Learning Multilingual Representation for Natural Language Understanding with Enhanced Cross-Lingual Supervision

Yinpeng Guo¹, Liangyou Li¹, Xin Jiang¹, and Qun Liu¹
¹Huawei Noah’s Ark Lab
{yinpeng.guo, liliangyou, jiang.xin, qun.liu}@huawei.com

Abstract

Recently, pre-training multilingual language models has shown great potential in learning multilingual representation, a crucial topic of natural language processing. Prior works generally use a single mixed attention (MA) module, following TLM (Conneau and Lample, 2019), for attending to intra-lingual and cross-lingual contexts equivalently and simultaneously. In this paper, we propose a network named decomposed attention (DA) as a replacement of MA. The DA consists of an intra-lingual attention (IA) and a cross-lingual attention (CA), which model intra-lingual and cross-lingual supervisions respectively. In addition, we introduce a language-adaptive re-weighting strategy during training to further boost the model’s performance. Experiments on various cross-lingual natural language understanding (NLU) tasks show that the proposed architecture and learning strategy significantly improve the model’s cross-lingual transferability.

1 Introduction

Representation learning has become an important research area in natural language processing (NLP) for a long time. It aims to build models to automatically discover features from raw text for downstream NLP tasks. Beyond representation learning for specific languages, multilingual representation learning is recently drawing more attention. It learns universal representation across languages and strongly improves performance of cross-lingual natural language understanding (NLU) tasks, especially zero-shot transfer on low-resource languages.

Although overall performance of cross-lingual transfer has been remarkably improved by previous works, most models are built upon mixed attention (MA) modules. As shown in Figure 1a in MA modules, intra-lingual context and cross-lingual context are attended without distinction.

Then, questions come out naturally: Why should intra-lingual context and cross-lingual context be mixed? Is the model’s capacity sufficient to capture intra-lingual and cross-lingual with a set of shared parameters? Is the model prefer intra-lingual context for word prediction?

In this paper, we propose decomposed attention (DA) which consists of an intra-lingual attention (IA) to learn from intra-lingual context and a following cross-lingual attention (CA) which depends on cross-lingual context. Figure 1b shows the architecture of DA. This kind of independent modeling improves model capacity and is helpful to capture cross-lingual information, as in neural machine translation (Vaswani et al., 2017). Figure 2 shows the attention heatmaps obtained from MA and DA, respectively. Comparing attentions between IA and CA in Figure 2a we find that MA prefers to distribute attentions on intra-lingual context than cross-lingual context. In other words, MA predict
words depending more on intra-lingual context. On the contrary, as shown in Figure 2b, DA utilizes cross-lingual context much better.

In addition to the above improvement in architectures, we further explore language-adaptive sample re-weighting which encourages the model to learn challenging examples for each language. The difficulty of an example is defined as its prediction loss. In our experiment, the re-weighting strategy is language-wise and only triggered when easy examples are well learned.

In general, the contributions of this paper include:

- We propose a DA network for multilingual representation learning, which consists of an intra-lingual attention and a cross-lingual attention.
- We propose a language-adaptive sample re-weighting strategy, which encourages the model to gradually learn difficult examples in language-wise.
- Experimental results on different types of multilingual NLU tasks and further analysis prove the effectiveness of the proposed DA network and re-weighting strategy.

2 Related Work

Encoder-Only Models In recent years, with great progresses in language model pre-training, multilingual representation learning methods based on pre-training are emerging in large numbers. Inspired by masked language modeling (MLM) (Devlin et al., 2019), mBERT proposed multilingual MLM (mMLM) to learn representation from a mixture of Wikipedia corpora from one hundred languages. XLM (Conneau and Lample, 2019) introduced parallel translation corpora to further incorporate a cross-lingual pre-training objective TLM with mMLM. TLM concatenates a pair of translation sentences as a single input sequence to Transformer encoder, then performs the masked word prediction objective as MLM. Subsequently, following mBERT and XLM, various multilingual pre-training models were proposed. XLM-Roberta (Conneau et al., 2020) extended XLM via enlarging model size and training corpora. Unicoder (Huang et al., 2019) introduced three additional cross-lingual pre-training tasks, i.e. cross-lingual word recovery, cross-lingual paraphrase identification and cross-lingual masked language model. ERNIE-M (Ouyang et al., 2021) leveraged cross-lingual alignments to augment multilingual training data from monolingual corpora. HICTL (Wei et al., 2021b) designed a hierarchical contrastive learning framework on parallel corpora, including word-level and sentence-level contrast.

Encoder-Decoder Models In addition to encoder-only models mentioned above, encoder-decoder models were also explored in multilingual representation learning. Multilingual BART (mBART) (Liu et al., 2020) and Multilingual T5 (mT5) (Xue et al., 2020) applied denoising auto-encoder strategies in multilingual pre-training. VECO (Luo et al., 2021) integrated MLM, TLM and translation pre-training and shared self-attention parameters of encoder and decoder.

Decoder-Only Models Besides, Guo et al. (2019) leveraged decoder-only Transformer network to learn multilingual language model for zero-shot paraphrase generation. However, most of these models adopt MA as their encoder’s default backbone.

3 Methodology

We adopt a Transformer encoder as the backbone of our model since we aim to tackle NLU tasks where decoder is unnecessary. In addition, because the proposed DA module is an alternative to the conventional MA module, which is the default backbone of current SOTA encoder-decoder models, it is feasible to apply the DA module in encoder-decoder models.

3.1 Multilingual Masked Language Modeling (MMLM)

Since our model is based on Transformer encoder, it is convenient to follow the multilingual masked language modeling (MMLM) objective proposed by mBERT. During training on monolingual corpora (e.g. Wikipedia), sentences from the same document are concatenated as a single sequence to fulfill
the maximum input length of the encoder. Data from different languages are mixed so that model parameters are trained for various languages. Before inputted into the model, words in the sequence are randomly replaced with probability of 15%, in which 80% is a [MASK] token, 10% is a random word and the rest remain unchanged. The training objective is to predict the masked words depending on other visible words in the same language as described in Figure 10, Appendix B. More specifically, the cross-entropy objective function of MMLM can be written as Equation 1:

$$L_{MMLM} = - \sum_{x \in D_m} \sum_{i \in [1,n]} \log p(x_i | X/x_i)$$  \hspace{1cm} (1)

where $D_m$ is the monolingual corpus, $n$ is the maximum sequence length, $X/x_i$ denotes all the input words in $X$ except for $x_i$ and $p(\cdot | \cdot)$ is the prediction probability.

In experiments, we apply MMLM objective on both baseline and our models.

### 3.2 Decomposed Attention (DA)

![Diagram](image)

Figure 1: Comparison between mixed attention (MA) and the proposed decomposed attention (DA): intra-lingual attention (IA) and cross-lingual attention (CA).

During training on cross-lingual parallel corpora, the input is a pair of masked translation sentences in two languages. For instance in Figure 1, the sentence $X = \{x_1, x_2, \ldots, x_n\}$ and its translation $Y = \{y_1, y_2, \ldots, y_n\}$ are the inputs. The objective function of MA and DA are identical, which aims to predict masked words conditioning on both intra-lingual and cross-lingual context, except for the masked word itself, as described in Equation 2:

$$L_{Trans} = - \sum_{(x,y) \in D_b} \left[ \sum_{i=1}^{n} \log p(x_i | C_{x_i}) + \sum_{j=1}^{n} \log p(y_j | C_{y_j}) \right]$$  \hspace{1cm} (2)

Where $D_b$ is the bilingual translation corpus, $C_{x_i} = (X \cup Y)/x_i$ is the context of $x_i$ and $C_{y_j} = (X \cup Y)/y_j$ is the context of $y_j$.

Although MA and DA share the same objective function, calculation details of their prediction probabilities are different. Since MA and DA are built on attention mechanism, we can formalize the computation of attention as Equation 3 following the notation from Vaswani et al. (2017).

$$\text{Attention}(Q, K, V) = \text{softmax}\left( \frac{QK^T}{\sqrt{d_k}} \right)V$$  \hspace{1cm} (3)

Where $Q \in \mathbb{R}^{n \times d_k}, K \in \mathbb{R}^{n \times d_k}, V \in \mathbb{R}^{n \times d_k}$ are the query, key and value matrices respectively and $\sqrt{d_k}$ is the dimension size. For simplicity, we use $\odot$ to represent common computations
According to Equation 6, FL encourages the model to focus more on words of lower prediction probability (higher CE loss). In another word, FL pushes the model to learn more from difficult examples. But in our preliminary experiments, we found that the learning difficulty (loss) is varied among languages or between data types, as depicted in Figure 11, Appendix C. We also observe that naive FL works not stably if directly applied with different initializations. Therefore, applying FL on all training samples at the same time will restrain the learning of low-frequent and difficult examples. This feature strengthens the model’s capability of learning cross-lingual supervision. Attention heatmap in Figure 2b also reveals that DA captures stronger cross-lingual context information than MA.

3.3 Language-Adaptive Sample Re-weighting

In addition to concerns on network architecture, data imbalance is another issue that limits the model’s multilinguality. Masked language modeling is a word-level multi-classification problem which consists of vocabulary size V classes. Due to the imbalance nature of word distribution, training examples of each class is extremely imbalance. For example the sentence "The cat went away." appears 29.20M and 1009.06M times in our Wikipedia corpus, while "cat", "went", "away" appear 0.61M, 0.67M, 0.42M times respectively. In this example, the maximum imbalance ratio (IR) is 2403. However, the key semantic information is usually given by less frequent words, e.g. "cat went away" in the example. It is beneficial for the model to focus more on the less frequent words.

Hence, we introduce focal loss (FL Lin et al. 2017) to address the class imbalance issue. It was proposed to tackle class imbalance problem in objective detection. As mentioned above, we formalize the masked language modeling as a word-wise multi-classification task. Then, the Equations 1 and 2 can be modified into

\[ \mathcal{L}_\eta = \alpha (1 - p_t)^\gamma \mathcal{L}_\eta \]  

(6)

Where \( p_t = p(t|C_t), t \in \{x_i, y_j\}, \eta \in \{MMLM, Trans\} \), \( \alpha \) is a hyper-parameter to scale the loss and \( \gamma \) is a hyper-parameter to balance the strength of focus on low-frequent and difficult examples. The higher \( \gamma \) is, the more focus on difficult examples. In our experiments, we set \( \alpha = 0.25 \) and \( \gamma = 2 \) as suggested by Lin et al. (2017). To further understand FL, we refer readers to its original paper (Lin et al. 2017).

According to Equation 6, FL encourages the model to focus more on words of lower prediction probability (higher CE loss). In another word, FL pushes the model to learn more from difficult examples. But in our preliminary experiments, we found that the learning difficulty (loss) is varied among languages or between data types, as depicted in Figure 11, Appendix C. We also observe that naive FL works not stably if directly applied with different initializations. Therefore, applying FL on all training samples at the same time will restrain the learning of low-frequent and difficult samples, such as those from low-resource languages and monolingual corpora.

On basis of above analysis, we design a language-wise and data-type-wise FL to perform language-adaptive sample re-weighting. Briefly speaking, we only apply FL to languages and data types which
have been learned well. Then, Equation 6 is modified into
\[ L_\eta = \alpha (1 - p_t) \gamma_\eta L_\eta \]
\[ \gamma_\eta = \begin{cases} 0 & L_{CE} > L_T \text{ or step } < S \\ \gamma & L_{CE} \leq L_T \text{ and step } \geq S \end{cases} \]

Where \( \gamma_\eta \) is a language-wise FL hyper-parameter, \( L_T \) and \( S \) denote threshold value of training loss and training step respectively. \( \gamma_\eta \) is updated during training for different languages and data types independently. Finally, the objective function of our model becomes
\[ L = \sum_l \sum_\eta \alpha (1 - p_t) \gamma_\eta L_\eta \]

4 Experiments

4.1 Dataset and Preprocessing

4.1.1 Pre-Training

The 12-layer (BASE) models are pre-trained on Wikipedia dumps\(^2\) and OpenSubtitles-2018 (Lison et al., 2018) corpus, which are under CC-BY-SA and BSD-licensed respectively. We collected 59 languages of Wikipedia dumps (20200520 version). We adopted opus-tools\(^3\) to fetch 422 language pairs of OpenSubtitles-2018 translation corpus covering languages in our Wikipedia data. Detailed statistics of our data are given in Appendix E. We extracted pure text from Wikipedia and OpenSubtitles using WikiExtractor\(^4\) and opus-tools respectively. Byte-level Byte-Pair Encoding (BBPE) (Wei et al., 2021a) was applied for tokenization. The vocabulary was created based on our Wikipedia data, consisting of 100618 tokens.

The 6-layer (MINI) models are pre-trained with 0.1% Wikipedia data and 0.01% OpenSubtitles-2018 data. The number of examples of each language pair is greater than one thousand. Other data configuration follows that of the BASE models.

4.1.2 Fine-Tuning

XNLI (Conneau et al., 2018) is a multilingual version of the natural language inference (NLI) task which is translated from MultiNLI (Williams et al., 2018).

PAWS-X (Yang et al., 2019) is a multilingual version of paraphrase identification task which is translated from PAWS-X (Zhang et al., 2019).

TyDiQA-GoldP is a SQuAD-like multilingual question answering task that derived from Clark et al. (2020). Given a context passage and a question, the task is to predict a answer span from the passage. It covers 9 typologically diverse languages.

UDPOS (Nivre et al., 2020) is a POS tagging task which covers 90 languages. In this paper, we follow the setting of Hu et al. (2020) that includes the 29 languages with validation data.

4.2 Training Details

For fair comparison, we train mBERT, MA and the proposed DA models on identical corpora. The motivation of this work is to explore a novel framework to leverage cross-lingual supervision better than the MA module. Most SOTA multilingual models are based on MA and this work is orthogonal to them. Hence, we only reproduce mBERT and MA as our baselines. Experiments are conducted on Nvidia V100 GPUs. Detailed hyper-parameter configuration is given in Appendix D.

4.2.1 Pre-Training

mBERT is trained only on Wikipedia corpus with masked word prediction objective and next sentence prediction (NSP) objective. Language embeddings are not adopted.

---

\(^{2}\)https://meta.wikimedia.org/wiki/List_of_Wikipedias

\(^{3}\)https://pypi.org/project/opustools-pkg/

\(^{4}\)https://github.com/attardi/wikiextractor
MA and DA-series models are trained on a mixture of Wikipedia and OpenSubtitles-2018 data. Both of them are only trained with masked word prediction objective. Batches of Wikipedia and OpenSubtitles-2018 examples are iteratively fed to the model. For each batch, only one type of examples is included. Language embeddings are trained with these models.

4.2.2 Fine-Tuning

All the evaluation of downstream NLU tasks are performed under the setting of zero-shot cross-lingual transfer. In more details, we only fine-tune the models using English training and development set. Then directly evaluate the models on testing set of the target language.

We adopt early stopping as the training stop criteria for all experiments. The training is stopped if without accuracy increase on development set for 50 epochs. In order to reduce randomness, the results are reported as the average score of 8 rounds of experiments with different random seeds.

For sentence pair classification, i.e. XNLI and PAWS-X, two sentences are concatenated as a single sequence before fed into the models. Then the contextual word representation of the [BOS] is used by downstream classification network. All of the models adopt identical classification network, following the original mBERT.

For question answering and structured prediction tasks, i.e. TyDiQA-GoldP and UDPOS, contextual representations of all words at the last layer are extracted as the input to downstream network. In TyDiQA-GoldP, a pair of question and its related answering passage is concatenated as models’ input. The model is trained to label the answer span’s start word and end word. Similarly, model is trained to label POS of each input word in UDPOS task. DA models only adopt IA layers for question answering and structured prediction tasks.

Fine-tuning hyper-parameters of all models are identical. Detailed configuration is provided in Appendix D.

4.3 Experimental Results

In this section, + FL and + Adapt-FL denote DA models trained with naive focal loss and the proposed language-adaptive sample re-weighting respectively. We adopt language code from ISO-639-1 as abbreviation of the tested languages.

4.3.1 Sentence Pair Classification

| Model  | en  | ar  | bg  | de  | el  | es  | fr  | hi  | ru  | sw  | th  | tr  | ur  | vi  | zh  | avg  |
|--------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|------|
| mBERT  | 80.2 | 64.3 | 69.7 | 69.2 | 67.2 | 72.9 | 71.6 | 58.7 | 68.4 | 55.6 | 49.8 | 62.2 | 58.1 | 69.1 | 65.4 | 65.5 |
| MA     | 81.7 | 70.4 | 75.9 | 73.9 | 73.6 | 77.0 | 76.1 | 61.5 | 73.3 | 47.9 | 68.8 | 71.6 | 60.7 | 73.6 | 68.8 | 70.3 |
| DA     | 82.0 | 71.2 | 76.3 | 74.1 | 73.9 | 77.4 | 76.8 | 61.8 | 73.5 | 46.6 | 68.9 | 71.3 | 59.6 | 74.5 | 69.2 | 70.5 |
| + FL   | 81.7 | 71.1 | 76.4 | 74.5 | 74.2 | 77.3 | 76.4 | 62.5 | 73.3 | 48.2 | 69.3 | 71.5 | 60.0 | 73.9 | 69.2 | 70.6 |
| + Adapt-FL | 81.8 | 70.9 | 76.3 | 75.0 | 74.1 | 77.6 | 76.6 | 62.9 | 73.3 | 48.1 | 69.0 | 71.7 | 60.2 | 74.4 | 69.3 | 70.8 |

Table 1: XNLI accuracy scores.

We first evaluate baselines and our models on the widely used multilingual classification task XNLI. As shown in Table 1, DA models consistently perform better than MA on most languages and on average. It demonstrates the effectiveness of the proposed DA module. With the proposed sample re-weighting strategy, DA's performance is further improved. Performance on relatively low-resource languages are consistently improved, e.g. Hindi (hi), Swahili (sw), Turkish (tr) and Urdu (ur).

We also notice that mBERT outperforms both DA and MA on Swahili. It’s because Swahili is not included in the OpenSubtitles-2018 translation corpus, but included in the Wikipedia monolingual corpus. Since DA and MA models are iteratively trained over translation and monolingual corpora, Swahili is underfit by the two models.

In order to diagnose DA's advantage on cross-lingual alignment, we evaluate DA and the baselines on a multilingual paraphrase identification task PAWS-X. Better cross-lingual alignment is supposed to benefit paraphrase identification more than natural language inference, since word alignment in paraphrase pair is stricter. As illustrated in Table 2, DA outperforms the baselines with a large gap on most languages.
Table 2: PAWS-X accuracy scores. *euro*: average score of German (de), Spanish (es) and French (fr). *asia*: average score of Japanese (ja), Korean (ko) and Chinese (zh).

This advantage is more obvious on Asian languages than on European languages. According to Table 2, MA beats mBERT on *euro* but is defeated on *asia*. It implies that given translation corpora, MA is only able to learn cross-lingual alignment from similar languages, such as the ones within the same family of the source language. However, DA is able to learn stronger alignment even between languages with different syntax. We further analyze this feature in Section 5.1. Besides, compared with naive FL, the improvement on *ja, ko and zh* brought by the language-adaptive sample re-weighting demonstrates its effectiveness again, since sample frequency of *asia* is generally lower than *euro* in the training set.

4.3.2 Question Answering

In addition to classification tasks, we evaluate the models on multilingual question answering. Both F1 and exact match (EM) scores are reported in Table 3. DA performs better than MA on all languages in terms of both F1 and EM scores. In this task, MA performs worse even than mBERT, which can be also observed (as XLM) in Hu et al. (2020). We guess the reason for MA's failure is: question answering is a task requires long-distance dependency, the model needs to link the question span (at the beginning of the sequence) to the answer span (in the middle of the sequence). But sentence-level translation corpora (e.g. OpenSubtitles) usually consist of short sequences, which limits the model's long-distance modeling capability, then brings side-effects for question answering. Especially for MA-based models that learn monolingual and translation corpora with mixed IA and CA. However, DA decomposes IA and CA so that IA is not affected by short translation sequences. Hence long-distance modeling capability is retained by DA's IA layers, then question answering can be addressed just as mBERT does. Note that in the experiments, we only adopt DA's IA layers for the question answering task. Besides, due to lack of Swahili translation data, MA and DA perform worse on Swahili again.

4.3.3 Structured Prediction

In addition to classification tasks, we evaluate the models on multilingual question answering. Both F1 and exact match (EM) scores are reported in Table 3. DA performs better than MA on all languages in terms of both F1 and EM scores. In this task, MA performs worse even than mBERT, which can be also observed (as XLM) in Hu et al. (2020). We guess the reason for MA's failure is: question answering is a task requires long-distance dependency, the model needs to link the question span (at the beginning of the sequence) to the answer span (in the middle of the sequence). But sentence-level translation corpora (e.g. OpenSubtitles) usually consist of short sequences, which limits the model's long-distance modeling capability, then brings side-effects for question answering. Especially for MA-based models that learn monolingual and translation corpora with mixed IA and CA. However, DA decomposes IA and CA so that IA is not affected by short translation sequences. Hence long-distance modeling capability is retained by DA's IA layers, then question answering can be addressed just as mBERT does. Note that in the experiments, we only adopt DA's IA layers for the question answering task. Besides, due to lack of Swahili translation data, MA and DA perform worse on Swahili again.

Table 4: Accuracy scores on UDPOS validation set, covering 29 languages.
Finally, we evaluate the models on the multilingual POS task UDPOS. Since DA only adopts IA layers during fine-tuning on UDPOS, the advantage of CA layers’ cross-lingual modeling may not be fully leveraged. Nevertheless, even with only IA layers, DA performs on par with MA as shown in Table 4.

It’s worth noting that both DA and MA are significantly better than mBERT, which implies cross-lingual supervision benefits the multilingual structured prediction task. Hence, it is very promising to further explore how to fully take advantage of DA’s network for structured prediction task.

5 Further Analysis

5.1 Decomposed Attention

Figure 2: Attention probability heatmap of MA and DA. Words attend from row to column. Upper-left: IA of English. Lower-right: IA of Chinese. Upper-right: CA of English-Chinese. Lower-left: CA of Chinese-English. Red blocks: reference alignments.

Figure 2 provides a visual comparison between MA and DA in both intralingual attention (IA) and cross-lingual attention (CA). Attention probabilities are colored at the same scale, [0, 0.15], as shown by the colorbars on the right side. We observe DA’s advantages as follows.

**Balance between IA and CA** As shown in Figure 2a, MA’s attention probabilities on intra-lingual context (upper-left and lower-right) are obviously higher than those on cross-lingual context (upper-right and lower-left). In contrast with MA, Figure 2b reveals significantly more balance attention between intra-lingual context and cross-lingual context, especially on the correctly aligned words. The behaviour of MA is easy to understand since learning alignment from intra-lingual context is easier than from cross-lingual context. However, because DA decomposes IA and CA, the model is forced to learn alignment from cross-lingual context.

**More accurate alignment** Although MA is able to provide correct intra-lingual alignments, it’s not as strong as DA. Besides, intra-lingual alignments are monotonic which is so easy to model that cannot discriminate models’ cross-lingual transferability. Hence, we study the cross-lingual alignment of MA and DA in more details. The example in Figure 2 is challenging for CA, since its word order in English and Chinese are very different. We labeled the reference alignment in Figure 2 in red blocks. To understand Figure 2, we define a correct alignment as the darkest color block in a row which hits the golden alignment word. Then from Figure 2b we can observe that DA captures almost all the cross-lingual alignment accurately, in both English-Chinese and Chinese-English attention directions. Even the long-distance prepositional phrase “中 (in)” is successfully aligned with high probabilities. On the contrary, MA totally fails most cross-lingual alignments, such as “Partial remains”, all “of’s, “cities” etc.
In more case studies, we find DA is especially powerful in capturing entities. Examples are provided in Appendix A. Figure 3 shows that DA highlights the name Maud Gonne, and also successfully captures the alignment “s' (的) continuous (无数次) 拒绝 (refusal)” while MA fails.

In conclusion, MA pay more attentions on intra-lingual context and cannot accurately capture cross-lingual alignment. MA tends to simply provide monotonic cross-lingual alignment, like the intra-lingual alignment. This issue may be due to its nature of mixed intra-lingual and cross-lingual attentions. The proposed DA network tackles MA’s defects and highly enhances model’s capability in learning cross-lingual supervision.

5.2 Effect of Model Size

| Model     | #Parameters |
|-----------|-------------|
| MA        | 114.4 M     |
| DA        | 117.8 M     |
| DA-reduce | 114.6 M     |
| DA-share  | 114.4 M     |

(a) MINI: 6 Transformer layers.

| Model     | #Parameters |
|-----------|-------------|
| MA        | 154.9 M     |
| DA        | 161.7 M     |

(b) BASE: 12 Transformer layers.

Table 5: Comparison of model size.

Although CA in the proposed model does not adopt attention projection matrices $W_Q^{CA}$, $W_K^{CA}$, $W_V^{CA}$, as shown in Equation 5, CA keeps parameters of the feedforward (FFN) layer and the subsequent layer normalization (LayerNorm). It slightly improves number of DA’s parameter. Table 5 illustrates the number of parameters of MA and DA models. In order to verify the effect of model size and exactly align DA’s model size to MA’s, we propose two methods to reduce DA’s model size: 1) DA-reduce: reduce the hidden dimension of Transformer layers from 768 to 720, but keep the dimension of embedding layer unchanged. 2) DA-share: share FFN and LayerNorm parameters between IA and CA. Take MINI models for example, as shown in Table 5a, number of parameters of DA-reduce and DA-share are controlled identical to the MA counterpart.

| Model           | en | ar | bg | de | el | es | fr | hi | ru | sw | th | tr | ur | vi | zh | avg |
|-----------------|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|-----|
| MA              | 71.2 | 53.6 | 56.8 | 56.9 | 54.9 | 59.7 | 59.4 | 51.2 | 55.4 | 43.0 | 53.4 | 51.4 | 51.3 | 57.5 | 52.5 | 55.2 |
| DA              | 74.2 | 54.4 | 58.9 | 59.5 | 55.0 | 63.4 | 61.3 | 53.5 | 57.8 | 39.4 | 53.4 | 51.3 | 52.8 | 58.3 | 52.0 | 56.3 |
| DA-reduce       | 74.0 | 54.8 | 59.7 | 60.1 | 57.0 | 63.6 | 62.3 | 53.3 | 58.2 | 41.2 | 54.0 | 53.2 | 52.8 | 59.1 | 51.9 | 57.0 |
| DA-share        | 73.7 | 55.2 | 58.9 | 59.1 | 57.4 | 62.9 | 61.8 | 53.1 | 57.4 | 41.8 | 54.9 | 52.0 | 52.7 | 58.4 | 51.9 | 56.8 |

Table 6: XNLI accuracy scores.

We further pre-train MINI models of MA, DA, DA-reduce and DA-share, then evaluate them on XNLI. The MINI pre-training configuration is described in Section 4.1.1. Fine-tuning setting follows the BASE models, as introduced in Section 4.2. The results in Table 6 demonstrate that the DA outperforms MA even with identical number of parameters.

6 Conclusions and Future Directions

In this paper, we explore the potential of making better use of cross-lingual supervision for learning multilingual representation in two aspects. In terms of network architecture, we propose a decomposed attention (DA) module equipped with independent intra-lingual attention (IA) and cross-lingual attention (CA). Compared to conventional, widely-used mixed attention (MA) network, DA significantly improves models’ learning ability on cross-lingual supervision. We further propose a language-adaptive sample re-weighting strategy for learning multilingual representation. The cross-lingual transferability of the model is further boosted by the re-weighting strategy. Experiments on various cross-lingual NLU tasks and further analysis prove the effectiveness of the proposed methods.

To our surprise, both DA-reduce and DA-share improves the performance of DA. The results imply that: 1) linear projection between embeddings and Transformer hidden representations may enhance model capability, 720 hidden dimensions is large enough for MINI models. 2) the shared FFN and LayerNorm encourage IA and CA to build a language-agnostic semantic space. We leave these conjectures for future research.
However, there are still questions required future research. First, as mentioned in Section 4.3 currently we only adopt DA's IA layers for structured prediction task. There should be more effective ways to explore DA's potential in this task. Another direction may be related to the re-weighting strategy. Currently, the its hyper-parameters is empirically set. We are curious if there is a theoretical method to decide it directly, without prior information of the training data. There are many interesting topics on multilingual representation learning, we leave these topics for future works.

References

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. Transactions of the Association for Computational Linguistics, 8:454–470.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In Advances in Neural Information Processing Systems, volume 32, pages 7059–7069. Curran Associates, Inc.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yinpeng Guo, Yi Liao, Xin Jiang, Qing Zhang, Yibo Zhang, and Qun Liu. 2019. Zero-shot paraphrase generation with multilingual language models. CoRR, abs/1911.03597.

Dan Hendrycks and Kevin Gimpel. 2016. Bridging nonlinearities and stochastic regularizers with gaussian error linear units. CoRR, abs/1606.08415.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Proceedings of the 37th International Conference on Machine Learning, volume 119 of Proceedings of Machine Learning Research, pages 4411–4421. PMLR.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Uncoder: A universal language encoder by pre-training with multiple cross-lingual tasks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2485–2494, Hong Kong, China. Association for Computational Linguistics.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. 2017. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV).

Pierre Lison, Jorg Tiedemann, and Milen Kouylekov. 2018. OpenSubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation.
Fuli Luo, Wei Wang, Jiahao Liu, Yijia Liu, Bin Bi, Songfang Huang, Fei Huang, and Luo Si. 2021. {VECO}: Variable encoder-decoder pre-training for cross-lingual understanding and generation.

Lei Mao. 2019. Use Focal Loss To Train Model Using Imbalanced Dataset. https://leimao.github.io/blog/Focal-Loss-Explained/ [Online; accessed 30-Jun-2019].

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In Proceedings of the 12th Language Resources and Evaluation Conference, pages 4034–4043, Marseille, France. European Language Resources Association.

Xuan Ouyang, Shuohuan Wang, Chao Pang, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-m: Enhanced multilingual representation by aligning cross-lingual semantics with monolingual corpora.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. CoRR, abs/1706.03762.

Junqiu Wei, Qun Liu, Yinpeng Guo, and Xin Jiang. 2021a. Training multilingual pre-trained language model with byte-level subwords.

Xiangpeng Wei, Yue Hu, Rongxiang Weng, Luxi Xing, Heng Yu, and Weihua Luo. 2021b. On learning universal representations across languages. In International Conference on Learning Representations.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mt5: A massively multilingual pre-trained text-to-text transformer.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A cross-lingual adversarial dataset for paraphrase identification. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. 2020. Large batch optimization for deep learning: Training bert in 76 minutes. In International Conference on Learning Representations.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.
A Attention Heatmaps

In this section, we provide more attention examples to reveal the advantage of DA over MA.

A.1 DA Highlights Entities

(a) Mixed attention.  
(b) Decomposed attention.  

Figure 3: Alignment of entity, name *Maud Gonne*.

(a) Mixed attention. 
(b) Decomposed attention. 

Figure 4: Alignment of entity, country *America*.

A.2 When Word Order is Different
(a) Mixed attention.  
(b) Decomposed attention.

Figure 5: Alignment of entity, currency $50,000.

(a) Mixed attention.  
(b) Decomposed attention.

Figure 6: Alignment of entity, time 21 days, 8 hours, 16 minutes and 24 seconds.
(a) Mixed attention. 

(b) Decomposed attention.

Figure 7

(a) Mixed attention. 

(b) Decomposed attention.

Figure 8
Figure 9

(a) Mixed attention.

(b) Decomposed attention.
B Multilingual Maksed Language Modeling

Figure 10: Multilingual masked language modeling (MMLM).

C Training Loss of Preliminary Experiment

Figure 11: Training CE loss of languages from different language families. Solid: loss on monolingual data. Dashed: loss on cross-lingual data. The experiment is conducted with MINI model as described in Section 4.1.1.
## D Hyper-Parameters

| Hyper-parameter          | Value                  |
|-------------------------|------------------------|
| Hidden size             | 768                    |
| FFN projection          | 3072                   |
| Positional embedding length | 128                 |
| Warmup steps            | 1800                   |
| Optimizer               | LAMB [You et al., 2020]|
| LAMB $\beta_1$         | $1e^{-4}$              |
| LAMB $\beta_2$         | 0.9                    |
| LAMB $\beta_3$         | 0.999                  |
| Learning rate           | $5e^{-5}$              |
| Dropout probability     | 0.1                    |
| Activation function     | GELU [Hendrycks and Gimpel, 2016] |
| Batch size              | 2048                   |
| Attention layers        | 12                     |
| Attention heads         | 12                     |
| Training steps          | 2011K                  |
| Re-weighting $L_T$      | 1.6                    |
| Re-weighting $S$        | 2000K                  |

Table 7: Hyper-parameters of pre-training BASE models.

## E Dataset Statistics

Figure 12: Statistics of Wikipedia corpus in our experiments.
Figure 13: Statistics of OpenSubtitles-2018 corpus in our experiments.