5812 0.657 0.6245 |
| query_relation                     | 0.6859 0.6823 0.6823 0.6787 0.6789 |
| key_relation                       | 0.6715 0.6715 0.6715 0.6751 0.6426 |
| value_relation                     | 0.6823 0.7004 0.6823 0.6715 0.6715 |
| bert-tiny                          | First Last Dilatation First-1 Last-1 |
| attention_mse_sum                 | 0.6173 0.5957 0.6029 0.6209 0.6101 |
| attention_ce_mean                 | 0.6173 0.6065 0.6065 0.6173 0.6173 |
| hidden_mse                        | 0.5993 0.6209 0.6101 0.6173 0.5884 |
| mmd                               | 0.6173 0.5957 0.5957 0.6209 0.6137 |
| gram                              | 0.6137 0.6065 0.6209 0.6173 0.6209 |
| cos                               | 0.5921 0.6209 0.6065 0.6065 0.5812 |
| pkd                               | 0.6137 0.6065 0.6245 0.6137 0.5993 |
| query_relation                    | 0.6354 0.6426 0.6426 0.639 0.6354 |
| key_relation                      | 0.639 0.6354 0.6354 0.6354 0.6101 |
| value_relation                    | 0.6245 0.6354 0.6245 0.639 0.6282 |
| bert-small                         | First Last Dilatation First-1 Last-1 |
| attention_mse_sum                 | 0.767 0.7383 0.7478 0.768 0.7766 |
| attention_ce_mean                 | 0.7718 0.7613 0.7603 0.7632 0.7881 |
| hidden_mse                        | 0.6932 0.6913 0.7066 0.768 0.7804 |
| mmd                               | 0.7574 0.7517 0.7565 0.7603 0.7766 |
| gram                              | 0.7747 0.7603 0.7641 0.7555 0.7804 |
| cos                               | 0.6942 0.6942 0.6989 0.7593 0.7718 |
| pkd                               | 0.6913 0.7066 0.7057 0.7661 0.768 |
| query_relation                    | 0.7709 0.7651 0.7718 0.7718 0.7745 |
| key_relation                      | 0.7718 0.7728 0.7728 0.7728 0.7756 |
| value_relation                    | 0.7689 0.7776 0.7737 0.7728 0.7795 |
| bert-mini                         | First Last Dilatation First-1 Last-1 |
| attention_mse_sum                 | 0.6922 0.6989 0.6932 0.7392 0.745 |
| attention_ce_mean                 | 0.7354 0.7354 0.7383 0.744 0.745 |
| hidden_mse                        | 0.6951 0.6951 0.6932 0.7383 0.743 |
| mmd                               | 0.7229 0.698 0.6951 0.744 0.7459 |
| gram                              | 0.7488 0.7354 0.744 0.7363 0.745 |
| cos                               | 0.6942 0.6913 0.6913 0.7248 0.7181 |
| pkd                               | 0.6913 0.6913 0.6913 0.743 0.7335 |
| query_relation                    | 0.7459 0.7469 0.744 0.744 0.7457 |
| key_relation                      | 0.7469 0.7411 0.745 0.745 0.745 |
| value_relation                    | 0.7392 0.7383 0.7469 0.743 0.743 |
| bert-tiny                         | First Last Dilatation First-1 Last-1 |
| attention_mse_sum                 | 0.6932 0.6913 0.6913 0.6913 0.6913 |
| attention_ce_mean                 | 0.6913 0.6913 0.6961 0.6913 0.6999 |
| hidden_mse                        | 0.6913 0.6913 0.6913 0.6913 0.6913 |
| mmd                               | 0.6913 0.6913 0.6932 0.6922 0.6989 |
| gram                              | 0.6913 0.6942 0.6913 0.6932 0.6951 |
| cos                               | 0.6951 0.6913 0.6913 0.6951 0.6913 |
| pkd                               | 0.6913 0.6922 0.6913 0.6913 0.6913 |
| query_relation                    | 0.6913 0.6913 0.6913 0.6913 0.6913 |
| key_relation                      | 0.6913 0.6913 0.6913 0.6913 0.6942 |
| value_relation                    | 0.6913 0.6913 0.6913 0.6913 0.6942 |

| STSB                               |
|------------------------------------|
| bert-small                         | First Last Dilatation First-1 Last-1 |
| attention_mse_sum                 | 0.8717 0.8731 0.8713 0.8728 0.8715 |
Table 5 continued from previous page

|                         | bert-mini | First | Last | Dilatation | First-1 | Last-1 |
|-------------------------|-----------|-------|------|------------|---------|--------|
| attention_ce_mean       | 0.8702    | 0.8699| 0.8651| 0.8646     | 0.8724  |
| hidden_mse              | 0.8689    | 0.8696| 0.8687| 0.8597     | 0.8735  |
| mmd                     | 0.8725    | 0.8708| 0.8693| 0.8705     | 0.8724  |
| gram                    | 0.8726    | 0.8718| 0.873 | 0.8716     | 0.8737  |
| cos                     | 0.8726    | 0.8661| 0.8692| 0.8626     | 0.8725  |
| pkd                     | 0.864     | 0.8672| 0.8678| 0.8594     | 0.873   |
| query_relation          | 0.8697    | 0.8711| 0.87  | 0.8715     | 0.8731  |
| key_relation            | 0.8731    | 0.8712| 0.8704| 0.8714     | 0.8733  |
| value_relation          | 0.87      | 0.8708| 0.8706| 0.8711     | 0.8718  |
|                         | bert-tiny | First | Last | Dilatation | First-1 | Last-1 |
| attention_mse_sum       | 0.8669    | 0.865 | 0.8653| 0.8556     | 0.8652  |
| attention_ce_mean       | 0.8656    | 0.8643| 0.8645| 0.8671     | 0.8634  |
| hidden_mse              | 0.8637    | 0.8613| 0.8494| 0.8675     | 0.8636  |
| mmd                     | 0.8574    | 0.8485| 0.866 | 0.8536     | 0.8627  |
| gram                    | 0.8651    | 0.8648| 0.8652| 0.8585     | 0.863   |
| cos                     | 0.8634    | 0.8533| 0.8633| 0.8598     | 0.8647  |
| pkd                     | 0.8402    | 0.8384| 0.8595| 0.8519     | 0.865   |
| query_relation          | 0.862     | 0.8622| 0.8611| 0.8627     | 0.8663  |
| key_relation            | 0.8631    | 0.8612| 0.8598| 0.8624     | 0.8674  |
| value_relation          | 0.8615    | 0.861 | 0.8604| 0.8632     | 0.8646  |
|                         | bert-small| First | Last | Dilatation | First-1 | Last-1 |
| attention_mse_sum       | 0.7761    | 0.7822| 0.6231| 0.7797     | 0.8165  |
| attention_ce_mean       | 0.7813    | 0.7816| 0.7871| 0.6145     | 0.8167  |
| hidden_mse              | 0.7796    | 0.7848| 0.7859| 0.6546     | 0.8166  |
| mmd                     | 0.7879    | 0.6161| 0.6075| 0.6266     | 0.8168  |
| gram                    | 0.7802    | 0.7758| 0.7795| 0.6325     | 0.8166  |
| cos                     | 0.6664    | 0.7857| 0.6735| 0.6776     | 0.8163  |
| pkd                     | 0.6745    | 0.6911| 0.7814| 0.6685     | 0.8156  |
| query_relation          | 0.7969    | 0.7969| 0.8058| 0.8065     | 0.8155  |
| key_relation            | 0.804     | 0.7969| 0.7969| 0.8065     | 0.8173  |
| value_relation          | 0.8065    | 0.797 | 0.7989| 0.797      | 0.8166  |
|                         | MNLI-mm   | First | Last | Dilatation | First-1 | Last-1 |
| attention_mse_sum       | 0.7861    | 0.7811| 0.787 | 0.7886     | 0.8008  |
| attention_ce_mean       | 0.788    | 0.7877| 0.7916| 0.79       | 0.8011  |
| hidden_mse              | 0.7764    | 0.776 | 0.779 | 0.7862     | 0.8003  |
| mmd                     | 0.7895    | 0.7911| 0.7866| 0.7876     | 0.7998  |
| gram                    | 0.7867    | 0.7881| 0.7881| 0.79       | 0.8001  |
| cos                     | 0.7679    | 0.7735| 0.7833| 0.7851     | 0.8014  |
| pkd                     | 0.7483    | 0.7883| 0.7936| 0.7872     | 0.8098  |
| query_relation          | 0.7916    | 0.7917| 0.7918| 0.7918     | 0.7995  |
| key_relation            | 0.7926    | 0.7916| 0.7894| 0.7923     | 0.7983  |
| value_relation          | 0.7912    | 0.7927| 0.7905| 0.7922     | 0.8008  |
|                         | bert-mini | First | Last | Dilatation | First-1 | Last-1 |
| attention_mse_sum       | 0.761     | 0.7457| 0.7522| 0.7717     | 0.7831  |
| attention_ce_mean       | 0.7722    | 0.774 | 0.7722| 0.7736     | 0.7829  |
| hidden_mse              | 0.7476    | 0.7492| 0.76  | 0.7674     | 0.7833  |
| mmd                     | 0.7724    | 0.7686| 0.773 | 0.7729     | 0.782   |
| gram                    | 0.7707    | 0.7653| 0.7727| 0.7723     | 0.7847  |
| cos                     | 0.7388    | 0.7485| 0.7459| 0.7632     | 0.7859  |
| pkd                     | 0.7294    | 0.7632| 0.7645| 0.7672     | 0.7928  |
| query_relation          | 0.7728    | 0.7749| 0.7726| 0.7736     | 0.7844  |
| key_relation            | 0.7735    | 0.7731| 0.7733| 0.7743     | 0.7833  |
| Table 5 continued from previous page |
|---------------------------------------|
|                                       |
| value_relation                        |
| value_relation                        |
| bert-tiny                             |
| First                                 |
| Last                                  |
| Dilatation                            |
| First-1                               |
| Last-1                                |
| 0.7731                                |
| 0.7728                                |
| 0.7743                                |
| 0.7732                                |
| 0.7826                                |
| attention_mse_sum                     |
| 0.7072                                |
| 0.7013                                |
| 0.702                          |
| 0.7036                                |
| 0.7181                                |
| attention_ce_mean                     |
| 0.702                                |
| 0.7063                                |
| 0.7035                                |
| 0.7024                                |
| 0.7191                                |
| hidden_mse                            |
| 0.6974                                |
| 0.7035                                |
| 0.6995                                |
| 0.7007                                |
| 0.7284                                |
| mmd                                  |
| 0.7043                                |
| 0.7063                                |
| 0.704                        |
| 0.7053                                |
| 0.7287                                |
| gram                                 |
| 0.7041                                |
| 0.7002                                |
| 0.7005                        |
| 0.7035                                |
| 0.7175                                |
| cos                                  |
| 0.6915                                |
| 0.6928                                |
| 0.6961                        |
| 0.6988                                |
| 0.7233                                |
| pkd                                  |
| 0.5794                                |
| 0.6989                                |
| 0.7002                        |
| 0.6969                                |
| 0.7321                                |
| query_relation                       |
| 0.711                                |
| 0.71                                |
| 0.71                        |
| 0.7112                                |
| 0.7282                                |
| key_relation                         |
| 0.709                                |
| 0.7088                                |
| 0.7083                        |
| 0.7104                                |
| 0.7281                                |
| value_relation                       |
| 0.7097                                |
| 0.7074                                |
| 0.7087                        |
| 0.7089                                |
| 0.7276                                |

| MNLI-m                                |
|---------------------------------------|
|                                       |
| bert-small                            |
| First                                 |
| Last                                  |
| Dilatation                            |
| First-1                               |
| Last-1                                |
| 0.7943                                |
| 0.7877                                |
| 0.7893                                |
| 0.7915                                |
| 0.8047                                |
| attention_ce_mean                     |
| 0.7915                                |
| 0.7917                                |
| 0.7946                                |
| 0.7947                                |
| 0.7954                                |
| hidden_mse                            |
| 0.7853                                |
| 0.7818                                |
| 0.7873                                |
| 0.7944                                |
| 0.8004                                |
| mmd                                  |
| 0.7911                                |
| 0.7921                                |
| 0.7887                                |
| 0.7885                                |
| 0.7991                                |
| gram                                 |
| 0.7949                                |
| 0.7912                                |
| 0.7959                        |
| 0.7896                                |
| 0.799                                |
| cos                                  |
| 0.781                                |
| 0.7824                                |
| 0.787                        |
| 0.7928                                |
| 0.8013                                |
| pkd                                  |
| 0.7686                                |
| 0.7905                                |
| 0.7947                        |
| 0.7924                                |
| 0.807                                |
| query_relation                       |
| 0.792                                |
| 0.7926                                |
| 0.7926                        |
| 0.7926                                |
| 0.793                                |
| key_relation                         |
| 0.7934                                |
| 0.7929                                |
| 0.791                        |
| 0.7938                                |
| 0.8011                                |
| value_relation                       |
| 0.7936                                |
| 0.7933                                |
| 0.7941                        |
| 0.7917                                |
| 0.7946                                |
| bert-mini                            |
| First                                 |
| Last                                  |
| Dilatation                            |
| First-1                               |
| Last-1                                |
| 0.7606                                |
| 0.7532                                |
| 0.7555                                |
| 0.7631                                |
| 0.7785                                |
| attention_mse_sum                     |
| 0.7643                                |
| 0.7661                                |
| 0.7627                                |
| 0.766                                |
| 0.7787                                |
| attention_ce_mean                     |
| 0.7575                                |
| 0.7555                                |
| 0.761                        |
| 0.7614                                |
| 0.7783                                |
| hidden_mse                            |
| 0.7655                                |
| 0.7628                                |
| 0.7624                        |
| 0.7678                                |
| 0.7788                                |
| mmd                                  |
| 0.7619                                |
| 0.7637                                |
| 0.7669                        |
| 0.7668                                |
| 0.7763                                |
| gram                                 |
| 0.7493                                |
| 0.7511                                |
| 0.7583                        |
| 0.7578                                |
| 0.7798                                |
| cos                                  |
| 0.7244                                |
| 0.7576                                |
| 0.7588                        |
| 0.7602                                |
| 0.7827                                |
| pkd                                  |
| 0.7694                                |
| 0.7689                                |
| 0.7698                        |
| 0.7697                                |
| 0.7744                                |
| query_relation                       |
| 0.7702                                |
| 0.767                                |
| 0.7694                        |
| 0.7699                                |
| 0.776                                |
| key_relation                         |
| 0.7691                                |
| 0.7676                                |
| 0.7665                        |
| 0.7693                                |
| 0.7753                                |
| value_relation                       |
| 0.7025                                |
| 0.6977                                |
| 0.6977                                |
| 0.7056                                |
| 0.7172                                |
| bert-tiny                            |
| First                                 |
| Last                                  |
| Dilatation                            |
| First-1                               |
| Last-1                                |
| 0.7025                                |
| 0.6977                                |
| 0.6977                                |
| 0.7056                                |
| 0.7172                                |
| attention_mse_sum                     |
| 0.7069                                |
| 0.7028                                |
| 0.7043                                |
| 0.7038                                |
| 0.7196                                |
| attention_ce_mean                     |
| 0.7028                                |
| 0.7049                                |
| 0.7033                                |
| 0.7018                                |
| 0.7252                                |
| hidden_mse                            |
| 0.7083                                |
| 0.7006                                |
| 0.702                        |
| 0.7043                                |
| 0.726                                |
| mmd                                  |
| 0.7035                                |
| 0.7052                                |
| 0.7042                        |
| 0.7038                                |
| 0.7158                                |
| gram                                 |
| 0.6977                                |
| 0.7018                                |
| 0.6993                        |
| 0.7016                                |
| 0.7261                                |
| cos                                  |
| 0.6772                                |
| 0.6968                                |
| 0.6975                        |
| 0.6997                                |
| 0.7202                                |
| pkd                                  |
| 0.7074                                |
| 0.7086                                |
| 0.7075                        |
| 0.7082                                |
| 0.70456                               |
| query_relation                       |
| 0.7088                                |
| 0.7087                                |
| 0.7087                        |
| 0.709                                |
| 0.7242                                |
| key_relation                         |
| 0.709                                |
| 0.708                                |
| 0.7092                        |
| 0.7106                                |
| 0.7169                                |

A.3.3 Double-Match Experiments

| MRPC                                |
|-------------------------------------|
|                                       |
| bert-small                          |
| First                               |
| Last                                |
| Dilatation                          |
| First-1                                   |
| Last-1                                |
|                                       |

21
|                                | bert-mini | First       | Last        | Dilatation | First-1     | Last-1     |
|--------------------------------|-----------|-------------|-------------|------------|-------------|------------|
| attention_ce_mean.cos         | 0.7623    | 0.7304      | 0.7451      | 0.8309     | 0.7745      |            |
| attention_ce_mean.gram        | 0.8431    | 0.8186      | 0.8407      | 0.8211     | 0.8186      |            |
| attention_ce_mean.hidden_mse  | 0.75      | 0.7475      | 0.7475      | 0.8309     | 0.8088      |            |
| attention_ce_mean.key_relation| 0.799     | 0.777       | 0.777       | 0.826      | 0.777       |            |
| attention_ce_mean.mmd         | 0.8211    | 0.8162      | 0.8137      | 0.8211     | 0.8039      |            |
| attention_ce_mean.pkd         | 0.6838    | 0.8162      | 0.777       | 0.8284     | 0.8137      |            |
| attention_ce_mean.query_relation| 0.7892   | 0.7745      | 0.7794      | 0.8162     | 0.7794      |            |
| attention_ce_mean.value_relation| 0.8039   | 0.8064      | 0.8088      | 0.8015     | 0.8113      |            |
| attention_mse_sum.cos         | 0.7623    | 0.7157      | 0.7377      | 0.8235     | 0.7966      |            |
| attention_mse_sum.gram        | 0.8064    | 0.7328      | 0.7549      | 0.8186     | 0.826       |            |
| attention_mse_sum.hidden_mse  | 0.7598    | 0.723       | 0.7475      | 0.8431     | 0.8211      |            |
| attention_mse_sum.key_relation| 0.8088    | 0.7304      | 0.7426      | 0.8162     | 0.7966      |            |
| attention_mse_sum.mmd         | 0.8113    | 0.7304      | 0.7353      | 0.8333     | 0.8235      |            |
| attention_mse_sum.pkd         | 0.6838    | 0.7451      | 0.7574      | 0.8358     | 0.8137      |            |
| attention_mse_sum.query_relation| 0.8113   | 0.7304      | 0.7377      | 0.8333     | 0.7974      |            |
| attention_mse_sum.value_relation| 0.8039   | 0.7549      | 0.7574      | 0.8358     | 0.8235      |            |
| cos.key_relation              | 0.7525    | 0.7353      | 0.7451      | 0.8235     | 0.7745      |            |
| cos.query_relation            | 0.7525    | 0.7328      | 0.7525      | 0.826      | 0.7549      |            |
| cos.value_relation            | 0.75      | 0.7377      | 0.7426      | 0.8309     | 0.777       |            |
| gram.key_relation             | 0.8309    | 0.7696      | 0.7672      | 0.8211     | 0.777       |            |
| gram.query_relation           | 0.7966    | 0.7574      | 0.7745      | 0.826      | 0.7721      |            |
| gram.value_relation           | 0.8137    | 0.8064      | 0.8064      | 0.8235     | 0.8309      |            |
| hidden_mse.key_relation       | 0.75      | 0.7549      | 0.7598      | 0.799      | 0.8015      |            |
| hidden_mse.query_relation     | 0.7525    | 0.7549      | 0.7598      | 0.799      | 0.8015      |            |
| hidden_mse.value_relation     | 0.7549    | 0.7549      | 0.7598      | 0.799      | 0.8015      |            |
| mmd.key_relation              | 0.8015    | 0.7721      | 0.7868      | 0.799      | 0.799       | 0.799      |
| mmd.query_relation            | 0.7941    | 0.7868      | 0.7868      | 0.799      | 0.799       | 0.799      |
| mmd.value_relation            | 0.8284    | 0.8088      | 0.799       | 0.799      | 0.799       | 0.799      |
| pkd.key_relation              | 0.777     | 0.7745      | 0.7794      | 0.8162     | 0.8088      |            |
| pkd.query_relation            | 0.7721    | 0.777       | 0.7843      | 0.8088     | 0.8309      |            |
| pkd.value_relation            | 0.7647    | 0.8211      | 0.8211      | 0.8088     | 0.8358      |            |
| mmd, key_relation | 0.8211 | 0.7574 | 0.7794 | 0.8235 | 0.7696 |
|------------------|--------|--------|--------|--------|--------|
| mmd, query_relation | 0.8064 | 0.7696 | 0.7745 | 0.8186 | 0.7745 |
| mmd, value_relation | 0.8235 | 0.8186 | 0.8186 | 0.826 | 0.8162 |
| pkd, key_relation | 0.6838 | 0.7696 | 0.7745 | 0.828 | 0.7868 |
| pkd, query_relation | 0.6863 | 0.7696 | 0.7672 | 0.8309 | 0.7941 |
| pkd, value_relation | 0.6863 | 0.8088 | 0.7917 | 0.824 | 0.8113 |

| bert-tiny | First | Last | Dilatation | First-1 | Last-1 |
|-----------|-------|------|------------|--------|--------|
| attention_ce_mean, cos | 0.7353 | 0.7206 | 0.7108 | 0.7402 | 0.7059 |
| attention_ce_mean, gram | 0.7206 | 0.7328 | 0.7304 | 0.7304 | 0.723 |
| attention_ce_mean, hidden mse | 0.7353 | 0.7304 | 0.7255 | 0.7353 | 0.7206 |
| attention_ce_mean, key_relation | 0.7279 | 0.7157 | 0.7206 | 0.723 | 0.7181 |
| attention_ce_mean, mmd | 0.7353 | 0.7451 | 0.7475 | 0.7426 | 0.7206 |
| attention_ce_mean, pkd | 0.6838 | 0.7475 | 0.7475 | 0.7598 | 0.7328 |
| attention_ce_mean, query_relation | 0.7255 | 0.7279 | 0.7206 | 0.7255 | 0.7279 |
| attention_ce_mean, value_relation | 0.7206 | 0.7377 | 0.7304 | 0.723 | 0.7377 |
| attention_mse_sum, cos | 0.7279 | 0.7206 | 0.7108 | 0.7426 | 0.723 |
| attention_mse_sum, gram | 0.7402 | 0.7255 | 0.7377 | 0.7279 | 0.7304 |
| attention_mse_sum, hidden mse | 0.7328 | 0.7206 | 0.7132 | 0.7353 | 0.7181 |
| attention_mse_sum, key_relation | 0.7451 | 0.723 | 0.7279 | 0.7206 | 0.7328 |
| attention_mse_sum, mmd | 0.7475 | 0.7353 | 0.7377 | 0.7451 | 0.7304 |
| attention_mse_sum, pkd | 0.6838 | 0.75 | 0.7647 | 0.7574 | 0.7402 |
| attention_mse_sum, query_relation | 0.7353 | 0.7328 | 0.723 | 0.7206 | 0.7328 |
| attention_mse_sum, value_relation | 0.7353 | 0.7328 | 0.7328 | 0.723 | 0.7451 |
| cos, key_relation | 0.7353 | 0.7206 | 0.7181 | 0.7426 | 0.7206 |
| cos, query_relation | 0.7353 | 0.7157 | 0.7108 | 0.7402 | 0.7157 |
| cos, value_relation | 0.7279 | 0.723 | 0.7206 | 0.7402 | 0.723 |
| gram, key_relation | 0.723 | 0.7279 | 0.7255 | 0.7279 | 0.7206 |
| gram, query_relation | 0.7279 | 0.7279 | 0.7279 | 0.7279 | 0.7206 |
| gram, value_relation | 0.7255 | 0.7279 | 0.7255 | 0.7353 | 0.7304 |
| hidden mse, key_relation | 0.7353 | 0.7304 | 0.7402 | 0.723 | 0.7206 |
| hidden mse, query_relation | 0.7426 | 0.7304 | 0.7377 | 0.7255 | 0.7206 |
| hidden mse, value_relation | 0.7279 | 0.7304 | 0.7255 | 0.723 | 0.7328 |
| mmd, key_relation | 0.7353 | 0.7574 | 0.7451 | 0.7353 | 0.7206 |
| mmd, query_relation | 0.7353 | 0.7525 | 0.7353 | 0.7304 | 0.7255 |
| mmd, value_relation | 0.7353 | 0.7525 | 0.7475 | 0.7451 | 0.7328 |
| pkd, key_relation | 0.6838 | 0.7525 | 0.7451 | 0.7574 | 0.7304 |
| pkd, query_relation | 0.6838 | 0.7475 | 0.7451 | 0.7549 | 0.7304 |
| pkd, value_relation | 0.6838 | 0.7426 | 0.7451 | 0.7598 | 0.7426 |

| bert-small | First | Last | Dilatation | First-1 | Last-1 |
|------------|-------|------|------------|--------|--------|
| attention_ce_mean, cos | 0.8853 | 0.8888 | 0.8865 | 0.8899 | 0.8979 |
| attention_ce_mean, gram | 0.8956 | 0.8922 | 0.8922 | 0.8979 | 0.8922 |
| attention_ce_mean, hidden mse | 0.8876 | 0.8956 | 0.8968 | 0.8911 | 0.8968 |
| attention_ce_mean, key_relation | 0.8956 | 0.8968 | 0.9014 | 0.8956 | 0.8979 |
| attention_ce_mean, mmd | 0.8922 | 0.8956 | 0.8888 | 0.8956 | 0.8945 |
| attention_ce_mean, pkd | 0.8739 | 0.8922 | 0.8922 | 0.8911 | 0.8945 |
| attention_ce_mean, query_relation | 0.9002 | 0.8876 | 0.8933 | 0.8979 | 0.8899 |
| attention_ce_mean, value_relation | 0.8979 | 0.8933 | 0.8968 | 0.8888 | 0.8991 |
| attention_mse_sum, cos | 0.8796 | 0.8899 | 0.9002 | 0.8933 | 0.8899 |
| attention_mse_sum, gram | 0.8956 | 0.9002 | 0.8956 | 0.8876 | 0.8968 |
| attention_mse_sum, hidden mse | 0.8865 | 0.8945 | 0.8956 | 0.8922 | 0.8979 |
| attention_mse_sum, key_relation | 0.8979 | 0.8979 | 0.8865 | 0.8933 | 0.8979 |
| attention_mse_sum, mmd | 0.8991 | 0.8979 | 0.8956 | 0.8945 | 0.8911 |
| attention_mse_sum, pkd | 0.8976 | 0.8899 | 0.8876 | 0.8945 | 0.9002 |
| attention_mse_sum, query_relation | 0.8888 | 0.8956 | 0.8933 | 0.8865 | 0.8956 |
| attention_mse_sum, value_relation | 0.8899 | 0.8968 | 0.8945 | 0.8853 | 0.8922 |
|                | First  | Last   | Dilatation | First-1 | Last-1 |
|----------------|--------|--------|------------|---------|--------|
| **cos, key_relation** | 0.883  | 0.8876 | 0.8888     | 0.8911  | 0.8911 |
| **cos, query_relation** | 0.8819 | 0.8899 | 0.8911     | 0.8945  | 0.8956 |
| **cos, value_relation** | 0.8876 | 0.8922 | 0.8888     | 0.9002  | 0.8933 |
| **gram, key_relation**  | 0.8933 | 0.8933 | 0.8899     | 0.8865  | 0.8888 |
| **gram, query_relation** | 0.8979 | 0.8979 | 0.8945     | 0.8853  | 0.8979 |
| **gram, value_relation** | 0.8991 | 0.9014 | 0.8979     | 0.8911  | 0.8922 |
| **hidden_mse, key_relation** | 0.8899 | 0.8899 | 0.8865     | 0.8876  | 0.906  |
| **hidden_mse, query_relation** | 0.8956 | 0.8865 | 0.8876     | 0.8945  | 0.9037 |
| **hidden_mse, value_relation** | 0.8922 | 0.8945 | 0.8922     | 0.8979  | 0.9025 |
| **mmd, key_relation**  | 0.8933 | 0.9002 | 0.8991     | 0.8865  | 0.8956 |
| **mmd, query_relation** | 0.8968 | 0.8979 | 0.8979     | 0.8968  | 0.8945 |
| **mmd, value_relation** | 0.8888 | 0.8911 | 0.8956     | 0.8979  | 0.8899 |
| **pkd, key_relation**  | 0.8819 | 0.8888 | 0.8899     | 0.8911  | 0.8956 |
| **pkd, query_relation** | 0.8761 | 0.8865 | 0.8876     | 0.8888  | 0.8933 |
| **pkd, value_relation** | 0.883  | 0.8933 | 0.8999     | 0.8911  | 0.8979 |

**bert-mini**

|                | First  | Last   | Dilatation | First-1 | Last-1 |
|----------------|--------|--------|------------|---------|--------|
| **attention_ce_mean, cos** | 0.8544 | 0.8452 | 0.8544     | 0.8704  | 0.8693 |
| **attention_ce_mean, gram** | 0.8761 | 0.8727 | 0.8727     | 0.8693  | 0.8693 |
| **attention_ce_mean, hidden_mse** | 0.867  | 0.8635 | 0.8704     | 0.8693  | 0.8727 |
| **attention_ce_mean, key_relation** | 0.8693 | 0.8727 | 0.867     | 0.867  | 0.8773 |
| **attention_ce_mean, mmd** | 0.867  | 0.8727 | 0.8693     | 0.8658  | 0.8704 |
| **attention_ce_mean, query_relation** | 0.8314 | 0.8578 | 0.8704     | 0.8647  | 0.8704 |
| **attention_ce_mean, value_relation** | 0.8704 | 0.8727 | 0.8739     | 0.8739  | 0.8704 |
| **attention_mse_sum, cos** | 0.8578 | 0.8463 | 0.844      | 0.8704  | 0.8704 |
| **attention_mse_sum, gram** | 0.8647 | 0.8761 | 0.8589     | 0.8704  | 0.8739 |
| **attention_mse_sum, hidden_mse** | 0.8635 | 0.8567 | 0.8601     | 0.8704  | 0.8773 |
| **attention_mse_sum, key_relation** | 0.867  | 0.8693 | 0.8658     | 0.8647  | 0.8739 |
| **attention_mse_sum, mmd** | 0.8635 | 0.875  | 0.8681     | 0.8658  | 0.8739 |
| **attention_mse_sum, query_relation** | 0.8211 | 0.8452 | 0.8486     | 0.867  | 0.8716 |
| **attention_mse_sum, value_relation** | 0.8704 | 0.8796 | 0.867      | 0.8681  | 0.8693 |
| **cos, key_relation** | 0.8658 | 0.8693 | 0.8739     | 0.8681  | 0.867  |
| **cos, query_relation** | 0.8612 | 0.8555 | 0.8635     | 0.8681  | 0.867  |
| **cos, value_relation** | 0.8555 | 0.8532 | 0.8658     | 0.8693  | 0.8704 |
| **gram, key_relation** | 0.8578 | 0.8498 | 0.8567     | 0.8658  | 0.8716 |
| **gram, query_relation** | 0.8727 | 0.867  | 0.8773     | 0.8716  | 0.8693 |
| **gram, value_relation** | 0.8739 | 0.8624 | 0.8727     | 0.8704  | 0.8693 |
| **hidden_mse, key_relation** | 0.8716 | 0.8727 | 0.8853     | 0.8704  | 0.8647 |
| **hidden_mse, query_relation** | 0.867  | 0.8589 | 0.8624     | 0.8693  | 0.8739 |
| **hidden_mse, value_relation** | 0.867  | 0.8589 | 0.867      | 0.8704  | 0.8693 |
| **mmd, key_relation** | 0.867  | 0.8567 | 0.8681     | 0.8681  | 0.8704 |
| **mmd, query_relation** | 0.867  | 0.8567 | 0.8681     | 0.8681  | 0.8704 |
| **mmd, value_relation** | 0.8624 | 0.8681 | 0.8693     | 0.8761  | 0.867  |
| **pkd, key_relation** | 0.8349 | 0.8498 | 0.8612     | 0.8612  | 0.8693 |
| **pkd, query_relation** | 0.8337 | 0.8417 | 0.8589     | 0.8612  | 0.875  |
| **pkd, value_relation** | 0.8394 | 0.8624 | 0.8681     | 0.8635  | 0.8704 |

**bert-tiny**

|                | First  | Last   | Dilatation | First-1 | Last-1 |
|----------------|--------|--------|------------|---------|--------|
| **attention_ce_mean, cos** | 0.8234 | 0.828  | 0.8314     | 0.8245  | 0.828  |
| **attention_ce_mean, gram** | 0.8234 | 0.8245 | 0.8234     | 0.8257  | 0.828  |
| **attention_ce_mean, hidden_mse** | 0.8257 | 0.8268 | 0.828      | 0.8257  | 0.8257 |
| **attention_ce_mean, key_relation** | 0.8234 | 0.8257 | 0.8234     | 0.8222  | 0.8245 |
| **attention_ce_mean, mmd** | 0.8257 | 0.8257 | 0.8245     | 0.8268  | 0.8245 |
| **attention_ce_mean, pkd** | 0.8291 | 0.8257 | 0.8326     | 0.8268  | 0.8326 |
| **attention_ce_mean, query_relation** | 0.8222 | 0.8234 | 0.8245     | 0.8245  | 0.8245 |
| **attention_ce_mean, value_relation** | 0.8234 | 0.8211 | 0.8211     | 0.8291  | 0.8257 |
| Table 6 continued from previous page |
|--------------------------------------|
| attention_mse_sum,cos               | 0.8268 | 0.8291 | 0.8291 | 0.8211 | 0.8245 |
| attention_mse_sum,gram              | 0.8268 | 0.8303 | 0.836  | 0.8257 | 0.8268 |
| attention_mse_sum,hidden_mse        | 0.8257 | 0.8257 | 0.828  | 0.8222 | 0.8268 |
| attention_mse_sum,key_relation      | 0.8245 | 0.8291 | 0.828  | 0.828  | 0.8234 |
| attention_mse_sum,mmd               | 0.8291 | 0.8211 | 0.8245 | 0.8291 | 0.8245 |
| attention_mse_sum,pkd               | 0.8211 | 0.8326 | 0.8314 | 0.8268 | 0.8303 |
| attention_mse_sum,query_relation    | 0.8268 | 0.8245 | 0.8245 | 0.8245 | 0.8257 |
| attention_mse_sum,value_relation    | 0.8245 | 0.828  | 0.8291 | 0.828  | 0.828  |
| cos,key_relation                     | 0.8257 | 0.8268 | 0.8268 | 0.828  | 0.828  |
| cos,query_relation                   | 0.8245 | 0.8245 | 0.8268 | 0.828  | 0.8291 |
| cos,value_relation                   | 0.8245 | 0.828  | 0.8291 | 0.8234 | 0.8268 |
| gram,key_relation                    | 0.8245 | 0.8234 | 0.8268 | 0.8268 | 0.8268 |
| gram,query_relation                  | 0.8303 | 0.8268 | 0.8245 | 0.8268 | 0.8245 |
| gram,value_relation                  | 0.8268 | 0.8268 | 0.8222 | 0.828  | 0.828  |
| hidden_mse,key_relation              | 0.8234 | 0.8291 | 0.8257 | 0.828  | 0.8245 |
| hidden_mse,query_relation            | 0.8245 | 0.828  | 0.8268 | 0.8268 | 0.8291 |
| hidden_mse,value_relation            | 0.828  | 0.8268 | 0.8268 | 0.828  | 0.828  |
| mmd,key_relation                     | 0.828  | 0.82  | 0.8234 | 0.8245 | 0.8268 |
| mmd,query_relation                   | 0.8257 | 0.8234 | 0.8257 | 0.8245 | 0.8257 |
| mmd,value_relation                   | 0.8222 | 0.8177 | 0.8211 | 0.8234 | 0.8234 |
| pkd,key_relation                     | 0.8268 | 0.8337 | 0.828  | 0.8268 | 0.8291 |
| pkd,query_relation                   | 0.8257 | 0.8314 | 0.828  | 0.828  | 0.8291 |
| pkd,value_relation                   | 0.8234 | 0.8326 | 0.8349 | 0.8234 | 0.836  |

| bert-small                           | First  | Last   | Dilatation | First-1 | Last-1  |
|--------------------------------------|--------|--------|------------|---------|---------|
| attention_ce_mean,cos               | 0.8936 | 0.8995 | 0.8969     | 0.8946  | 0.8977  |
| attention_ce_mean,gram              | 0.8916 | 0.8953 | 0.8925     | 0.8922  | 0.8925  |
| attention_ce_mean,hidden_mse        | 0.8946 | 0.8974 | 0.8993     | 0.8936  | 0.8968  |
| attention_ce_mean,key_relation      | 0.8956 | 0.8953 | 0.8957     | 0.8923  | 0.8925  |
| attention_ce_mean,mmd               | 0.8956 | 0.8952 | 0.8955     | 0.893   | 0.8922  |
| attention_ce_mean,pkd               | 0.8921 | 0.8989 | 0.899      | 0.8936  | 0.8952  |
| attention_ce_mean,query_relation    | 0.8951 | 0.8956 | 0.8947     | 0.8951  | 0.8957  |
| attention_ce_mean,value_relation    | 0.8915 | 0.8931 | 0.8907     | 0.8958  | 0.8932  |
| attention_mse_sum,cos               | 0.8916 | 0.8956 | 0.8966     | 0.8951  | 0.8966  |
| attention_mse_sum,gram              | 0.892  | 0.8964 | 0.894      | 0.8954  | 0.8952  |
| attention_mse_sum,hidden_mse        | 0.8923 | 0.896  | 0.8956     | 0.8953  | 0.8982  |
| attention_mse_sum,key_relation      | 0.8945 | 0.8897 | 0.8932     | 0.8922  | 0.8961  |
| attention_mse_sum,mmd               | 0.8943 | 0.8932 | 0.8933     | 0.8929  | 0.8966  |
| attention_mse_sum,pkd               | 0.8923 | 0.8978 | 0.8997     | 0.8921  | 0.8969  |
| attention_mse_sum,query_relation    | 0.8924 | 0.8916 | 0.8919     | 0.8918  | 0.8955  |
| attention_mse_sum,value_relation    | 0.8973 | 0.8919 | 0.8971     | 0.8912  | 0.8961  |
| cos,key_relation                     | 0.8937 | 0.8958 | 0.8972     | 0.8927  | 0.8987  |
| cos,query_relation                   | 0.8905 | 0.8968 | 0.899      | 0.8956  | 0.895  |
| cos,value_relation                   | 0.8922 | 0.8982 | 0.8969     | 0.8925  | 0.8937  |
| gram,key_relation                    | 0.8915 | 0.8949 | 0.8937     | 0.893   | 0.891   |
| gram,query_relation                  | 0.8926 | 0.8942 | 0.8962     | 0.8918  | 0.893   |
| gram,value_relation                  | 0.8938 | 0.8952 | 0.8935     | 0.8926  | 0.8928  |
| hidden_mse,key_relation              | 0.8938 | 0.8976 | 0.8989     | 0.8918  | 0.8929  |
| hidden_mse,query_relation            | 0.8941 | 0.8985 | 0.8991     | 0.8917  | 0.8937  |
| hidden_mse,value_relation            | 0.8917 | 0.8979 | 0.8965     | 0.8921  | 0.8933  |
| mmd,key_relation                     | 0.8922 | 0.8923 | 0.8927     | 0.8914  | 0.8959  |
| mmd,query_relation                   | 0.8924 | 0.8926 | 0.8921     | 0.8928  | 0.8932  |
| mmd,value_relation                   | 0.8915 | 0.8932 | 0.8945     | 0.893   | 0.8926  |
| pkd,key_relation                     | 0.8914 | 0.8982 | 0.8987     | 0.8917  | 0.896   |
| pkd,query_relation                   | 0.8914 | 0.8978 | 0.9002     | 0.892   | 0.8948  |
| pkd,value_relation                   | 0.8939 | 0.8979 | 0.9005     | 0.8949  | 0.8959  |
| bert-mini         | First | Last   | Dilatation | First-1 | Last-1 |
|-------------------|-------|--------|------------|---------|--------|
| attention_ce_mean.cos | 0.8885 | 0.8938 | 0.8918     | 0.8904  | 0.8919 |
| attention_ce_mean,gram | 0.8867 | 0.8912 | 0.8923     | 0.8866  | 0.8868 |
| attention_ce_mean,hidden_mse | 0.8879 | 0.8908 | 0.8917     | 0.8882  | 0.8913 |
| attention_ce_mean,key_relation | 0.8904 | 0.8867 | 0.8888     | 0.8901  | 0.8886 |
| attention_ce_mean,pkd | 0.8888 | 0.8867 | 0.8699     | 0.8869  | 0.8888 |
| attention_ce_mean,query_relation | 0.8886 | 0.8899 | 0.8959     | 0.8882  | 0.8893 |
| attention_mse_sum,cos | 0.8879 | 0.8908 | 0.887     | 0.8908  | 0.8873 |
| attention_mse_sum,gram | 0.8876 | 0.8897 | 0.8875     | 0.8903  | 0.8904 |
| attention_mse_sum,hidden_mse | 0.8841 | 0.8888 | 0.891     | 0.8906  | 0.8888 |
| attention_mse_sum,key_relation | 0.8862 | 0.8892 | 0.8859     | 0.8913  | 0.8909 |
| attention_mse_sum,pkd | 0.8884 | 0.8901 | 0.8893     | 0.892   | 0.8913 |
| attention_mse_sum,query_relation | 0.8884 | 0.8951 | 0.8949     | 0.8899  | 0.8932 |
| attention_mse_sum,value_relation | 0.8904 | 0.8881 | 0.8856     | 0.8893  | 0.8904 |
| cos,key_relation | 0.8867 | 0.892 | 0.8942     | 0.8909  | 0.8897 |
| cos,query_relation | 0.8848 | 0.892 | 0.8924     | 0.888  | 0.8891 |
| cos,value_relation | 0.8876 | 0.8898 | 0.8927     | 0.8892  | 0.8925 |
| gram,key_relation | 0.8904 | 0.8926 | 0.8906     | 0.8877  | 0.8878 |
| gram,query_relation | 0.8881 | 0.8927 | 0.8988     | 0.8864  | 0.888 |
| gram,value_relation | 0.8874 | 0.8911 | 0.8881     | 0.8858  | 0.8895 |
| hidden_mse,key_relation | 0.8909 | 0.8923 | 0.8952     | 0.887  | 0.8902 |
| hidden_mse,query_relation | 0.8886 | 0.8931 | 0.8952     | 0.8881  | 0.8891 |
| hidden_mse,value_relation | 0.8904 | 0.8922 | 0.8908     | 0.8882  | 0.8907 |
| mmd,key_relation | 0.8884 | 0.8899 | 0.8868     | 0.8906  | 0.8904 |
| mmd,query_relation | 0.8876 | 0.8904 | 0.8893     | 0.8874  | 0.8919 |
| mmd,value_relation | 0.8872 | 0.8892 | 0.8889     | 0.889  | 0.8877 |
| pkd,key_relation | 0.8876 | 0.8929 | 0.8943     | 0.8893  | 0.8915 |
| pkd,query_relation | 0.8856 | 0.8929 | 0.8949     | 0.8889  | 0.8917 |
| pkd,value_relation | 0.8847 | 0.8942 | 0.8956     | 0.8895  | 0.8926 |

| bert-tiny         | First | Last   | Dilatation | First-1 | Last-1 |
|-------------------|-------|--------|------------|---------|--------|
| attention_ce_mean.cos | 0.8645 | 0.8702 | 0.868     | 0.8688  | 0.8707 |
| attention_ce_mean,gram | 0.8701 | 0.8671 | 0.8695     | 0.8691  | 0.8695 |
| attention_ce_mean,hidden_mse | 0.8691 | 0.8661 | 0.8681     | 0.8694  | 0.8722 |
| attention_ce_mean,key_relation | 0.8724 | 0.8717 | 0.8718     | 0.8719  | 0.8694 |
| attention_ce_mean,mmd | 0.8707 | 0.8683 | 0.87     | 0.8712  | 0.8725 |
| attention_ce_mean,pkd | 0.8659 | 0.871 | 0.8706     | 0.8678  | 0.8733 |
| attention_ce_mean,query_relation | 0.869 | 0.8717 | 0.871     | 0.8676  | 0.8712 |
| attention_ce_mean,value_relation | 0.8689 | 0.8721 | 0.872     | 0.8707  | 0.8688 |
| attention_mse_sum,cos | 0.8653 | 0.864 | 0.8677     | 0.8704  | 0.8712 |
| attention_mse_sum,gram | 0.87 | 0.8697 | 0.8637     | 0.8721  | 0.8715 |
| attention_mse_sum,hidden_mse | 0.8678 | 0.8616 | 0.865     | 0.8699  | 0.8713 |
| attention_mse_sum,key_relation | 0.8699 | 0.869 | 0.8704     | 0.8718  | 0.8708 |
| attention_mse_sum,mmd | 0.8689 | 0.8637 | 0.8657     | 0.867  | 0.8705 |
| attention_mse_sum,pkd | 0.8686 | 0.8696 | 0.8724     | 0.869  | 0.8718 |
| attention_mse_sum,query_relation | 0.8692 | 0.8698 | 0.8675     | 0.8724  | 0.871 |
| attention_mse_sum,value_relation | 0.8708 | 0.8695 | 0.8684     | 0.8707  | 0.8665 |
| cos,key_relation | 0.8669 | 0.8691 | 0.8656     | 0.8679  | 0.8701 |
| cos,query_relation | 0.8659 | 0.8716 | 0.8648     | 0.8653  | 0.8698 |
| cos,value_relation | 0.8658 | 0.8682 | 0.8667     | 0.8711  | 0.8697 |
| gram,key_relation | 0.8692 | 0.8682 | 0.872     | 0.8695  | 0.869 |
| gram,query_relation | 0.8681 | 0.8689 | 0.8681     | 0.8726  | 0.8699 |
| gram,value_relation | 0.8696 | 0.8726 | 0.867     | 0.8683  | 0.87 |
| hidden_mse,key_relation | 0.8682 | 0.8673 | 0.8682     | 0.867  | 0.8705 |
Table 6 continued from previous page

| qnli          | bert-small       | First                      | Last                      | Dilatation               | First-1                      | Last-1                      |
|---------------|------------------|----------------------------|---------------------------|---------------------------|-----------------------------|-----------------------------|
| attention_ce_mean.cos | 0.8521 | 0.8504 | 0.8574 | 0.8686 | 0.868 | 0.8686 | 0.868 |
| attention_ce_mean.gram | 0.8677 | 0.857 | 0.864 | 0.8717 | 0.871 | 0.8711 | 0.871 |
| attention_ce_mean.hidden_mse | 0.8554 | 0.8506 | 0.855 | 0.8667 | 0.866 | 0.8662 | 0.8662 |
| attention_ce_mean.key_relation | 0.872 | 0.8717 | 0.8666 | 0.8724 | 0.872 | 0.87 |
| attention_ce_mean.mmd | 0.8722 | 0.8656 | 0.8713 | 0.8722 | 0.872 | 0.8724 | 0.8724 |
| attention_ce_mean.pkd | 0.8272 | 0.8647 | 0.8627 | 0.8711 | 0.871 | 0.8682 | 0.8682 |
| attention_ce_mean.query_relation | 0.8744 | 0.8699 | 0.8678 | 0.8719 | 0.871 | 0.8724 | 0.8724 |
| attention_ce_mean.value_relation | 0.8704 | 0.8735 | 0.8752 | 0.8744 | 0.874 | 0.8706 | 0.8706 |
| attention_mse_sum.cos | 0.849 | 0.8351 | 0.8534 | 0.8678 | 0.867 | 0.871 | 0.871 |
| attention_mse_sum.gram | 0.8602 | 0.8547 | 0.8708 | 0.8708 | 0.8708 | 0.8739 | 0.8739 |
| attention_mse_sum.hidden_mse | 0.8492 | 0.8444 | 0.8512 | 0.8667 | 0.866 | 0.8735 | 0.8735 |
| attention_mse_sum.key_relation | 0.8666 | 0.8651 | 0.8682 | 0.8715 | 0.871 | 0.8742 | 0.8742 |
| attention_mse_sum.mmd | 0.866 | 0.8592 | 0.8728 | 0.8726 | 0.872 | 0.8741 | 0.8741 |
| attention_mse_sum.pkd | 0.8395 | 0.858 | 0.8603 | 0.8728 | 0.872 | 0.87 | 0.87 |
| attention_mse_sum.query_relation | 0.8724 | 0.8622 | 0.8717 | 0.8711 | 0.871 | 0.8735 | 0.8735 |
| attention_mse_sum.value_relation | 0.8618 | 0.8678 | 0.8699 | 0.8722 | 0.872 | 0.8713 | 0.8713 |
| cos.key_relation | 0.8508 | 0.8473 | 0.8554 | 0.8688 | 0.868 | 0.8655 | 0.8655 |
| cos.query_relation | 0.8523 | 0.8528 | 0.8552 | 0.8678 | 0.867 | 0.8658 | 0.8658 |
| cos.value_relation | 0.8499 | 0.8504 | 0.8558 | 0.87 | 0.8622 | 0.8622 | 0.8622 |
| gram.key_relation | 0.871 | 0.857 | 0.8684 | 0.8722 | 0.872 | 0.8691 | 0.8691 |
| gram.query_relation | 0.8689 | 0.8602 | 0.8651 | 0.8722 | 0.872 | 0.871 | 0.871 |
| gram.value_relation | 0.8711 | 0.8569 | 0.8629 | 0.8761 | 0.876 | 0.8717 | 0.8717 |
| hidden_mse.key_relation | 0.8525 | 0.8495 | 0.8563 | 0.8667 | 0.866 | 0.8689 | 0.8689 |
| hidden_mse.query_relation | 0.8541 | 0.8526 | 0.8552 | 0.8673 | 0.867 | 0.8667 | 0.8667 |
| hidden_mse.value_relation | 0.8528 | 0.8514 | 0.8536 | 0.8664 | 0.866 | 0.8656 | 0.8656 |
| mmd.key_relation | 0.8702 | 0.8653 | 0.8726 | 0.875 | 0.8719 | 0.8719 | 0.8719 |
| mmd.query_relation | 0.8686 | 0.8653 | 0.8669 | 0.8717 | 0.872 | 0.8724 | 0.8724 |
| mmd.value_relation | 0.87 | 0.8673 | 0.8695 | 0.8719 | 0.871 | 0.8691 | 0.8691 |
| pkd.key_relation | 0.8312 | 0.8603 | 0.8629 | 0.8684 | 0.868 | 0.8671 | 0.8671 |
| pkd.query_relation | 0.8318 | 0.8633 | 0.8634 | 0.8713 | 0.871 | 0.8653 | 0.8653 |
| pkd.value_relation | 0.838 | 0.8625 | 0.8633 | 0.8706 | 0.870 | 0.8678 | 0.8678 |

| bert-mini       | First                      | Last                      | Dilatation               | First-1                      | Last-1                      |
|-----------------|----------------------------|---------------------------|---------------------------|-----------------------------|-----------------------------|
| attention_ce_mean.cos | 0.8349 | 0.8263 | 0.8351 | 0.844 | 0.844 | 0.8422 | 0.8422 |
| attention_ce_mean.gram | 0.8439 | 0.8433 | 0.8406 | 0.8442 | 0.844 | 0.8455 | 0.8455 |
| attention_ce_mean.hidden_mse | 0.8367 | 0.8358 | 0.838 | 0.8444 | 0.844 | 0.8431 | 0.8431 |
| attention_ce_mean.key_relation | 0.8448 | 0.8402 | 0.8426 | 0.8439 | 0.843 | 0.8446 | 0.8446 |
| attention_ce_mean.mmd | 0.8455 | 0.8455 | 0.8415 | 0.8437 | 0.843 | 0.8468 | 0.8468 |
| attention_ce_mean.pkd | 0.8179 | 0.8426 | 0.8431 | 0.8407 | 0.840 | 0.8479 | 0.8479 |
| attention_ce_mean.query_relation | 0.8486 | 0.8409 | 0.8411 | 0.8444 | 0.844 | 0.8428 | 0.8428 |
| attention_ce_mean.value_relation | 0.845 | 0.8413 | 0.8418 | 0.8433 | 0.843 | 0.8439 | 0.8439 |
| attention_mse_sum.cos | 0.8298 | 0.8188 | 0.8234 | 0.8409 | 0.840 | 0.8439 | 0.8439 |
| attention_mse_sum.gram | 0.8382 | 0.8342 | 0.8356 | 0.8481 | 0.848 | 0.845 | 0.845 |
| attention_mse_sum.hidden_mse | 0.8309 | 0.8272 | 0.8281 | 0.8462 | 0.846 | 0.8411 | 0.8411 |
| attention_mse_sum.key_relation | 0.8373 | 0.8391 | 0.8349 | 0.8435 | 0.843 | 0.8422 | 0.8422 |
| attention_mse_sum.mmd | 0.8396 | 0.8342 | 0.8331 | 0.8435 | 0.843 | 0.8417 | 0.8417 |
| attention_mse_sum.pkd | 0.8226 | 0.8384 | 0.842 | 0.84 | 0.8484 | 0.8484 | 0.8484 |
| Table 6 continued from previous page |
|--------------------------------------|
| attention_mse_sum,query_relation     | 0.8417 | 0.8367 | 0.8378 | 0.8435 | 0.8424 |
| attention_mse_sum,value_relation     | 0.8393 | 0.8365 | 0.8378 | 0.8435 | 0.8424 |
| cos,key_relation                     | 0.8371 | 0.8338 | 0.8384 | 0.8387 | 0.8431 |
| cos,query_relation                   | 0.8367 | 0.832 | 0.8365 | 0.8389 | 0.8415 |
| cos,value_relation                    | 0.8351 | 0.8321 | 0.8369 | 0.8387 | 0.8431 |
| gram,key_relation                     | 0.8424 | 0.8411 | 0.8413 | 0.8446 | 0.844 |
| gram,query_relation                   | 0.8435 | 0.8389 | 0.8393 | 0.8439 | 0.8422 |
| gram,value_relation                   | 0.8411 | 0.8406 | 0.8422 | 0.8457 | 0.845 |
| hidden_mse,key_relation               | 0.8413 | 0.8353 | 0.8406 | 0.8387 | 0.8418 |
| hidden_mse,query_relation             | 0.8411 | 0.8386 | 0.8437 | 0.8391 | 0.8411 |
| hidden_mse,value_relation             | 0.838 | 0.838 | 0.8411 | 0.8387 | 0.8413 |
| mmd,key_relation                      | 0.842 | 0.8 | 0.8439 | 0.8444 | 0.8442 |
| mmd,query_relation                    | 0.8437 | 0.8429 | 0.8415 | 0.8437 | 0.8435 |
| mmd,value_relation                    | 0.8422 | 0.8413 | 0.8413 | 0.8426 | 0.8437 |
| pkd,key_relation                      | 0.8212 | 0.8415 | 0.8415 | 0.8471 | 0.8477 |
| pkd,query_relation                    | 0.8234 | 0.8437 | 0.8428 | 0.8418 | 0.8462 |
| pkd,value_relation                    | 0.8213 | 0.8426 | 0.8415 | 0.8429 | 0.847 |

| bert-tiny First Last Dilatation First-1 Last-1 |
|-----------------------------------------------|
| attention_ce_mean.cos                        | 0.7829 | 0.7798 | 0.7809 | 0.7917 | 0.7899 |
| attention_ce_mean,gram                       | 0.7946 | 0.7889 | 0.7972 | 0.7954 | 0.7963 |
| attention_ce_mean,hidden_mse                 | 0.7838 | 0.7818 | 0.7825 | 0.7895 | 0.7939 |
| attention_ce_mean,key_relation               | 0.7966 | 0.7959 | 0.7975 | 0.7964 | 0.7974 |
| attention_ce_mean,mmd                        | 0.7981 | 0.7924 | 0.7959 | 0.7959 | 0.7963 |
| attention_ce_mean,pkd                        | 0.7778 | 0.7911 | 0.7893 | 0.7866 | 0.7913 |
| attention_mse_sum,cos                        | 0.7822 | 0.7811 | 0.7759 | 0.7853 | 0.7878 |
| attention_mse_sum,gram                       | 0.7959 | 0.7921 | 0.7939 | 0.793 | 0.7952 |
| attention_mse_sum,hidden_mse                 | 0.7866 | 0.7833 | 0.7827 | 0.7866 | 0.791 |
| attention_mse_sum,key_relation               | 0.7943 | 0.7966 | 0.7952 | 0.7937 | 0.7946 |
| attention_mse_sum,mmd                        | 0.7974 | 0.791 | 0.7955 | 0.7941 | 0.7974 |
| attention_mse_sum,pkd                        | 0.7747 | 0.7932 | 0.788 | 0.7911 | 0.7932 |
| attention_mse_sum,query_relation             | 0.7957 | 0.7955 | 0.7935 | 0.7948 | 0.7943 |
| attention_mse_sum,value_relation             | 0.7913 | 0.7934 | 0.7922 | 0.7965 | 0.7952 |
| cos,key_relation                              | 0.7844 | 0.782 | 0.7822 | 0.7908 | 0.7886 |
| cos,query_relation                           | 0.7844 | 0.7838 | 0.782 | 0.793 | 0.7871 |
| cos,value_relation                           | 0.7873 | 0.7814 | 0.7818 | 0.7886 | 0.7878 |
| gram,key_relation                             | 0.7983 | 0.7902 | 0.7957 | 0.7966 | 0.7957 |
| gram,query_relation                           | 0.797 | 0.7924 | 0.7964 | 0.8005 | 0.7952 |
| gram,value_relation                           | 0.7957 | 0.7884 | 0.7922 | 0.7966 | 0.7932 |
| hidden_mse,key_relation                      | 0.7895 | 0.7902 | 0.7913 | 0.7911 | 0.7917 |
| hidden_mse,query_relation                    | 0.7906 | 0.7849 | 0.7827 | 0.7937 | 0.7933 |
| hidden_mse,value_relation                    | 0.786 | 0.7864 | 0.7908 | 0.7893 | 0.79 |
| mmd,key_relation                              | 0.7979 | 0.7904 | 0.7977 | 0.7964 | 0.7972 |
| mmd,query_relation                           | 0.797 | 0.7944 | 0.7977 | 0.7977 | 0.7946 |
| mmd,value_relation                           | 0.7941 | 0.7932 | 0.7933 | 0.7948 | 0.795 |
| pkd,key_relation                              | 0.7838 | 0.7933 | 0.7911 | 0.7933 | 0.7919 |
| pkd,query_relation                           | 0.7849 | 0.7917 | 0.7904 | 0.7888 | 0.7922 |
| pkd,value_relation                           | 0.7811 | 0.7917 | 0.79 | 0.7926 | 0.7915 |

| bert-small First Last Dilatation First-1 Last-1 |
|-----------------------------------------------|
| attention_ce_mean.cos                        | 0.5704 | 0.556 | 0.5523 | 0.6606 | 0.6534 |
| attention_ce_mean,gram                       | 0.6715 | 0.6715 | 0.6751 | 0.6895 | 0.6895 |
| attention_ce_mean,hidden_mse                 | 0.5596 | 0.556 | 0.5596 | 0.6534 | 0.6354 |
| attention_ce_mean,key_relation               | 0.6534 | 0.6462 | 0.6498 | 0.6751 | 0.6534 |
| attention_ce_mean,mmd                        | 0.6643 | 0.657 | 0.6534 | 0.6859 | 0.6751 |

| rte First Last Dilatation First-1 Last-1 |
|----------------------------------------|
| attention_ce_mean.cos                  | 0.5704 | 0.556 | 0.5523 | 0.6606 | 0.6534 |
| attention_ce_mean,gram                 | 0.6715 | 0.6715 | 0.6751 | 0.6895 | 0.6895 |
| attention_ce_mean,hidden_mse           | 0.5596 | 0.556 | 0.5596 | 0.6534 | 0.6354 |
| attention_ce_mean,key_relation         | 0.6534 | 0.6462 | 0.6498 | 0.6751 | 0.6534 |
| attention_ce_mean,mmd                  | 0.6643 | 0.657 | 0.6534 | 0.6859 | 0.6751 |
Table 6 continued from previous page

| Term                        | bert-mini | First     | Last      | Dilatation | First-1 | Last-1 |
|-----------------------------|-----------|-----------|-----------|------------|---------|--------|
| attention_ce_mean, cos      |           | 0.5884    | 0.5596    | 0.5704     | 0.6679  | 0.6065 |
| attention_ce_mean, gram     |           | 0.6779    | 0.6534    | 0.6498     | 0.6787  | 0.6751 |
| attention_ce_mean, hidden_mse|         | 0.5993    | 0.556     | 0.5596     | 0.657   | 0.6498 |
| attention_ce_mean, key_relation|     | 0.6462    | 0.6209    | 0.6282     | 0.6787  | 0.639  |
| attention_ce_mean, mmd       |           | 0.6462    | 0.6715    | 0.657      | 0.6751  | 0.6606 |
| attention_ce_mean, pkd       |           | 0.5596    | 0.5884    | 0.6029     | 0.6534  | 0.6209 |
| attention_ce_mean, query_relation|     | 0.6426    | 0.6209    | 0.6354     | 0.6787  | 0.639  |
| attention_ce_mean, value_relation| | 0.6426    | 0.6498    | 0.6426     | 0.6895  | 0.6606 |
| attention_mse_sum, cos       |           | 0.5921    | 0.5415    | 0.5415     | 0.6426  | 0.6245 |
| attention_mse_sum, gram      |           | 0.6245    | 0.6209    | 0.6245     | 0.6643  | 0.6643 |
| attention_mse_sum, hidden_mse|         | 0.5632    | 0.5451    | 0.5848     | 0.6498  | 0.6426 |
| attention_mse_sum, key_relation|     | 0.6282    | 0.6065    | 0.5812     | 0.6787  | 0.6462 |
| attention_mse_sum, mmd       |           | 0.6462    | 0.6173    | 0.6137     | 0.6643  | 0.6534 |
| attention_mse_sum, pkd       |           | 0.5523    | 0.6245    | 0.6137     | 0.6426  | 0.6282 |
| attention_mse_sum, query_relation|     | 0.6426    | 0.6137    | 0.5812     | 0.6751  | 0.6534 |
| attention_mse_sum, value_relation| | 0.6318    | 0.6282    | 0.6209     | 0.6751  | 0.6534 |
| cos, key_relation            |           | 0.5921    | 0.5596    | 0.5812     | 0.657   | 0.6173 |
| cos, query_relation          |           | 0.5921    | 0.5523    | 0.5668     | 0.6498  | 0.6101 |
| cos, value_relation          |           | 0.5848    | 0.574     | 0.5704     | 0.6462  | 0.6101 |
| gram, key_relation           |           | 0.657     | 0.6245    | 0.6318     | 0.6787  | 0.639  |
| gram, query_relation         |           | 0.6679    | 0.639     | 0.6462     | 0.6787  | 0.6209 |
| gram, value_relation         |           | 0.6498    | 0.6426    | 0.6498     | 0.6787  | 0.6643 |
| hidden_mse, key_relation     |           | 0.5884    | 0.5632    | 0.5776     | 0.6643  | 0.6426 |
| hidden_mse, query_relation   |           | 0.5921    | 0.5596    | 0.5848     | 0.6498  | 0.6426 |
| hidden_mse, value_relation   |           | 0.5884    | 0.5884    | 0.5921     | 0.6715  | 0.6462 |
| mmd, key_relation            |           | 0.6462    | 0.6173    | 0.5812     | 0.6751  | 0.6354 |
| mmd, query_relation          |           | 0.6426    | 0.6137    | 0.5812     | 0.6751  | 0.6354 |
| mmd, value_relation          |           | 0.6318    | 0.6282    | 0.6209     | 0.6751  | 0.6534 |
| pkd, key_relation            |           | 0.5921    | 0.5596    | 0.5812     | 0.657   | 0.6173 |
| pkd, query_relation          |           | 0.5921    | 0.5523    | 0.5668     | 0.6498  | 0.6101 |
| pkd, value_relation          |           | 0.5848    | 0.574     | 0.5704     | 0.6462  | 0.6101 |
| bert-mini                   |           | 0.5921    | 0.5596    | 0.5812     | 0.657   | 0.6173 |
|                          |           | 0.5921    | 0.5523    | 0.5668     | 0.6498  | 0.6101 |
|                          |           | 0.5848    | 0.574     | 0.5704     | 0.6462  | 0.6101 |
|                          |           | 0.657     | 0.6245    | 0.6318     | 0.6787  | 0.639  |
|                          |           | 0.6498    | 0.6426    | 0.6498     | 0.6787  | 0.6643 |
|                          |           | 0.5884    | 0.5632    | 0.5776     | 0.6606  | 0.6354 |
|                          |           | 0.5921    | 0.5596    | 0.5848     | 0.6498  | 0.6101 |
|                          |           | 0.5884    | 0.5884    | 0.5921     | 0.6715  | 0.6462 |
|                          |           | 0.6534    | 0.639     | 0.6462     | 0.6787  | 0.6318 |
|                          |           | 0.657     | 0.6462    | 0.6498     | 0.6715  | 0.6354 |
|                          |           | 0.6354    | 0.6534    | 0.657      | 0.6751  | 0.6643 |
|                          |           | 0.556     | 0.6029    | 0.639      | 0.6498  | 0.6354 |
|                          |           | 0.556     | 0.5993    | 0.6137     | 0.6534  | 0.6318 |
Table 6 continued from previous page

|                            | bert-tiny | bert-small |
|---------------------------|-----------|------------|
|                            | First     | Last       | Dilatation | First-1 | Last-1 |
| pkd, value_relation       | 0.556     | 0.610      | 0.606      | 0.6534  | 0.6209 |
| attention_ce_mean, cos    | 0.6282    | 0.5776     | 0.6173     | 0.6101  | 0.5776 |
| attention_ce_mean, gram   | 0.6209    | 0.6137     | 0.6209     | 0.6101  | 0.6173 |
| attention_ce_mean, hidden_mse | 0.5921   | 0.5921     | 0.5921     | 0.5993  | 0.5921 |
| attention_ce_mean, key_relation | 0.6173 | 0.6173     | 0.6137     | 0.6173  | 0.6065 |
| attention_ce_mean, mmd    | 0.6354    | 0.5993     | 0.6173     | 0.6282  | 0.6209 |
| attention_ce_mean, pkd    | 0.5523    | 0.5921     | 0.5993     | 0.574   | 0.5957 |
| attention_ce_mean, query_relation | 0.6137 | 0.5957     | 0.5921     | 0.6173  | 0.5957 |
| attention_ce_mean, value_relation | 0.6209 | 0.6245     | 0.6318     | 0.6245  | 0.6245 |
| attention_mse_sum, cos    | 0.5993    | 0.5596     | 0.5848     | 0.6209  | 0.5848 |
| attention_mse_sum, gram   | 0.6101    | 0.6029     | 0.6245     | 0.6173  | 0.6065 |
| attention_mse_sum, hidden_mse | 0.5993 | 0.5884     | 0.5921     | 0.6137  | 0.5921 |
| attention_mse_sum, key_relation | 0.6137 | 0.6029     | 0.6101     | 0.6245  | 0.6065 |
| attention_mse_sum, mmd    | 0.6354    | 0.6245     | 0.6245     | 0.6282  | 0.6065 |
| attention_mse_sum, pkd    | 0.556     | 0.5921     | 0.5884     | 0.5812  | 0.5957 |
| attention_mse_sum, query_relation | 0.6173 | 0.6029     | 0.6101     | 0.6245  | 0.6029 |
| attention_mse_sum, value_relation | 0.6101 | 0.6029     | 0.6245     | 0.6137  | 0.6173 |
| cos, key_relation         | 0.6173    | 0.5957     | 0.6173     | 0.6137  | 0.5812 |
| cos, query_relation       | 0.6137    | 0.5848     | 0.6029     | 0.6173  | 0.5848 |
| cos, value_relation       | 0.6173    | 0.5884     | 0.5884     | 0.6101  | 0.5776 |
| gram, key_relation        | 0.6209    | 0.6245     | 0.6209     | 0.6282  | 0.5993 |
| gram, query_relation      | 0.6029    | 0.6029     | 0.5993     | 0.6245  | 0.5957 |
| gram, value_relation      | 0.6245    | 0.639      | 0.6245     | 0.6209  | 0.6245 |
| hidden_mse, key_relation  | 0.5957    | 0.5812     | 0.5884     | 0.6137  | 0.5848 |
| hidden_mse, query_relation | 0.5921    | 0.5848    | 0.5921     | 0.6065  | 0.5812 |
| hidden_mse, value_relation | 0.6029   | 0.5993     | 0.5993     | 0.6137  | 0.5921 |
| mmd, key_relation         | 0.6318    | 0.6101     | 0.6245     | 0.6318  | 0.6137 |
| mmd, query_relation       | 0.6318    | 0.6065     | 0.6101     | 0.6318  | 0.6065 |
| mmd, value_relation       | 0.6462    | 0.6101     | 0.6209     | 0.6354  | 0.6245 |
| pkd, key_relation         | 0.5523    | 0.6029     | 0.6029     | 0.5812  | 0.5921 |
| pkd, query_relation       | 0.5451    | 0.5993     | 0.5957     | 0.5848  | 0.5993 |
| pkd, value_relation       | 0.5487    | 0.6065     | 0.5993     | 0.5704  | 0.5957 |

**cola**

|                            | bert-small | bert-tiny |
|---------------------------|-----------|-----------|
|                            | First     | Last       | Dilatation | First-1 | Last-1 |
| attention_ce_mean, cos    | 0.7181    | 0.7018     | 0.7133     | 0.7747  | 0.7728 |
| attention_ce_mean, gram   | 0.7747    | 0.7804     | 0.7785     | 0.7747  | 0.7795 |
| attention_ce_mean, hidden_mse | 0.7565 | 0.72      | 0.7402     | 0.7728  | 0.768 |
| attention_ce_mean, key_relation | 0.7747 | 0.7613     | 0.767      | 0.7833  | 0.7728 |
| attention_ce_mean, mmd    | 0.7804    | 0.7709     | 0.7689     | 0.7776  | 0.7747 |
| attention_ce_mean, pkd    | 0.7143    | 0.7469     | 0.7296     | 0.7756  | 0.768 |
| attention_ce_mean, query_relation | 0.7747 | 0.7728     | 0.768      | 0.7747  | 0.7824 |
| attention_ce_mean, value_relation | 0.7785 | 0.7814     | 0.7814     | 0.7776  | 0.7756 |
| attention_mse_sum, cos    | 0.7277    | 0.7124     | 0.7306     | 0.7766  | 0.7689 |
| attention_mse_sum, gram   | 0.7737    | 0.767      | 0.7689     | 0.7766  | 0.7776 |
| attention_mse_sum, hidden_mse | 0.7651 | 0.7287     | 0.7392     | 0.7737  | 0.7689 |
| attention_mse_sum, key_relation | 0.7766 | 0.7641     | 0.7709     | 0.7747  | 0.7689 |
| attention_mse_sum, mmd    | 0.7766    | 0.7718     | 0.7718     | 0.7804  | 0.7737 |
| attention_mse_sum, pkd    | 0.7181    | 0.7344     | 0.743      | 0.7728  | 0.7709 |
| attention_mse_sum, query_relation | 0.7766 | 0.768      | 0.7737     | 0.7728  | 0.7766 |
| attention_mse_sum, value_relation | 0.7776 | 0.7574     | 0.7718     | 0.7766  | 0.7709 |
| cos, key_relation         | 0.7296    | 0.7085     | 0.721      | 0.7737  | 0.7641 |
| cos, query_relation       | 0.7354    | 0.7152     | 0.7191     | 0.7737  | 0.767 |
| cos, value_relation       | 0.7191    | 0.7028     | 0.7162     | 0.7747  | 0.767 |
| gram, key_relation        | 0.7747    | 0.7651     | 0.768      | 0.7756  | 0.7728 |
| gram, query_relation      | 0.7737    | 0.7718     | 0.7661     | 0.7766  | 0.7709 |
| Table 6 continued from previous page |
|--------------------------------------|
| gram, value_relation | 0.7728 | 0.7766 | 0.7766 | 0.7747 | 0.7795 |
| hidden_mse, key_relation | 0.7584 | 0.72 | 0.7421 | 0.7766 | 0.767 |
| hidden_mse, query_relation | 0.7661 | 0.7296 | 0.743 | 0.7747 | 0.7718 |
| hidden_mse, value_relation | 0.7536 | 0.7152 | 0.7373 | 0.7718 | 0.7709 |
| mmd, key_relation | 0.7824 | 0.767 | 0.7709 | 0.7804 | 0.7709 |
| mmd, query_relation | 0.7756 | 0.7709 | 0.7814 | 0.7737 | 0.7737 |
| mmd, value_relation | 0.7728 | 0.7737 | 0.7766 | 0.7756 | 0.7804 |
| pkd, key_relation | 0.7076 | 0.7277 | 0.7306 | 0.7766 | 0.7661 |
| pkd, query_relation | 0.7114 | 0.7421 | 0.7315 | 0.7766 | 0.7689 |
| pkd, value_relation | 0.7306 | 0.7411 | 0.7344 | 0.7728 | 0.7728 |

| bert-mini | First | Last | Dilatation | First-1 | Last-1 |
|-----------|-------|------|------------|--------|-------|
| attention_ce_mean.cos | 0.698 | 0.6922 | 0.6951 | 0.7536 | 0.7239 |
| attention_ce_mean.gram | 0.7459 | 0.7469 | 0.7488 | 0.7517 | 0.7488 |
| attention_ce_mean.hidden_mse | 0.7066 | 0.7037 | 0.7105 | 0.7469 | 0.743 |
| attention_ce_mean.key_relation | 0.7421 | 0.7392 | 0.7536 | 0.7402 | 0.7383 |
| attention_ce_mean.mmd | 0.7421 | 0.7335 | 0.7392 | 0.7411 | 0.7507 |
| attention_ce_mean.pkd | 0.6913 | 0.697 | 0.6922 | 0.7469 | 0.7373 |
| attention_ce_mean.query_relation | 0.7478 | 0.7478 | 0.7565 | 0.744 | 0.745 |
| attention_mse_sum.cos | 0.6913 | 0.6913 | 0.6932 | 0.7498 | 0.7267 |
| attention_mse_sum.gram | 0.7383 | 0.7028 | 0.697 | 0.7459 | 0.7421 |
| attention_mse_sum.hidden_mse | 0.698 | 0.6942 | 0.6922 | 0.7507 | 0.7402 |
| attention_mse_sum.key_relation | 0.7421 | 0.721 | 0.7066 | 0.7411 | 0.7402 |
| attention_mse_sum.mmd | 0.7306 | 0.7018 | 0.7085 | 0.745 | 0.745 |
| attention_mse_sum.pkd | 0.6913 | 0.6942 | 0.6922 | 0.7469 | 0.7354 |
| attention_mse_sum.query_relation | 0.7469 | 0.7191 | 0.7066 | 0.7421 | 0.7411 |
| attention_mse_sum.value_relation | 0.7459 | 0.7315 | 0.72 | 0.7402 | 0.7383 |
| cos, key_relation | 0.6932 | 0.6989 | 0.6999 | 0.7488 | 0.7172 |
| cos, query_relation | 0.6961 | 0.6942 | 0.6913 | 0.7584 | 0.7267 |
| cos, value relation | 0.6989 | 0.6932 | 0.698 | 0.743 | 0.7277 |
| gram, key_relation | 0.745 | 0.7392 | 0.7478 | 0.745 | 0.7507 |
| gram, query_relation | 0.745 | 0.7267 | 0.7383 | 0.745 | 0.7469 |
| gram, value_relation | 0.7565 | 0.7421 | 0.7469 | 0.7488 | 0.743 |
| hidden_mse.key_relation | 0.7133 | 0.7018 | 0.7037 | 0.7478 | 0.7402 |
| hidden_mse.query_relation | 0.7066 | 0.697 | 0.6989 | 0.7478 | 0.7383 |
| hidden_mse.value_relation | 0.7009 | 0.7037 | 0.7028 | 0.7507 | 0.7402 |
| mmd.key_relation | 0.7478 | 0.7363 | 0.7383 | 0.7421 | 0.7459 |
| mmd.query_relation | 0.7507 | 0.7143 | 0.7335 | 0.743 | 0.7469 |
| mmd.value_relation | 0.7507 | 0.7277 | 0.7296 | 0.7411 | 0.7402 |
| pkd, key_relation | 0.6913 | 0.7018 | 0.697 | 0.7488 | 0.7315 |
| pkd, query_relation | 0.6932 | 0.6999 | 0.697 | 0.7402 | 0.7344 |
| pkd, value_relation | 0.697 | 0.7009 | 0.7057 | 0.743 | 0.7383 |

| bert-tiny | First | Last | Dilatation | First-1 | Last-1 |
|-----------|-------|------|------------|--------|-------|
| attention_ce_mean.cos | 0.6913 | 0.6913 | 0.6913 | 0.6913 | 0.6913 |
| attention_ce_mean.gram | 0.6932 | 0.6961 | 0.6932 | 0.6913 | 0.697 |
| attention_ce_mean.hidden_mse | 0.6913 | 0.6913 | 0.6913 | 0.6913 | 0.6913 |
| attention_ce_mean.key_relation | 0.6922 | 0.6913 | 0.6951 | 0.6989 | 0.6913 |
| attention_ce_mean.mmd | 0.697 | 0.6922 | 0.6922 | 0.6913 | 0.6942 |
| attention_ce_mean.pkd | 0.6913 | 0.6913 | 0.6913 | 0.6913 | 0.6913 |
| attention_ce_mean.query_relation | 0.6989 | 0.698 | 0.6951 | 0.6999 | 0.6942 |
| attention_ce_mean.value_relation | 0.6913 | 0.698 | 0.6942 | 0.697 | 0.6913 |
| attention_mse_sum.cos | 0.6942 | 0.6913 | 0.6913 | 0.6913 | 0.6913 |
| attention_mse_sum.gram | 0.6913 | 0.6913 | 0.6922 | 0.6913 | 0.6913 |
| attention_mse_sum.hidden_mse | 0.6913 | 0.6932 | 0.6913 | 0.6922 | 0.6913 |
| attention_mse_sum.key_relation | 0.6922 | 0.6913 | 0.6913 | 0.6961 | 0.6913 |
| attention_mse_sum.mmd | 0.6922 | 0.6913 | 0.6913 | 0.6913 | 0.6913 |
### Table 6 continued from previous page

|                         | bert-small                      | bert-mini                      |
|-------------------------|--------------------------------|--------------------------------|
|                         | First | Last  | Dilatation | First-1 | Last-1 |
| attention_ce_mean.cos   | 0.8722 | 0.8715 | 0.8712     | 0.874   | 0.8721 |
| attention_ce_mean,gram  | 0.8732 | 0.8717 | 0.8731     | 0.8724  | 0.8745 |
| attention_ce_mean,hidden_mse | 0.8723 | 0.8703 | 0.8695     | 0.8736  | 0.872  |
| attention_ce_mean,key_relation | 0.8748 | 0.8724 | 0.8721     | 0.8741  | 0.8735 |
| attention_ce_mean,mmd   | 0.8721 | 0.8748 | 0.8737     | 0.8726  | 0.8736 |
| attention_ce_mean,pkd   | 0.8665 | 0.869   | 0.8694     | 0.8727  | 0.8726 |
| attention_ce_mean,query_relation | 0.8722 | 0.8729 | 0.8731     | 0.8738  | 0.8721 |
| attention_ce_mean,value_relation | 0.8735 | 0.872 | 0.8737    | 0.8717  | 0.8735 |
| attention_mse_sum,cos   | 0.87 | 0.8686 | 0.8695     | 0.8748  | 0.8739 |
| attention_mse_sum,gram  | 0.8719 | 0.8745 | 0.8744     | 0.8745  | 0.8749 |
| attention_mse_sum,hidden_mse | 0.8684 | 0.8703 | 0.8706     | 0.8724  | 0.8728 |
| attention_mse_sum,key_relation | 0.8734 | 0.8732 | 0.875      | 0.8731  | 0.8737 |
| attention_mse_sum,mmd   | 0.874 | 0.8726 | 0.872      | 0.873   | 0.8735 |
| attention_mse_sum,pkd   | 0.8666 | 0.8686 | 0.8685     | 0.8706  | 0.8738 |
| attention_mse_sum,query_relation | 0.8722 | 0.8724 | 0.8734     | 0.8727  | 0.8749 |
| attention_mse_sum,value_relation | 0.8727 | 0.8727 | 0.8748    | 0.8732  | 0.8753 |
| cos,key_relation        | 0.8727 | 0.87  | 0.87      | 0.8729  | 0.8708 |
| cos,query_relation      | 0.8721 | 0.8697 | 0.8698     | 0.873   | 0.8718 |
| cos,value_relation      | 0.8722 | 0.8704 | 0.8685     | 0.8751  | 0.8728 |
| gram,key_relation       | 0.8732 | 0.8722 | 0.8712     | 0.8725  | 0.8754 |
| gram,query_relation     | 0.8728 | 0.8732 | 0.8723     | 0.8729  | 0.8724 |
| gram,value_relation     | 0.8727 | 0.8717 | 0.8729     | 0.873   | 0.8721 |
| hidden_mse,key_relation | 0.8716 | 0.8697 | 0.8695     | 0.8736  | 0.8717 |
| hidden_mse,query_relation | 0.8724 | 0.8718 | 0.8712     | 0.8751  | 0.8723 |
| hidden_mse,value_relation | 0.8707 | 0.8711 | 0.8707     | 0.8712  | 0.8726 |
| mmd,key_relation        | 0.8742 | 0.8729 | 0.8736     | 0.8753  | 0.8745 |
| mmd,query_relation      | 0.8721 | 0.8717 | 0.8723     | 0.8756  | 0.8734 |
| mmd,value_relation      | 0.8733 | 0.8752 | 0.8731     | 0.8731  | 0.8741 |
| pkd,key_relation        | 0.8672 | 0.8693 | 0.8677     | 0.8716  | 0.8733 |
| pkd,query_relation      | 0.8678 | 0.8708 | 0.8684     | 0.8719  | 0.8742 |
| pkd,value_relation      | 0.8664 | 0.8694 | 0.8689     | 0.8708  | 0.8742 |

|                         | First       | Last       | Dilatation | First-1  | Last-1 |
|-------------------------|-------------|------------|------------|----------|--------|
| attention_ce_mean.cos   | 0.8648      | 0.8541     | 0.8566     | 0.8661   | 0.8661 |
| attention_ce_mean,gram  | 0.865       | 0.8646     | 0.8635     | 0.8624   | 0.8636 |
| attention_ce_mean,hidden_mse | 0.8667 | 0.8605 | 0.8625     | 0.8648   | 0.8638 |
| attention_ce_mean,key_relation | 0.8659 | 0.866 | 0.8652     | 0.8644   | 0.8643 |

32
| Metric                  | bert-tiny | First | Last | Dilatation | First-1 | Last-1 |
|-------------------------|-----------|-------|------|------------|---------|--------|
| attention_ce_mean,cos   | 0.8178    | 0.8189| 0.8154| 0.8175     | 0.8166  |
| attention_ce_mean,gram  | 0.8165    | 0.814 | 0.816 | 0.8168     | 0.8164  |
| attention_ce_mean,key_relation | 0.8153 | 0.8157 | 0.8157 | 0.8167 | 0.8158 |
| attention_ce_mean,mmd   | 0.817    | 0.8158 | 0.8148 | 0.8155 | 0.8156 |
| attention_ce_mean,pkd   | 0.8156   | 0.8155 | 0.8155 | 0.8149 | 0.8155 |
| attention_ce_mean,value_relation | 0.8155 | 0.8153 | 0.8153 | 0.8155 | 0.8157 |
| attention_mse_sum,cos   | 0.8169    | 0.8155 | 0.8174 | 0.8171 | 0.8154 |
| attention_mse_sum,gram  | 0.8172    | 0.8146 | 0.8173 | 0.8169 | 0.8161 |
| attention_mse_sum,key_relation | 0.8173 | 0.8147 | 0.8142 | 0.8144 | 0.8156 |
| attention_mse_sum,mmd   | 0.8155   | 0.8154 | 0.8155 | 0.8154 | 0.8156 |
| attention_mse_sum,pkd   | 0.8154   | 0.814 | 0.8152 | 0.815 | 0.8159 |
| attention_mse_sum,value_relation | 0.817 | 0.8152 | 0.8151 | 0.817 | 0.8163 |
| cos,key_relation        | 0.8169    | 0.816 | 0.8167 | 0.8167 | 0.8149 |
| cos,query_relation      | 0.8174    | 0.8156 | 0.8152 | 0.8175 | 0.8169 |
| cos,value_relation      | 0.8178    | 0.8158 | 0.818 | 0.8173 | 0.8142 |
| gram,key_relation       | 0.8167    | 0.8171 | 0.8151 | 0.8142 | 0.8154 |
| gram,query_relation     | 0.8154    | 0.8171 | 0.817 | 0.8168 | 0.817 |
| gram,value_relation     | 0.8158    | 0.816 | 0.8149 | 0.8153 | 0.8156 |
| hidden_mse,key_relation | 0.8154   | 0.8142 | 0.8164 | 0.8141 | 0.8143 |
| hidden_mse,query_relation | 0.8173 | 0.8143 | 0.8138 | 0.8167 | 0.8144 |
| hidden_mse,value_relation | 0.8138 | 0.8149 | 0.8148 | 0.8142 | 0.8148 |
| mmd,key_relation        | 0.8164    | 0.8146 | 0.815 | 0.8147 | 0.8156 |
| mmd,query_relation      | 0.8155    | 0.8145 | 0.8149 | 0.8152 | 0.8172 |
| mmd,value_relation      | 0.8162    | 0.816 | 0.8142 | 0.8152 | 0.8152 |
| pkd,key_relation        | 0.8129    | 0.8144 | 0.8128 | 0.8145 | 0.8164 |

Table 6 continued from previous page
| Metric                          | bert-small  | bert-mini  |
|------|----------------|-------------|
|      | First          | Last        | Dilatation | First-1 | Last-1 |
|      |                |             |            |         |        |
| pkd,query_relation            | 0.8163      | 0.8147      | 0.8132     | 0.8145  | 0.8169 |
| pkd,value_relation            | 0.814       | 0.8141      | 0.8101     | 0.8144  | 0.8155 |
| mnli_mismatched               |             |             |            |         |        |
| attention_ce_mean.cos         | 0.7888      | 0.8099      | 0.8111     | 0.7968  | 0.8063 |
| attention_ce_mean.gram        | 0.801       | 0.8037      | 0.8027     | 0.8017  | 0.8012 |
| attention_ce_mean.hidden_mse  | 0.79        | 0.8067      | 0.8094     | 0.7984  | 0.8018 |
| attention_ce_mean,key_relation| 0.7988      | 0.7952      | 0.8034     | 0.803   | 0.7987 |
| attention_ce_mean.mmd         | 0.7981      | 0.7992      | 0.804      | 0.7986  | 0.8016 |
| attention_ce_mean.pkd         | 0.7855      | 0.8128      | 0.8164     | 0.8009  | 0.8063 |
| attention_ce_mean.query_relation | 0.8001  | 0.7988      | 0.7986     | 0.7988  | 0.8004 |
| attention_ce_mean,value_relation | 0.7971 | 0.7977      | 0.8003     | 0.7986  | 0.8007 |
| attention_mse_sum.cos         | 0.7884      | 0.8082      | 0.8131     | 0.7972  | 0.8066 |
| attention_mse_sum.gram        | 0.7977      | 0.8039      | 0.8013     | 0.799   | 0.8013 |
| attention_mse_sum.hidden_mse  | 0.7866      | 0.8056      | 0.8086     | 0.7981  | 0.8027 |
| attention_mse_sum,key_relation| 0.7986      | 0.7924      | 0.7989     | 0.8032  | 0.8002 |
| attention_mse_sum.mmd         | 0.7964      | 0.7974      | 0.7986     | 0.8036  | 0.8007 |
| attention_mse_sum.pkd         | 0.7874      | 0.8094      | 0.8155     | 0.8     | 0.8076 |
| attention_mse_sum.query_relation | 0.7988  | 0.7914      | 0.7962     | 0.8001  | 0.8003 |
| attention_mse_sum,value_relation | 0.7987 | 0.7924      | 0.7994     | 0.8005  | 0.8009 |
| cos,key_relation               | 0.7917      | 0.8118      | 0.8123     | 0.7981  | 0.8038 |
| cos,query_relation             | 0.7922      | 0.8111      | 0.8119     | 0.7977  | 0.8044 |
| cos,value_relation              | 0.7892      | 0.8108      | 0.8111     | 0.7988  | 0.8059 |
| gram,key_relation               | 0.8007      | 0.7961      | 0.8024     | 0.8019  | 0.8011 |
| gram,query_relation             | 0.7972      | 0.7974      | 0.7995     | 0.8     | 0.7991 |
| gram,value_relation             | 0.799       | 0.8037      | 0.8002     | 0.8017  | 0.7986 |
| hidden_mse,key_relation         | 0.7932      | 0.8053      | 0.8056     | 0.7999  | 0.8007 |
| hidden_mse,query_relation       | 0.7935      | 0.8095      | 0.8069     | 0.8011  | 0.8009 |
| hidden_mse,value_relation       | 0.7909      | 0.8067      | 0.8082     | 0.797   | 0.8016 |
| mmd,key_relation                | 0.7971      | 0.7997      | 0.7976     | 0.8022  | 0.7991 |
| mmd,query_relation              | 0.801       | 0.7975      | 0.7987     | 0.8036  | 0.8027 |
| mmd,value_relation              | 0.7978      | 0.7986      | 0.7995     | 0.803   | 0.8005 |
| pkd,key_relation                | 0.7869      | 0.8104      | 0.8116     | 0.8     | 0.8037 |
| pkd,query_relation              | 0.7888      | 0.8081      | 0.8115     | 0.8032  | 0.8009 |
| pkd,value_relation              | 0.7847      | 0.8103      | 0.8154     | 0.8002  | 0.8046 |

Table 6 continued from previous page
| Table 6 continued from previous page |
|--------------------------------------|
| gram,query_relation                  | 0.7836 0.7852 0.7857 0.7836 0.7839 |
| gram,value_relation                  | 0.7854 0.7859 0.7863 0.7825 0.7839 |
| hidden_mse,key_relation              | 0.7797 0.7947 0.7879 0.7822 0.7854 |
| hidden_mse,query_relation            | 0.7771 0.7952 0.7935 0.7831 0.7849 |
| hidden_mse,value_relation            | 0.7796 0.7928 0.7889 0.7807 0.783 |
| mmd,key_relation                     | 0.7812 0.784 0.7847 0.7836 0.7809 |
| mmd,query_relation                   | 0.7851 0.7845 0.7845 0.7852 0.7869 |
| mmd,value_relation                   | 0.7818 0.7823 0.7824 0.7832 0.7835 |
| pkd,key_relation                     | 0.7694 0.7954 0.8001 0.7834 0.7892 |
| pkd,query_relation                   | 0.7704 0.7911 0.7971 0.7833 0.791 |
| pkd,value_relation                   | 0.7709 0.7957 0.7965 0.7846 0.7903 |
| bert-tiny                            |                         |                     |
| attention_ce_mean,cos                | 0.7203 0.7337 0.7297 0.7281 0.7289 |
| attention_ce_mean,gram               | 0.7289 0.7307 0.7286 0.7253 0.7301 |
| attention_ce_mean,hidden_mse         | 0.7294 0.7317 0.7328 0.7271 0.7285 |
| attention_ce_mean,key_relation       | 0.7298 0.7235 0.7279 0.7241 0.7257 |
| attention_ce_mean,mmd                | 0.7256 0.7231 0.7312 0.728 0.729 |
| attention_ce_mean,pkd                | 0.7174 0.7295 0.731 0.7264 0.727 |
| attention_ce_mean,query_relation     | 0.7307 0.7252 0.7271 0.7295 0.7267 |
| attention_ce_mean,value_relation     | 0.7227 0.7237 0.725 0.729 0.7292 |
| attention_mse_sum,cos                | 0.7231 0.7155 0.7237 0.7315 0.7271 |
| attention_mse_sum,gram               | 0.7227 0.7257 0.7279 0.7286 0.7255 |
| attention_mse_sum,hidden_mse         | 0.7259 0.7194 0.7258 0.723 0.7257 |
| attention_mse_sum,key_relation       | 0.7257 0.7272 0.7257 0.7252 0.7262 |
| attention_mse_sum,mmd                | 0.7249 0.7242 0.7266 0.7251 0.7262 |
| attention_mse_sum,pkd                | 0.7032 0.7319 0.73 0.7216 0.7262 |
| attention_mse_sum,query_relation     | 0.7264 0.7269 0.7247 0.7245 0.7281 |
| attention_mse_sum,value_relation     | 0.7269 0.7252 0.7233 0.7282 0.7282 |
| cos,key_relation                      | 0.7257 0.7304 0.7299 0.7259 0.7289 |
| cos,query_relation                   | 0.7224 0.7283 0.7307 0.7255 0.727 |
| cos,value_relation                   | 0.7215 0.7317 0.7286 0.7267 0.7301 |
| gram,key_relation                    | 0.7263 0.7259 0.7277 0.7303 0.7251 |
| gram,query_relation                  | 0.7246 0.7249 0.7301 0.7309 0.7291 |
| gram,value_relation                  | 0.7242 0.7277 0.7271 0.7293 0.728 |
| hidden_mse,key_relation              | 0.7227 0.7238 0.7264 0.722 0.7253 |
| hidden_mse,query_relation            | 0.7215 0.7321 0.7255 0.7206 0.7295 |
| hidden_mse,value_relation            | 0.7273 0.7249 0.7257 0.7217 0.7262 |
| mmd,key_relation                     | 0.7257 0.7235 0.7225 0.728 0.7244 |
| mmd,query_relation                   | 0.7248 0.7231 0.729 0.726 0.7283 |
| mmd,value_relation                   | 0.725 0.7198 0.727 0.7242 0.7241 |
| pkd,key_relation                     | 0.7147 0.7303 0.7306 0.7213 0.7261 |
| pkd,query_relation                   | 0.7135 0.727 0.7253 0.7257 0.7313 |
| pkd,value_relation                   | 0.716 0.7275 0.7299 0.7216 0.731 |
| mnl_match                            |                         |                     |
| bert-small                           |                         |                     |
| attention_ce_mean,cos                | 0.7985 0.8022 0.8039 0.7985 0.7984 |
| attention_ce_mean,gram               | 0.8034 0.8049 0.8046 0.8021 0.8008 |
| attention_ce_mean,hidden_mse         | 0.7945 0.8015 0.8038 0.8022 0.7992 |
| attention_ce_mean,key_relation       | 0.8009 0.7987 0.7964 0.8016 0.7985 |
| attention_ce_mean,mmd                | 0.8001 0.7987 0.7996 0.7988 0.8013 |
| attention_ce_mean,pkd                | 0.7862 0.8071 0.8115 0.7989 0.807 |
| attention_ce_mean,query_relation     | 0.8007 0.8017 0.798 0.8003 0.8005 |
| attention_ce_mean,value_relation     | 0.7999 0.7992 0.8008 0.7999 0.8013 |
| attention_mse_sum,cos                | 0.7911 0.8019 0.8038 0.7984 0.8041 |
| attention_mse_sum,gram               | 0.7979 0.7978 0.8025 0.7982 0.8021 |
| attention_mse_sum,hidden_mse         | 0.7944 0.8015 0.8041 0.7981 0.802 |

35
Table 6 continued from previous page

| Term                                | bert-mini | bert-tiny |
|-------------------------------------|-----------|-----------|
| attention_mse_sum, key_relation     | 0.7995    | 0.7223    |
| attention_mse_sum, mmd              | 0.8008    | 0.7251    |
| attention_mse_sum, pkd              | 0.7941    | 0.7226    |
| attention_mse_sum, query_relation   | 0.8019    | 0.7922    |
| attention_mse_sum, value_relation   | 0.7983    | 0.7962    |
| cos, key_relation                   | 0.7933    | 0.7375    |
| cos, query_relation                 | 0.794    | 0.7372    |
| cos, value_relation                 | 0.7938    | 0.7373    |
| gram, key_relation                  | 0.8037    | 0.7377    |
| gram, query_relation                | 0.7993    | 0.7368    |
| gram, value_relation                | 0.8033    | 0.7372    |
| hidden_mse, key_relation            | 0.7956    | 0.7372    |
| hidden_mse, query_relation          | 0.7969    | 0.7373    |
| hidden_mse, value_relation          | 0.7952    | 0.7368    |
| mmd, key_relation                   | 0.7995    | 0.7372    |
| mmd, query_relation                 | 0.7979    | 0.7368    |
| mmd, value_relation                 | 0.7996    | 0.7372    |
| pkd, key_relation                   | 0.7945    | 0.7372    |
| pkd, query_relation                 | 0.7893    | 0.7372    |
| pkd, value_relation                 | 0.7912    | 0.7372    |
| bert-mini First Last Dilatation First-1 Last-1 |           |           |
| attention_ce_mean, cos              | 0.7751    | 0.7751    |
| attention_ce_mean, gram             | 0.7745    | 0.7745    |
| attention_ce_mean, hidden_mse       | 0.7751    | 0.7751    |
| attention_ce_mean, key_relation     | 0.7749    | 0.7749    |
| attention_ce_mean, mmd              | 0.7771    | 0.7771    |
| attention_ce_mean, pkd              | 0.7701    | 0.7701    |
| attention_ce_mean, query_relation   | 0.7773    | 0.7773    |
| attention_ce_mean, value_relation   | 0.777    | 0.777    |
| attention_mse_sum, cos              | 0.7707    | 0.7707    |
| attention_mse_sum, gram             | 0.7768    | 0.7768    |
| attention_mse_sum, hidden_mse       | 0.7782    | 0.7782    |
| attention_mse_sum, key_relation     | 0.7739    | 0.7739    |
| attention_mse_sum, mmd              | 0.7734    | 0.7734    |
| attention_mse_sum, pkd              | 0.765    | 0.765    |
| attention_mse_sum, query_relation   | 0.7768    | 0.7768    |
| attention_mse_sum, value_relation   | 0.7751    | 0.7751    |
| cos, key_relation                   | 0.7737    | 0.7737    |
| cos, query_relation                 | 0.7734    | 0.7734    |
| cos, value_relation                 | 0.7751    | 0.7751    |
| gram, key_relation                  | 0.7775    | 0.7775    |
| gram, query_relation                | 0.7792    | 0.7792    |
| gram, value_relation                | 0.7727    | 0.7727    |
| hidden_mse, key_relation            | 0.7743    | 0.7743    |
| hidden_mse, query_relation          | 0.7752    | 0.7752    |
| hidden_mse, value_relation          | 0.7761    | 0.7761    |
| mmd, key_relation                   | 0.7787    | 0.7787    |
| mmd, query_relation                 | 0.7748    | 0.7748    |
| mmd, value_relation                 | 0.7753    | 0.7753    |
| pkd, key_relation                   | 0.7695    | 0.7695    |
| pkd, query_relation                 | 0.7722    | 0.7722    |
| pkd, value_relation                 | 0.7683    | 0.7683    |
| bert-tiny First Last Dilatation First-1 Last-1 |           |           |
| attention_ce_mean, cos              | 0.7223    | 0.7223    |
| attention_ce_mean, gram             | 0.7251    | 0.7251    |
| attention_ce_mean, hidden_mse       | 0.7226    | 0.7226    |
Table 6 continued from previous page

| Score | attention_ce_mean, key_relation | attention_ce_mean, mmd | attention_ce_mean, pkd | attention_ce_mean, query_relation | attention_ce_mean, value_relation | attention_mse_sum, cos | attention_mse_sum, gram | attention_mse_sum, hidden_mse | attention_mse_sum, key_relation | attention_mse_sum, mmd | attention_mse_sum, pkd | attention_mse_sum, query_relation | attention_mse_sum, value_relation |
|-------|-------------------------------|------------------------|------------------------|----------------------------------|----------------------------------|------------------------|------------------------|-------------------------------|----------------------------------|------------------------|------------------------|----------------------------------|----------------------------------|
|       | 0.7241                        | 0.725                  | 0.7257                 | 0.7223                           | 0.7248                           | 0.7246                  | 0.7274                  | 0.7266                         | 0.7258                           | 0.7242                 | 0.7117                  | 0.7207                           | 0.7261                         | 0.72                      |
|       |                               |                        |                        |                                  |                                  |                        |                        |                               |                                  |                       |                        |                                  |                                  |                        |
|       |                               |                        |                        |                                  |                                  |                        |                        |                               |                                  |                       |                        |                                  |                                  |                        |

### A.3.4 Best Practices Experiments

Table 7: Best Practice for BERT\textsubscript{mini}

|          | #para (M) | SST-2 | STS-B | QQP | MRPC | RTE | MNLI | QNLI | average |
|----------|-----------|-------|-------|-----|------|-----|------|------|---------|
| BERT\textsubscript{base} | 109       | 0.923 | 0.88  | 0.909 | 0.877 | 0.725 | 0.845 | 0.915 | 0.8715 |
| Previous SOTA (ELECTRA-small) | 14        | 0.912 | 0.875 | 0.89 | 0.88 | 0.667 | 0.813 | 0.884 | 0.8513 |
| Ours (BERT\textsubscript{mini}) | 11        | 0.91  | 0.873 | 0.903 | 0.874 | 0.7003 | 0.797 | 0.805 | 0.872 | 0.8554 |

### A.3.5 Weighted Single-Match Experiments

Table 8: Weighted Single Match Experiments

| MRPC     | bert-small First | Last | Dilatation | First1 | Last1 |
|----------|------------------|------|------------|--------|-------|
| attention_mse_sum | 0.8015 | 0.7868 | 0.8162 | 0.7966 | 0.8284 |
| attention_ce_mean | 0.799  | 0.799 | 0.8088 | 0.799  | 0.799  |
| hidden_mse | 0.75    | 0.7328 | 0.7377 | 0.7745 | 0.7892 |
| mmd | 0.7966 | 0.8407 | 0.8407 | 0.8113 | 0.8235 |
| cos | 0.7843 | 0.7157 | 0.7402 | 0.826  | 0.8186 |
| pkd | 0.7696 | 0.7549 | 0.7574 | 0.7941 | 0.7941 |
| query_relation | 0.8137 | 0.7745 | 0.7794 | 0.826  | 0.8015 |
|                | bert-mini | bert-tiny | bert-small | SST-2 |
|----------------|-----------|-----------|------------|-------|
| key_relation   | 0.8431    | 0.8162    | 0.8257     | 0.8268 |
| value_relation | 0.8382    | 0.8162    | 0.8257     | 0.8268 |
| bert-mini      | First     | Last      | Dilatation | First  |
| attention_mse_sum | 0.7892   | 0.723     | 0.7402     | 0.7279 |
| attention_ce_mean | 0.8211   | 0.8137    | 0.8235     | 0.7279 |
| hidden_mse     | 0.723     | 0.7108    | 0.7181     | 0.7328 |
| mmd            | 0.7647    | 0.8113    | 0.7892     | 0.8186 |
| gram           | 0.6863    | 0.7206    | 0.7157     | 0.7181 |
| cos            | 0.7525    | 0.7402    | 0.7402     | 0.7451 |
| pkd            | 0.6838    | 0.8186    | 0.7745     | 0.7475 |
| query_relation | 0.799     | 0.7794    | 0.777      | 0.7623 |
| key_relation   | 0.8162    | 0.7623    | 0.777      | 0.7623 |
| value_relation | 0.8186    | 0.7892    | 0.799      | 0.7892 |
| bert-tiny      | First     | Last      | Dilatation | First  |
| attention_mse_sum | 0.7426   | 0.7157    | 0.7279     | 0.7279 |
| attention_ce_mean | 0.7255   | 0.7279    | 0.7255     | 0.7279 |
| hidden_mse     | 0.723     | 0.723     | 0.7255     | 0.7255 |
| mmd            | 0.7353    | 0.723     | 0.7328     | 0.7328 |
| gram           | 0.7181    | 0.7059    | 0.7206     | 0.7181 |
| cos            | 0.7328    | 0.7255    | 0.7181     | 0.7181 |
| pkd            | 0.6838    | 0.7475    | 0.7451     | 0.7475 |
| query_relation | 0.7402    | 0.7549    | 0.775      | 0.7549 |
| key_relation   | 0.7525    | 0.7623    | 0.7598     | 0.7598 |
| value_relation | 0.7353    | 0.7549    | 0.7475     | 0.7549 |
| bert-small     | First     | Last      | Dilatation | First  |
| attention_mse_sum | 0.8933   | 0.8888    | 0.8922     | 0.8922 |
| attention_ce_mean | 0.8899   | 0.8899    | 0.8865     | 0.8922 |
| hidden_mse     | 0.8761    | 0.8853    | 0.883      | 0.8922 |
| mmd            | 0.8899    | 0.8979    | 0.9025     | 0.8933 |
| gram           | 0.8807    | 0.8842    | 0.8956     | 0.8979 |
| cos            | 0.8807    | 0.8899    | 0.8888     | 0.8999 |
| pkd            | 0.8819    | 0.8933    | 0.8888     | 0.8865 |
| query_relation | 0.8922    | 0.8945    | 0.8922     | 0.8933 |
| key_relation   | 0.8911    | 0.8876    | 0.8945     | 0.8933 |
| value_relation | 0.8876    | 0.8842    | 0.8853     | 0.8865 |
| bert-mini      | First     | Last      | Dilatation | First  |
| attention_mse_sum | 0.8693   | 0.8704    | 0.8555     | 0.8704 |
| attention_ce_mean | 0.8716   | 0.8681    | 0.8716     | 0.8716 |
| hidden_mse     | 0.8394    | 0.8509    | 0.8647     | 0.8819 |
| mmd            | 0.8658    | 0.8727    | 0.8635     | 0.8761 |
| gram           | 0.8601    | 0.8291    | 0.8555     | 0.8601 |
| cos            | 0.8647    | 0.8612    | 0.8567     | 0.8693 |
| pkd            | 0.8291    | 0.8704    | 0.8681     | 0.8635 |
| query_relation | 0.8716    | 0.8589    | 0.8727     | 0.8658 |
| key_relation   | 0.8704    | 0.8647    | 0.8681     | 0.8681 |
| value_relation | 0.8716    | 0.875     | 0.8727     | 0.8761 |
| bert-tiny      | First     | Last      | Dilatation | First  |
| attention_mse_sum | 0.8257   | 0.8314    | 0.8349     | 0.8291 |
| attention_ce_mean | 0.828    | 0.828     | 0.828      | 0.828 |
| hidden_mse     | 0.8291    | 0.8268    | 0.8291     | 0.8303 |
| mmd            | 0.8257    | 0.8314    | 0.8257     | 0.8257 |
| gram           | 0.8268    | 0.8314    | 0.836      | 0.8326 |
|     | QQP | QNLI |
|-----|-----|-----|
| **cos** | 0.8257 | 0.8288 |
| **pkd** | 0.8188 | 0.8291 |
| **query relation** | 0.828 | 0.8291 |
| **key relation** | 0.8326 | 0.8337 |
| **value relation** | 0.8234 | 0.8257 |

| bert-small | First | Last | Dilatation | First1 | Last1 |
|-------------|-------|------|------------|--------|-------|
| attention_mse_sum | 0.8932 | 0.8899 | 0.8937 | 0.8915 | 0.8953 |
| attention_ce_mean | 0.8922 | 0.891 | 0.8918 | 0.8913 | 0.8929 |
| hidden_mse | 0.8887 | 0.899 | 0.8982 | 0.8916 | 0.8961 |
| mmd | 0.8925 | 0.8935 | 0.8951 | 0.8922 | 0.895 |
| gram | 0.8926 | 0.8972 | 0.8994 | 0.8914 | 0.8967 |
| cos | 0.8918 | 0.8963 | 0.8965 | 0.893 | 0.8953 |
| pkd | 0.8909 | 0.8971 | 0.8975 | 0.8922 | 0.8972 |
| query_relation | 0.8755 | 0.8756 | 0.8758 | 0.8774 | 0.8801 |
| key_relation | 0.8763 | 0.8811 | 0.8801 | 0.8756 | 0.8795 |
| value_relation | 0.881 | 0.8765 | 0.8737 | 0.8787 | 0.8775 |

| bert-mini | First | Last | Dilatation | First1 | Last1 |
|-------------|-------|------|------------|--------|-------|
| attention_mse_sum | 0.8865 | 0.8768 | 0.882 | 0.8861 | 0.888 |
| attention_ce_mean | 0.8881 | 0.8883 | 0.8855 | 0.8864 | 0.8897 |
| hidden_mse | 0.8842 | 0.8893 | 0.8927 | 0.8886 | 0.8906 |
| mmd | 0.8897 | 0.8825 | 0.892 | 0.8856 | 0.8918 |
| gram | 0.8888 | 0.8829 | 0.8949 | 0.8884 | 0.8923 |
| cos | 0.8874 | 0.8921 | 0.8933 | 0.8881 | 0.8893 |
| pkd | 0.8863 | 0.8939 | 0.8959 | 0.8886 | 0.8936 |
| query_relation | 0.8742 | 0.8751 | 0.8749 | 0.8733 | 0.8743 |
| key_relation | 0.8754 | 0.878 | 0.8763 | 0.8762 | 0.877 |
| value_relation | 0.8775 | 0.8757 | 0.8752 | 0.8762 | 0.8743 |

| bert-tiny | First | Last | Dilatation | First1 | Last1 |
|-------------|-------|------|------------|--------|-------|
| attention_mse_sum | 0.8688 | 0.8651 | 0.8657 | 0.8684 | 0.8689 |
| attention_ce_mean | 0.8699 | 0.8682 | 0.8687 | 0.8682 | 0.8699 |
| hidden_mse | 0.8633 | 0.8666 | 0.8624 | 0.8656 | 0.871 |
| mmd | 0.8657 | 0.8471 | 0.8633 | 0.8651 | 0.8684 |
| gram | 0.8662 | 0.8664 | 0.8712 | 0.8657 | 0.8734 |
| cos | 0.8696 | 0.8694 | 0.8686 | 0.8681 | 0.8707 |
| pkd | 0.869 | 0.8718 | 0.8698 | 0.8618 | 0.8687 |
| query_relation | 0.8644 | 0.8618 | 0.8602 | 0.8619 | 0.8633 |
| key_relation | 0.8597 | 0.8642 | 0.857 | 0.8612 | 0.8569 |
| value_relation | 0.8619 | 0.8564 | 0.8588 | 0.864 | 0.8602 |

| bert-small | First | Last | Dilatation | First1 | Last1 |
|-------------|-------|------|------------|--------|-------|
| attention_mse_sum | 0.8592 | 0.8627 | 0.8717 | 0.8722 | 0.8742 |
| attention_ce_mean | 0.8728 | 0.8728 | 0.8704 | 0.8722 | 0.8704 |
| hidden_mse | 0.8287 | 0.8256 | 0.842 | 0.8508 | 0.8612 |
| mmd | 0.8583 | 0.8592 | 0.8634 | 0.8715 | 0.8744 |
| gram | 0.8514 | 0.8298 | 0.8426 | 0.8739 | 0.8741 |
| cos | 0.8528 | 0.8528 | 0.8572 | 0.8693 | 0.864 |
| pkd | 0.8298 | 0.8678 | 0.8605 | 0.8717 | 0.8651 |
| query_relation | 0.8726 | 0.8689 | 0.8742 | 0.8719 | 0.8715 |
| key_relation | 0.8691 | 0.864 | 0.8644 | 0.8686 | 0.8667 |
| value_relation | 0.8662 | 0.8667 | 0.864 | 0.8684 | 0.8684 |

| bert-mini | First | Last | Dilatation | First1 | Last1 |
|-------------|-------|------|------------|--------|-------|
| attention_mse_sum | 0.8369 | 0.8272 | 0.8298 | 0.8455 | 0.8442 |
|                     | bert-tiny | First | Last | Dilatation | First1 | Last1 |
|---------------------|-----------|-------|------|------------|--------|-------|
| attention_ce_mean   | 0.8437    | 0.844 | 0.844 | 0.8442     | 0.8439 |
| hidden_mse          | 0.8045    | 0.7884| 0.8105| 0.838      | 0.8391 |
| mmd                 | 0.8356    | 0.816 | 0.8371| 0.8422     | 0.8457 |
| gram                | 0.831     | 0.6447| 0.8133| 0.8446     | 0.844  |
| cos                 | 0.8334    | 0.827 | 0.8354| 0.8387     | 0.8426 |
| pkd                 | 0.8195    | 0.8428| 0.8439| 0.8418     | 0.8473 |
| query_relation      | 0.8473    | 0.8437| 0.8444| 0.8475     | 0.8477 |
| key_relation        | 0.8448    | 0.8424| 0.8402| 0.8459     | 0.8445 |
| value_relation      | 0.8459    | 0.8387| 0.8386| 0.8424     | 0.8418 |

|                     | bert-small | First | Last | Dilatation | First1 | Last1 |
|---------------------|------------|-------|------|------------|--------|-------|
| attention_mse_sum   | 0.7932     | 0.7924| 0.7877| 0.7915     | 0.7946 |
| attention_ce_mean   | 0.7968     | 0.7977| 0.797 | 0.7966     | 0.797  |
| hidden_mse          | 0.7728     | 0.7712| 0.7712| 0.7877     | 0.7825 |
| mmd                 | 0.7871     | 0.78  | 0.7867| 0.791      | 0.7917 |
| gram                | 0.7752     | 0.7593| 0.7748| 0.7922     | 0.7899 |
| cos                 | 0.7833     | 0.7811| 0.7816| 0.7913     | 0.791  |
| pkd                 | 0.7791     | 0.79  | 0.7899| 0.7941     | 0.7904 |
| query_relation      | 0.8089     | 0.8076| 0.8083| 0.8089     | 0.8098 |
| key_relation        | 0.8062     | 0.7988| 0.7977| 0.8069     | 0.8018 |
| value_relation      | 0.8052     | 0.7985| 0.801 | 0.804      | 0.8025 |

|                     | bert-mini  | First | Last | Dilatation | First1 | Last1 |
|---------------------|------------|-------|------|------------|--------|-------|
| attention_mse_sum   | 0.639      | 0.6065| 0.5957| 0.6751     | 0.6643 |
| attention_ce_mean   | 0.6787     | 0.6787| 0.6751| 0.6787     | 0.6823 |
| hidden_mse          | 0.5451     | 0.5596| 0.556 | 0.6245     | 0.6354 |
| mmd                 | 0.6173     | 0.6245| 0.6245| 0.6643     | 0.6498 |
| gram                | 0.5921     | 0.5596| 0.574 | 0.6643     | 0.6643 |
| cos                 | 0.5632     | 0.5668| 0.556 | 0.6787     | 0.6354 |
| pkd                 | 0.5776     | 0.6209| 0.6065| 0.6462     | 0.6426 |
| query_relation      | 0.6751     | 0.6462| 0.6534| 0.6643     | 0.6498 |
| key_relation        | 0.6498     | 0.6282| 0.6137| 0.6715     | 0.657  |
| value_relation      | 0.639      | 0.6823| 0.6839| 0.6606     | 0.6751 |

|                     | RTE        | bert-tiny | First | Last | Dilatation | First1 | Last1 |
|---------------------|------------|-----------|-------|------|------------|--------|-------|
| attention_mse_sum   | 0.6282     | 0.5632   | 0.5921| 0.657| 0.6534     |
| attention_ce_mean   | 0.6679     | 0.6679   | 0.6643| 0.6787| 0.6787     |
| hidden_mse          | 0.5776     | 0.5632   | 0.556 | 0.6462| 0.6209     |
| mmd                 | 0.6282     | 0.6354   | 0.6462| 0.6498| 0.6282     |
| gram                | 0.574     | 0.5415   | 0.5596| 0.6895| 0.6354     |
| cos                 | 0.5884     | 0.5632   | 0.5668| 0.6679| 0.6209     |
| pkd                 | 0.5632     | 0.5921   | 0.5993| 0.657 | 0.6245     |
| query_relation      | 0.6462     | 0.639    | 0.6426| 0.6679| 0.6354     |
| key_relation        | 0.6534     | 0.6282   | 0.639 | 0.6679| 0.6173     |
| value_relation      | 0.6462     | 0.6354   | 0.6282| 0.6498| 0.6498     |

|                     | bert-tiny  | First | Last | Dilatation | First1 | Last1 |
|---------------------|------------|-------|------|------------|--------|-------|
| attention_mse_sum   | 0.6101     | 0.6137| 0.6065| 0.6282     | 0.5957 |
| attention_ce_mean   | 0.6318     | 0.6209| 0.6137| 0.6173     | 0.6209 |
| hidden_mse          | 0.5921     | 0.5812| 0.5812| 0.6101     | 0.5957 |
| mmd                 | 0.6101     | 0.6173| 0.6029| 0.6173     | 0.639  |
| gram                | 0.5812     | 0.5704| 0.5812| 0.6137     | 0.5848 |
| cos                 | 0.6065     | 0.5884| 0.6065| 0.6101     | 0.5884 |
| pkd                 | 0.5451     | 0.5993| 0.5957| 0.574      | 0.5993 |
| query_relation      | 0.6354     | 0.6101| 0.6029| 0.6282     | 0.6029 |
| key_relation        | 0.6426     | 0.6245| 0.6173| 0.6209     | 0.6245 |
|         | bert-small | bert-mini | bert-tiny |         |
|---------|------------|-----------|-----------|---------|
|         | First      | Last      | Dilatation| First1  | Last1   |
| value_relation | 0.6245  | 0.6029  | 0.5921     | 0.6426  | 0.6065  |
| CoLA    |            |           |            |         |         |
| attention_mse_sum | 0.7747  | 0.7574  | 0.7747     | 0.7747  | 0.7737  |
| attention_ce_mean | 0.7766  | 0.7747  | 0.7776     | 0.7766  | 0.7728  |
| hidden_mse | 0.697   | 0.6961  | 0.6961     | 0.7824  | 0.7584  |
| mmd     | 0.745     | 0.7459  | 0.767      | 0.7955  | 0.767   |
| gram    | 0.7306    | 0.7018  | 0.7114     | 0.7689  | 0.7728  |
| cos     | 0.7181    | 0.7114  | 0.7133     | 0.7776  | 0.7718  |
| pkd     | 0.7095    | 0.7373  | 0.7277     | 0.7709  | 0.768   |
| query_relation | 0.7766  | 0.7699  | 0.7689     | 0.7785  | 0.7747  |
| key_relation | 0.7709  | 0.7603  | 0.7709     | 0.7824  | 0.7603  |
| value_relation | 0.7776  | 0.7756  | 0.7689     | 0.7814  | 0.7689  |
|         |            |           |            |         |         |
| STS-B   |            |           |            |         |         |
| attention_mse_sum | 0.6922  | 0.697   | 0.7057     | 0.7469  | 0.7335  |
| attention_ce_mean | 0.745   | 0.7421  | 0.7469     | 0.7478  | 0.743   |
| hidden_mse | 0.6932  | 0.6961  | 0.6951     | 0.7344  | 0.7229  |
| mmd     | 0.6942    | 0.7152  | 0.7162     | 0.7517  | 0.7354  |
| gram    | 0.6942    | 0.6913  | 0.6913     | 0.745   | 0.7344  |
| cos     | 0.6932    | 0.6942  | 0.698      | 0.7478  | 0.7181  |
| pkd     | 0.6913    | 0.6999  | 0.6942     | 0.7392  | 0.7335  |
| query_relation | 0.7555  | 0.7325  | 0.7488     | 0.7555  | 0.743   |
| key_relation | 0.7507  | 0.7277  | 0.7507     | 0.7565  | 0.7421  |
| value_relation | 0.7593  | 0.7411  | 0.7392     | 0.7507  | 0.7546  |
|         |            |           |            |         |         |
| bert-tiny | First      | Last      | Dilatation| First1  | Last1   |
| attention_mse_sum | 0.6942  | 0.6913  | 0.6913     | 0.6932  | 0.6913  |
| attention_ce_mean | 0.697   | 0.6913  | 0.6951     | 0.6961  | 0.697   |
| hidden_mse | 0.6913  | 0.6913  | 0.6922     | 0.6913  | 0.6913  |
| mmd     | 0.6922    | 0.6922  | 0.6922     | 0.6913  | 0.6913  |
| gram    | 0.6913    | 0.6932  | 0.6913     | 0.6913  | 0.6932  |
| cos     | 0.6922    | 0.6913  | 0.6922     | 0.6913  | 0.6913  |
| pkd     | 0.6913    | 0.6913  | 0.6961     | 0.6913  | 0.6913  |
| query_relation | 0.6913  | 0.6922  | 0.6922     | 0.6913  | 0.6951  |
| key_relation | 0.6942  | 0.6913  | 0.6913     | 0.6913  | 0.6913  |
| value_relation | 0.6932  | 0.6913  | 0.6951     | 0.6961  | 0.6951  |
|         |            |           |            |         |         |
| STS-B   |            |           |            |         |         |
| attention_mse_sum | 0.8731  | 0.8705  | 0.8727     | 0.8739  | 0.8745  |
| attention_ce_mean | 0.8731  | 0.8735  | 0.8727     | 0.8732  | 0.8725  |
| hidden_mse | 0.8656  | 0.8646  | 0.8642     | 0.8718  | 0.8717  |
| mmd     | 0.8727    | 0.8678  | 0.8685     | 0.8752  | 0.8748  |
| gram    | 0.8715    | 0.8458  | 0.862      | 0.8728  | 0.8723  |
| cos     | 0.8724    | 0.8708  | 0.8694     | 0.874   | 0.8733  |
| pkd     | 0.8663    | 0.8698  | 0.8693     | 0.8726  | 0.873   |
| query_relation | 0.8773  | 0.8772  | 0.876      | 0.8762  | 0.8753  |
| key_relation | 0.8766  | 0.8734  | 0.8748     | 0.8772  | 0.8744  |
| value_relation | 0.8745  | 0.8757  | 0.8741     | 0.877   | 0.8752  |
|         |            |           |            |         |         |
| bert-mini | First      | Last      | Dilatation| First1  | Last1   |
| attention_mse_sum | 0.8674  | 0.8249  | 0.8431     | 0.8655  | 0.8638  |
| attention_ce_mean | 0.865   | 0.8641  | 0.8629     | 0.8627  | 0.8656  |
| hidden_mse | 0.8549  | 0.8455  | 0.8511     | 0.8665  | 0.8675  |
| mmd     | 0.8657    | 0.8608  | 0.8654     | 0.8678  | 0.866   |
| gram    | 0.8551    | 0.7472  | 0.8283     | 0.8677  | 0.8643  |
|                | bert-tiny |               |               | bert-small |               |               |
|----------------|-----------|---------------|---------------|-----------|---------------|---------------|
|                | First     | Last          | Dilatation    | First     | Last          | Dilatation    |
| cos            | 0.8626    | 0.8552        | 0.8568        | 0.8684    | 0.8667        |               |
| pkd            | 0.8558    | 0.8591        | 0.855         | 0.8607    | 0.865         |               |
| query_relation | 0.8693    | 0.8686        | 0.8679        | 0.8688    | 0.8684        |               |
| key_relation   | 0.869     | 0.8689        | 0.8683        | 0.8688    | 0.8697        |               |
| value_relation | 0.8696    | 0.869         | 0.8693        | 0.8688    | 0.8701        |               |
|                |           |               |               |           |               |               |
| attention_mse_sum | 0.8164 | 0.8139        | 0.8171        | 0.8163    | 0.8161        |               |
| attention_ce_mean | 0.8168 | 0.8168        | 0.8168        | 0.8156    | 0.8168        |               |
| hidden_mse     | 0.8176    | 0.8191        | 0.8169        | 0.8192    | 0.8165        |               |
| mmd            | 0.8095    | 0.812         | 0.8123        | 0.8119    | 0.8187        |               |
| gram           | 0.8185    | 0.8046        | 0.8105        | 0.8149    | 0.8176        |               |
| cos            | 0.8181    | 0.8185        | 0.8184        | 0.8175    | 0.8163        |               |
| pkd            | 0.8146    | 0.8156        | 0.8145        | 0.8151    | 0.8126        |               |
| query_relation | 0.8229    | 0.823         | 0.8231        | 0.8227    | 0.823         |               |
| key_relation   | 0.8187    | 0.8213        | 0.821         | 0.8194    | 0.822         |               |
| value_relation | 0.8212    | 0.8167        | 0.8169        | 0.8203    | 0.8214        |               |

**MNLI-mm**

|                | bert-small |               |               | bert-mini |               |               |
|----------------|------------|---------------|---------------|-----------|---------------|---------------|
|                | First      | Last          | Dilatation    | First     | Last          | Dilatation    |
| attention_mse_sum | 0.7965 | 0.7938        | 0.8004        | 0.8001    | 0.8036        |               |
| attention_ce_mean | 0.7991 | 0.7993        | 0.8003        | 0.8008    | 0.8027        |               |
| hidden_mse      | 0.784     | 0.8166        | 0.814         | 0.7944    | 0.8087        |               |
| mmd             | 0.7947    | 0.802         | 0.8034        | 0.7975    | 0.8049        |               |
| gram            | 0.7859    | 0.7966        | 0.8043        | 0.8002    | 0.8064        |               |
| cos             | 0.7908    | 0.8086        | 0.8102        | 0.7974    | 0.8037        |               |
| pkd             | 0.7903    | 0.8135        | 0.8127        | 0.7992    | 0.8098        |               |
| query_relation  | 0.7973    | 0.7957        | 0.7952        | 0.7994    | 0.7984        |               |
| key_relation    | 0.8012    | 0.7912        | 0.7957        | 0.7982    | 0.7971        |               |
| value_relation  | 0.7957    | 0.7915        | 0.7931        | 0.7956    | 0.7937        |               |

**MNLI-m**

|                | bert-tiny |               |               | bert-small |               |               |
|----------------|-----------|---------------|---------------|-----------|---------------|---------------|
|                | First      | Last          | Dilatation    | First     | Last          | Dilatation    |
| attention_mse_sum | 0.7249 | 0.7156        | 0.7079        | 0.7236    | 0.7248        |               |
| attention_ce_mean | 0.7272 | 0.724         | 0.7275        | 0.7231    | 0.7281        |               |
| hidden_mse      | 0.7189    | 0.7304        | 0.7288        | 0.7228    | 0.729         |               |
| mmd             | 0.7234    | 0.6882        | 0.7191        | 0.724     | 0.7269        |               |
| gram            | 0.7224    | 0.7023        | 0.7245        | 0.7253    | 0.7316        |               |
| cos             | 0.7234    | 0.7273        | 0.7297        | 0.7252    | 0.7317        |               |
| pkd             | 0.7148    | 0.7268        | 0.7306        | 0.7254    | 0.7321        |               |
| query_relation  | 0.7246    | 0.7266        | 0.7243        | 0.7248    | 0.723         |               |
| key_relation    | 0.7265    | 0.72           | 0.723         | 0.7259    | 0.721         |               |
| value_relation  | 0.7248    | 0.7165        | 0.7145        | 0.727     | 0.7215        |               |

|                | bert-small |               |               |               |
|----------------|-----------|---------------|---------------|---------------|
|                | First      | Last          | Dilatation    |               |
| attention_mse_sum | 0.7942 | 0.7923        | 0.8001        | 0.8015    | 0.8043        |
|                     | First     | Last      | Dilatation | First1    | Last1     |
|---------------------|-----------|-----------|------------|-----------|-----------|
| attention_ce_mean   | 0.7978    | 0.8003    | 0.8006     | 0.801     | 0.8021    |
| hidden_mse          | 0.7912    | 0.8057    | 0.8093     | 0.7955    | 0.8017    |
| mmd                 | 0.7963    | 0.8016    | 0.8056     | 0.8012    | 0.802     |
| gram                | 0.7977    | 0.8025    | 0.8098     | 0.7996    | 0.8059    |
| cos                 | 0.796     | 0.8016    | 0.8052     | 0.7981    | 0.8021    |
| pkd                 | 0.7902    | 0.8095    | 0.8084     | 0.7961    | 0.807     |
| query_relation      | 0.7911    | 0.7874    | 0.7905     | 0.791     | 0.7954    |
| key_relation        | 0.7909    | 0.7927    | 0.7934     | 0.7894    | 0.7943    |
| value_relation      | 0.7878    | 0.7925    | 0.7907     | 0.79      | 0.7927    |
| bert-mini           |           |           |            |           |           |
| attention_mse_sum   | 0.767     | 0.7513    | 0.7617     | 0.775     | 0.7801    |
| attention_ce_mean   | 0.7779    | 0.7776    | 0.7748     | 0.7749    | 0.7789    |
| hidden_mse          | 0.7723    | 0.7869    | 0.7851     | 0.7785    | 0.7819    |
| mmd                 | 0.775     | 0.7787    | 0.7836     | 0.7805    | 0.7779    |
| gram                | 0.7754    | 0.7763    | 0.7882     | 0.7764    | 0.7823    |
| cos                 | 0.7752    | 0.7815    | 0.7826     | 0.7764    | 0.7794    |
| pkd                 | 0.7715    | 0.7858    | 0.7863     | 0.7769    | 0.7828    |
| query_relation      | 0.7682    | 0.7705    | 0.7721     | 0.7679    | 0.7705    |
| key_relation        | 0.7693    | 0.7725    | 0.7686     | 0.7689    | 0.7721    |
| value_relation      | 0.7697    | 0.7658    | 0.7686     | 0.771     | 0.771     |
| bert-tiny           |           |           |            |           |           |
| attention_mse_sum   | 0.7224    | 0.7166    | 0.7174     | 0.7226    | 0.7218    |
| attention_ce_mean   | 0.7252    | 0.7244    | 0.7253     | 0.725     | 0.7238    |
| hidden_mse          | 0.7187    | 0.7215    | 0.7194     | 0.7188    | 0.7243    |
| mmd                 | 0.7178    | 0.7047    | 0.7222     | 0.7208    | 0.7203    |
| gram                | 0.7202    | 0.7208    | 0.7231     | 0.7245    | 0.7253    |
| cos                 | 0.7211    | 0.7259    | 0.7227     | 0.7228    | 0.7246    |
| pkd                 | 0.7112    | 0.7227    | 0.7241     | 0.7205    | 0.7256    |
| query_relation      | 0.7261    | 0.7242    | 0.7231     | 0.7261    | 0.7223    |
| key_relation        | 0.7246    | 0.7271    | 0.7264     | 0.7251    | 0.7283    |
| value_relation      | 0.7285    | 0.7148    | 0.7244     | 0.7279    | 0.7261    |