Relaxed Attention for Transformer Models

Timo Lohrenz, Björn Möller, Zhengyang Li, Tim Fingscheidt
Institut for Communications Technology
Technische Universität Braunschweig
Braunschweig, Germany
{t.lohrenz, b.moeller, zhengyang.li, t.fingscheidt}@tu-braunschweig.de

Abstract—The powerful modeling capabilities of all-attention-based transformer architectures often cause overfitting and—for natural language processing tasks—lead to an implicitly learned internal language model in the autoregressive transformer decoder complicating the integration of external language models. In this paper, we explore relaxed attention, a simple and easy-to-implement smoothing of the attention weights, yielding a twofold improvement to the general transformer architecture. First, relaxed attention provides regularization when applied to the self-attention layers in the encoder. Second, we show that it naturally supports the integration of an external language model as it suppresses the implicitly learned internal language model by relaxing the cross attention in the decoder. We demonstrate the benefit of relaxed attention across several tasks from different applications with clear improvement in combination with recent benchmark approaches using various transformer model variants and sizes. Specifically, we exceed the former state-of-the-art performance of 26.90% word error rate on the largest public lip-reading LRS3 benchmark with a word error rate of 26.31%, as well as we achieve a top-performing BLEU score of 37.67 on the IWSLT14 (DE → EN) machine translation task without external language models and virtually no additional model parameters.

Index Terms—transformer, attention, relaxation, regularizer, internal language model

I. INTRODUCTION

Early encoder-decoder models emerged from machine translation, where the encoder compressed the entire source language sentence into a fixed-length embedding vector [1]. This is particularly difficult for very long sentences [2], as the fixed-length embedding vector is only a limited-capacity representation. The use of attention, introduced in [3], enabled the computation of variable-length weight distributions over the input sequence and soon turned out to be advantageous for far more applications than just neural machine translation (NMT), e.g., automatic speech recognition (ASR) [4]–[6], language modeling and understanding [7], object detection [8], and image classification [9], [10]. Soon the most prominent attention-based encoder-decoder (AED) model emerged, namely the transformer [11] model. Without the use of any recurrence, it entirely relies on self-attention in the encoder to model temporal dependencies in the input and cross attention in the decoder to extract relevant timesteps thereof during the autoregressive decoding process. While transformers in language modeling tasks are well-suited for upsampling the model size and depth without any saturation when large amounts of data are present [7], [12]–[14], they are also susceptible to overfit and require strong regularization to learn at all [15]–[17]. In our previous study exclusively on ASR [18], we showed that regularization by smoothing the attention weights in the decoder’s cross attention, dubbed relaxed attention, improves performance when the transformer model is combined with an external language model (LM) but, for reasons yet to be explored, does not help without a LM.

In this work, we take on the idea of relaxed attention to expand it to self-attention in the encoder layers, providing regularization already in the transformer encoder. Thereby, we increase the method’s versatility as it becomes applicable to encoder-only transformers, which are common in several non-sequence tasks such as image classification or pre-trained bidirectional encoder representation by transformer (BERT) [7] models. Our main contributions are summarized as follows:

• We introduce relaxed self-attention in the transformer encoder to improve generalization and develop fuzzy relaxation as a variant thereof.

• Beyond relaxed self-attention, we extensively investigate the capability of relaxed cross attention in the decoder of sequence-to-sequence transformer models and show that the improvement is due to better external language model integration as it suppresses the influence of the internal language model.

• We show improvements of both relaxed attention variants on a variety of tasks including automatic speech recognition, lip-reading, machine translation, and image classification. On the lip-reading and machine translation task we report a new state of the art and top-performing result, respectively.

The paper is structured as follows: After a summary of related work and a brief overview of the original transformer model in Sections II and III, we introduce the relaxed attention approach in Section IV, followed by the experimental evaluation including results and discussion in Section V. Section VI concludes the paper.

II. RELATED WORK

a) Regularization of Transformer Models: In this work, we introduce a regularization method to the self-attention function [11], which is fundamental to transformer models. Several regularization approaches proposed for such networks in the past are related to the network output of transformer
When attention-based encoder-decoder networks were first used to prevent too narrow attention weight distributions (RNN)-based AED models, we include it as a reference method only on text transcriptions from the acoustic training data. They compressed the probability-like outputs, but didn’t take into account the input sequence length, despite the authors’ observation that longer sentences require less smoothing of the attention weights. Even though this method dubbed smooth focus was so far only applied to recurrent neural network (RNN)-based AED models, we include it as a reference method in our simulations as it is the closest to our relaxed attention approach.

b) Internal Language Model Handling: For many sequence-to-sequence tasks the integration of language models (LMs) to AED models is of dual use: First, LMs leverage the use of large amounts of additional text-only data to improve performance. Second, LMs can be utilized to adapt acoustic models to domains which differ from the original acoustic training data. Several techniques exist to combine models to domains which differ from the original acoustic training data domain. Second, LMs can be utilized to adapt acoustic models to domains which differ from the original acoustic training data domain. Several techniques exist to combine models to domains which differ from the original acoustic training data domain. Second, LMs can be utilized to adapt acoustic models to domains which differ from the original acoustic training data domain.

In our previous work in [18], relaxed attention was used to prevent too narrow attention weight distributions in the cross attention during training which only yielded improvements with an external language model in ASR. Here, however, we apply this approach to the self-attention function to reduce over-fitting already in the encoder and investigate if relaxed self-attention also helps when applied during both training and test (matched inference). In addition we include a variety of the aforementioned—proven to be effective—regularization methods as baselines and show that relaxed attention is able to further improve performance yielding complementarity to other regularization methods. When attention-based encoder-decoder networks were first applied to ASR, [4] proposed a modified softmax function to smooth the attention weights in the cross attention between encoder and decoder by replacing the exponential function in the standard softmax function with a sigmoid. Thereby, they compressed the probability-like outputs, but didn’t take into account the input sequence length, despite the authors’ observation that longer sentences require less smoothing of the attention weights. Even though this method dubbed smooth focus was so far only applied to recurrent neural network (RNN)-based AED models, we include it as a reference method in our simulations as it is the closest to our relaxed attention approach.

III. THE ENCODER-DECODER TRANSFORMER MODEL

In this section we briefly recapitulate the original transformer architecture from [11] consisting of encoder and decoder as shown in Figure 1. Please note that here we describe the transformer architecture exactly as used for the investigated sequence-to-sequence tasks (i.e., automatic speech recognition, lip-reading, and machine translation). In our simulations we investigate the hypothesis that relaxed cross attention successfully suppresses the internal language model, which in contrast to the aforementioned methods does not—apart from a single hyperparameter—require any additional models [34], parameters [33], or adaptation trainings [35], but weakens the internal language model during training of the transformer thus supporting shallow fusion [32]. In addition, we will introduce relaxed self-attention, which improves performance in many applications even without the use of an explicit LM.
V. PROPOSED RELAXED ATTENTION

Scaled dot-product multi-head attention (see Figure 3) is typically used in two variants in the encoder-decoder transformer model [11]: First, it is used in the encoder as self-attention to model positional (e.g., temporal) dependencies in the preprocessed input sequence indexed with $t \in \{1, \ldots, T\}$ of length $T$. Second, it is used in the decoder as cross attention (also often referred to as encoder-decoder or source target attention), which draws the decoder’s attention to relevant parts in the encoded input sequence $h^T_t \in \mathbb{R}^{T \times d}$ for decoding at output sequence index $\ell \in \{1, \ldots, L\}$ with model dimension $d$. In case of self-attention\(^1\), all multi-head attention inputs (key $K$, value $V$, query $Q$) are the same, i.e., $K = V = Q$, with query input $Q \in \mathbb{R}^{L \times d}$ of length $\tilde{L} = T$. For cross attention, key and value inputs, $K \in \mathbb{R}^{T \times d}$ and $V \in \mathbb{R}^{T \times d}$, respectively, stem from the encoder output $h^T_t$ yielding $K = V = h^T_t$, while the query input $Q$ comes from the previous decoder layer with $\tilde{L} = 1$ during inference and $\tilde{L} = L$ during training, where for the latter all $L$ tokens of the target sequence are processed in parallel. The attention weights $G_i(Q, K) \in \mathbb{I}^{L \times T}$ for the scaled dot-product multi-head attention sum up to one across the query input length $\tilde{L}$ after the softmax activation function and thus can be interpreted as a probabilistic weighting applied to the value input projection $Y_i \in \mathbb{R}^{T \times d/N_h}$, with $N_h$ being the number of attention heads each indexed with $i \in \mathcal{N}_h$.

Relaxed attention follows the basic principle of regularization by introducing some stress into the training process. Shown as red box in Figure 3 it modifies the standard attention weights $G_i(Q, K)$ that draw attention to the encoded input sequence. Our relaxed attention weights for scaled dot-product attention are defined as simple as

$$\tilde{G}_i(Q, K) = \left(1 - \gamma\right)G_i(Q, K) + \gamma \frac{1}{T}, \quad i \in \mathcal{N}_h,$$  \hfill (1)

gradually injecting a uniform distribution (with $1$ here being an $L \times T$ matrix of ones), thereby smoothing the distribution of attention weights and imposing some of the attention weight mass across the entire input sequence. Please note that the normalizing division by $T$ in equation (1) is different for each file and minibatch for sequence tasks due to varying input sequence lengths. This provides a natural and rich variation

\(^1\)Note that masked self-attention is also used in the decoder to attend to prefix output tokens (cf. decoder blocks in Fig. 1 or [11]), while in this work we focus on the multi-head attention variants that attend to the input domain.
of the effective height of the injected uniform distribution and prevents compensation of the relaxation by the learning process. The injection of the uniform distribution is controlled by a relaxation coefficient $\gamma \in [0, 1]$, which is a constant single hyperparameter for all respective attention layers.

While in our previous work [18] we only investigated relaxed cross attention, only for automatic speech recognition and only during training, in our work (i) we propose relaxed cross attention and self-attention, (ii) during training and during inference (matched inference), (iii) we investigate their application to automatic speech recognition, lip-reading, machine translation, and image classification, and (iv) we introduce fuzzy relaxation for the image classification task, where we randomly draw the relaxation coefficient from a normal distribution $\gamma \sim \mathcal{N}(x; \mu = \gamma_0, \sigma^2)$, with the initially set $\gamma_0$ being the mean $\mu$. For this specific task, the variable sequence length $T$ in equation (1) is substituted by a constant number of image patch tokens $M^2$, thereby omitting the aforementioned natural variation. Fuzzy relaxation re-establishes this variation of the relaxation by randomizing $\gamma$ during training, while for the matched inference case, the relaxation coefficient is kept fixed at $\gamma = \mu \gamma_0$ during inference.

V. EXPERIMENTAL VALIDATION AND DISCUSSION

A. Application to Automatic Speech Recognition

a) Task and Datasets: Automatic speech recognition transforms recorded speech signals into a sequence of text tokens. We investigate our relaxed attention method on the Librispeech dataset [37] with the clean and other conditions of the dev and test subsets. We measure system performance in terms of word error rate $\text{WER} = 1 - \frac{N - D - I - S}{N}$, depending on the number of words $N$, deletions $D$, insertions $I$ and substitutions $S$. All raw speech signals are sampled at 16 kHz and analyzed with a 25 ms window at a frame shift of 10 ms. As common in ASR, we also use an external language model trained on the text labels of the 960h training set as well as on the text-only Librispeech language model training corpus, the latter containing sentences from a total amount of 14,500 books from project Gutenberg [37] which are accessible under public-domain. The Librispeech ASR corpus is available under the very permissive Creative Commons BY 4.0 license.

b) Models and Training: For training with 100h and 960h of training data, we trained standard encoder-decoder transformer models [11] from scratch in the small and base configuration, comprising 19.3M and 69.8M parameters, respectively. As common for ASR, filterbank features are extracted for each time frame $t$ and then preprocessed by a four-layer convolutional neural network, each using $3 \times 3$ filter kernels (cf. preprocessing block in Figure 1, Section III). All hyperparameters were set according to the recipes available in the fairseq based espresso toolkit$^2$ [38] except the relaxation coefficients $\gamma$, which have been tuned on the joint clean and other portions of the dev set for both, relaxed cross attention and relaxed self-attention. As additional regularization we use

SpecAugment [39], label smoothing [19] and dropout [27] during training.

c) Results and Discussion: For both, the 100h and 960h training data cases in Table I, the resimulated baselines (using training scripts from [38]) yield similar results as in [18] using a standard transformer approach. The smooth focus method [4] has a higher WER compared to the baseline on the small training data case, but yields small improvements on some clean settings for the 960h training data case. Compared to smooth focus, relaxed self- and cross attention adapt to the length $T$ of the input sequence, with the latter yielding solid WER reduction across all dev and test conditions when an LM is used (right-hand side of Table I), thereby confirming the results of [18]. In Section V-E, we show that the strong improvement with LM using relaxed cross attention is due to improved internal language model suppression. Without an LM, both the resimulated baseline and relaxed cross attention approaches are outperformed by our new relaxed self-attention in all dev and test conditions for both training data cases. Specifically, the WER across the test conditions of the 960h case for relaxed self-attention improved by a relative 9% (clean) and 5% (other) compared to the resimulated baseline, yielding complementary regularization of our method to the other employed regularization methods. Note that in all aforementioned cases, relaxed attention is best when used only in training. Only in a very specific case on the dev set, however, “matched inference”, i.e., relaxed self-attention in training and test, is slightly ahead of using it in training only. In Sections V-F and V-G, we provide an ablation study on the related attention dropout and investigate initialization seed robustness, respectively.

\begin{table}[h]
\centering
\caption{Automatic speech recognition results in terms of WER (\%) on the Librispeech task using standard encoder-decoder transformer models. Attention relaxation is applied in training only, except for “matched inference” (attention relaxation in training and test). We separately use the 100h and 960h training datasets and highlight the respective best results for each size in \textbf{bold} font.}
\begin{tabular}{lcccccc}
\hline
\textbf{Approach} & \multicolumn{2}{c}{\textbf{without LM}} & \multicolumn{2}{c}{\textbf{with LM}} \\
 & \textbf{dev} & \textbf{test} & \textbf{dev} & \textbf{test} \\
\hline
\textbf{100h training data} & & & & & & \\
Baseline [18], resim.) & 13.98 & 28.71 & 14.82 & 29.31 & 10.62 & 24.19 & 12.06 & 25.56 \\
+ smooth focus [4] & 14.60 & 28.73 & 15.50 & 30.78 & 10.83 & 24.86 & 12.11 & 26.46 \\
+ matched inference & 13.91 & 28.70 & 14.70 & 30.10 & \bf{9.33} & \bf{22.16} & \bf{10.62} & \bf{23.04} \\
+ relaxed cross attention & 14.30 & 29.03 & 15.15 & 30.09 & 11.04 & 25.19 & 12.16 & 26.36 \\
+ matched self-attention & 13.48 & \bf{27.87} & \bf{14.20} & \bf{28.96} & 10.22 & 23.53 & 11.04 & 24.55 \\
+ matched inference & \bf{13.43} & 28.00 & 14.46 & 29.23 & 10.01 & 23.96 & 11.19 & 25.32 \\
\hline
\textbf{960h training data} & & & & & & \\
Baseline [18], resim.) & 3.92 & 9.00 & 4.47 & 9.23 & 3.73 & 8.52 & 4.40 & 8.95 \\
+ smooth focus [4] & 4.11 & 9.42 & 4.35 & 9.63 & 3.70 & 9.18 & 4.31 & 9.33 \\
+ matched inference & 3.95 & 9.33 & 4.28 & 9.45 & \bf{3.44} & \bf{7.74} & \bf{3.58} & \bf{8.35} \\
+ relaxed cross attention & 3.96 & 9.29 & 4.20 & 9.40 & 3.69 & 8.95 & 4.21 & 9.46 \\
+ matched self-attention & \bf{3.82} & \bf{8.50} & \bf{4.05} & \bf{8.71} & 3.52 & 8.03 & 4.17 & 8.51 \\
+ matched inference & \bf{3.79} & 9.12 & 4.09 & 9.07 & 3.35 & 8.28 & 3.91 & 8.50 \\
\hline
\end{tabular}
\end{table}

$^2$ASR training recipes at \url{https://github.com/freewym/espresso}
TABLE II: Automatic lip-reading results in terms of WER (%) on the LRS3 task using various sequence-to-sequence topologies (top segment basalins) or AV-HuBERT encoders (lower three segments) pre-trained on unlabeled English data from Voxceleb2 and fine-tuned with a joint transformer decoder on the given amount of fine-tuning training data. We also use self-training (bottom segment) by creating pseudo-labels for the 1,326 h of unlabeled data and using these for fine-tuning. Attention relaxation is applied in training only, except for “matched inference” (relaxation in training and test). Best results for each of the three fine-tuning setups are in **bold** font.

| Labeled data (fine-tuning) | Approach | without LM | with LM |
|---------------------------|----------|------------|--------|
|                           |          | dev       | test   | dev   | test |

**Sequence-to-sequence baselines** without unlabeled data

1,362 h + 157 h
- Afouras et al. [40], 2018 — 59.90 — 58.90
- Ma et al. [42], 2021 — 46.90
- Ma et al. [42], 2021 — 43.30

33,000 h
- Makino et al. [43], 2019 — 33.60
- with 334 h of unlabeled data

433 h
- Afouras et al. [44], 2020 — 59.80

**Pre-trained AV-HuBERT encoders** with 1,326 h of unlabeled data

30 h
- Baseline (Shi et al. [45]) — 46.10 —
- Baseline (45), resimulated 47.36 45.90 47.61 45.33
- + smooth focus (4) 47.08 45.80 46.69 45.38
- + relaxed cross attention 45.92 44.00 45.11 42.68
- + matched inference 46.55 45.25 46.46 45.39
- + relaxed self-attention 46.90 45.47 46.95 44.64
- + matched inference 46.85 45.04 46.68 44.69

433 h
- Baseline (Shi et al. [45]) — 28.60 —
- Baseline (45), resimulated 21.90 20.52 21.61 28.97
- + smooth focus (4) 21.87 20.25 21.29 28.86
- + relaxed cross attention 22.12 20.99 21.05 28.05
- + matched inference 21.11 20.29 21.55 28.55
- + relaxed self-attention 21.89 20.96 21.25 28.55
- + matched inference 21.86 20.84 21.24 28.48

1,326 h
- Baseline (Shi et al. [45]) — 26.90 —
- Baseline (45), resimulated 17.92 26.73 17.18 26.50
- + smooth focus (4) 17.42 26.78 17.22 26.29
- + relaxed cross attention 17.40 26.57 16.92 25.51
- + matched inference 17.71 26.43 17.48 25.95
- + relaxed self-attention 17.54 26.31 17.12 26.17
- + matched inference 17.65 26.40 17.16 26.06

B. Application to Lip-Reading

a) Task and Datasets: Automatic lip-reading strives to extract spoken text from an image sequence of talking face recordings. We evaluate lip-reading performance in terms of WER on the test partition of the Lip Reading Sentences 3 (LRS3) dataset consisting of a total of 1,321 recorded videos of English utterances sourced from TED talks [46]. To investigate the performance of the relaxed attention approach on recently successful self-supervised learning approaches, we closely follow the training setup from [45] and use audio-visual hidden unit BERT (AV-HuBERT) encoder models pre-trained on the English subset of the Voxceleb2 dataset [47], containing a total amount of 1,326 hours of unlabeled video recordings. For some experiments we also use an external language model trained on the joint text data from LRS3 and the Librispeech language model training text corpus. LRS3 is publicly available under the TED terms of use as well as the Creative Commons BY-NC-ND 4.0 license.

b) Models and Training: We use AV-HuBERT models\(^3\), introduced recently by [45], which receive image and acoustic frames for pre-training by unlabeled training data to iteratively learn contextualized feature representations \(h^T\). For fine-tuning and inference, only the video input is used and preprocessed (cf. preprocessing layer in Figure 1, Section III) with a 3D convolutional layer and a subsequent ResNet-18 [48], [49] architecture. The models fine-tuned on 30 h of LRS3 training data use the base configuration of the downloaded AV-HuBERT encoder and have a total of 160M parameters. Models fine-tuned on 433 h of LRS3 training data (with or without self-training) use the large AV-HuBERT encoder and comprise 477M parameters in total. As additional regularization methods we use label smoothing [19], LayerDrop [30], as well as dropout [27]. For final experiments, we use the self-training [50] method, where an AV-HuBERT model fine-tuned on 433 h of LRS3 training data is inferred to generate pseudo-labels for the 1,326 h of unlabeled Voxceleb2 data. These were then used together with the true labels from the LRS3 training data to fine-tune the pre-trained AV-HuBERT model. Relaxed attention was only used during this final fine-tuning, and relaxation coefficient \(\gamma\) of each relaxed attention approach was optimized on the development set for each corresponding amount of fine-tuning data.

c) Results and Discussion: The upper segment of Table II shows various baselines on LRS3, whereby [43] reached 33.60% WER w/o LM, using 33,000 h of YouTube training data, and [42] achieved 43.30% with LM and 157 h of additional data from the Lip Reading in the Wild dataset [51]. By leveraging pre-training of AV-HuBERT models, [45] report state of the art so far on LRS3 in three cases with 1,326 h unlabeled pre-training data plus 30 h, plus 433 h, plus 433 h + 1,326 h of fine-tuning data, respectively, the latter using self-training to leverage the pre-training data using pseudo-labels. See also our resimulated numbers of that approach. Note that as it is common practice on the LRS3 task to not even report performance on dev condition, we also formulate performance claims on the test set. Smooth focus [4] helps a bit in 4 out of the 6 total test conditions. Without a language model—adding virtually no parameters and only marginally more complexity during training—our relaxed self-attention achieves WERs of 45.04% vs. 45.93% from [45], resimulated, and 28.84% vs. 29.52% from [45], resimulated, in the 30 h and 433 h fine-tuning cases, respectively, with matched inference (relaxation in training and test). With self-training (433 h + 1,326 h), relaxed self-attention without matched inference even achieves 26.31% WER compared to the best lip-reading WER of 26.90% from [45] thus setting

---

\(^3\)Pre-trained AV-HuBERT models and fine-tuning code downloaded from https://github.com/facebookresearch/av_hubert
TABLE III: Neural machine translation results in terms of BLEU scores on the IWSLT14 task (DE → EN) using encoder-decoder transformer models with cutoff augmentation [23]. Attention relaxation is applied in training only, except for "matched inference" (attention relaxation in training and test). Best results across all approaches are in bold font, second best underlined.

| Approach                        | without LM | with LM (training transcripts only) | with extended LM (additional data) |
|---------------------------------|------------|-------------------------------------|----------------------------------|
|                                 | test       | test                                | test                             |
| Vaswani et al. [11], 2017       | 34.40      | —                                   | —                                |
| Fan et al. [30], 2020           | 34.50      | —                                   | —                                |
| Wu et al. [52], 2019            | 35.20      | —                                   | —                                |
| Wu et al. [53], 2021            | 36.88      | —                                   | —                                |
| Liang et al. [22], 2021         | 37.25      | —                                   | —                                |
| Shen et al. [23], 2020          | 37.60      | —                                   | —                                |
| Baseline ([23], resimulated)    | 37.42      | 37.42                               | 37.62                            |
| + smooth focus ([41])           | 37.42      | 37.52                               | 37.67                            |
| + relaxed cross attention       | 37.56      | 37.53                               | 37.96                            |
| + matched inference             | 37.60      | 37.64                               | 37.57                            |
| + relaxed self-attention        | 37.57      | 37.49                               | 37.74                            |
| + matched inference             | 37.67      | 37.67                               | 37.71                            |

A new state of the art for LRS3. With an additional LM, similar to the ASR task in Section V-A, relaxed cross attention yields consistent improvement on the test set compared to the resimulated baseline in all three fine-tuning cases (i.e., 42.68% vs. 45.33%, 28.05% vs. 28.97%, and 25.51% vs. 26.50%, respectively). We show that this is also caused by the improved internal language model handling for this task in Section V-E and investigate robustness towards different initialization seeds in Section V-G.

C. Application to Machine Translation

a) Task and Datasets: Neural machine translation (NMT) models use neural networks that translate an input text sequence from a source language to a different target language. For our particular experiments on relaxed attention we use data from the well-known IWSLT14 translation challenge [54], choosing the German-to-English (DE → EN) subtask and report performance in terms of BLEU scores [55]. For training of an external LM we either use the IWSLT14 target language transcripts (160k utterances) or the MuST-C dataset, the latter contains 47% additional transcripts (236k utterances) from TED talks and is available under the Creative Commons BY–NC–ND 4.0 international license [56].

b) Models and Training: For training we use the standard encoder-decoder transformer model from [11] in the base configuration with 36.7M parameters and apply cutoff augmentation, which first randomly masks input positions and feature dimensions of the embedded input tokens and second uses a divergence loss to minimize the difference in predictions for different input masks4 [23]. The joint dictionary for source and target language comprises 10k tokens generated with SentencePiece [36] and preprocessed with an embedding layer (cf. preprocessing layer in Figure 1, Section III). As in the previous tasks, to investigate relaxed attention with LM, we trained two transformer LMs of equal size: One LM trained with IWSLT14 training transcripts and an extended LM trained on the MuST-C dataset, respectively. For both, relaxed cross attention and relaxed self-attention, the relaxation coefficient γ has been tuned on the development set.

c) Results and Discussion: In the upper segment of Table III, we report BLEU scores for recent transformer-based approaches to NMT, whereof we choose the strong approach from [23] using cutoff augmentation as a baseline and report a somewhat lower BLEU score of 37.42 in our resimulation. Smooth focus here achieves comparable performance to that baseline with small gains when LMs are used. We observe that a LM trained only with the target language training transcripts of the translation model yields no additional information compared to the internally learned language model and thus does not improve performance for most approaches, even the relaxed cross attention that has been strong (with LM) in previous tasks. However, in case of a strong extended LM trained with additional data, relaxed cross attention (only during training again) yields the best performance of 37.96 BLEU, as it suppresses the internal LM. The best performance for the common case without LM is achieved with our relaxed self-attention approach applied during training and test, slightly outperforming the previous state-of-the-art BLEU score without additional training data (37.60, [23]), with a score of 37.67, exceeding the resimulated baseline even by 0.25 BLEU. We note, that in [57] (only available as preprint) the authors also chose the model of [23] as baseline but were able to reproduce the result of 37.60 BLEU. They report a BLEU score of 37.78 by simply applying a modified learning rate schedule achieving a somewhat smaller improvement of 0.18 BLEU absolute vs. their baseline. Without claiming a new state of the art, we note that both, our and their method are top-performing on the IWSLT14 task. An ablation study on attention dropout is provided in Section V-F and we show robustness of the best relaxed self-attention method towards different initialization seeds in Section V-G.

D. Application to Image Classification

a) Task and Datasets: Image classification is a fundamental task in computer vision aiming at recognizing the primary content of images and differs significantly from the previous sequence-to-sequence tasks as it uses a type of an attention-based encoder-only (decoder-less) transformer model, which recently dominate vision benchmarks. To investigate if relaxed attention is also applicable to such tasks, we evaluate performance in terms of classification accuracy (%) on the computationally less demanding CIFAR-100 dataset [58]. For each of its 100 classes, it contains 500 and 100 images for training and test, respectively, and is publicly available without a specified license. As initialization, we use a model pre-trained on the ImageNet-1k dataset [59], which contains 1.28M training images from 1,000 classes and is also available for research purposes upon agreement of the terms of access.

4Code available from https://github.com/dinghanshen/cutoff
TABLE IV: Image classification results in terms of accuracy (%) on the CIFAR-100 task using encoder-only transformer models. Relaxed self-attention is applied in training only, except for "matched inference" (relaxation in training and test). All reference methods have roughly the same model size and complexity. Best results across all Swin-T approaches are in **bold** font, second best underlined.

| Approach | w/ pre-training | w/o pre-training |
|----------|----------------|-----------------|
| batch size | batch size | batch size |
| 1024 | 128 | 128 |
| test | test | test |

Other transformers

| Approach | w/ pre-training | w/o pre-training |
|----------|----------------|-----------------|
| batch size | batch size | batch size |
| 1024 | 128 | 128 |
| test | test | test |

Swin-T transformers

| Approach | w/ pre-training | w/o pre-training |
|----------|----------------|-----------------|
| batch size | batch size | batch size |
| 1024 | 128 | 128 |
| test | test | test |

b) Models and Training: For our experiments we use the vanilla Swin-T transformer model [10] as baseline—a recently established vision transformer comprising 29M parameters using localized attention. For training settings we follow [10]. For some experiments we downloaded the official ImageNet-1k pre-trained model⁵ and report results after fine-tuning for 100 epochs on CIFAR-100 training data. With or without pre-training, relaxed self-attention is applied only during fine-tuning.

We investigate the interaction of our relaxed self-attention approach with other regularization methods by omitting already employed (i.e., the well-known stochastic depth method [60]) or adding recently proposed (i.e., the dense relative localization loss $\mathcal{L}_{\text{drloc}}$ [61]) approaches. For fair comparison and following common practice as in [61]–[63], we report results of our relaxed self-attention approaches after roughly optimizing test accuracy with a small grid search over $\gamma$ values (and $\sigma^2$ for fuzzy relaxation after the optimal $\gamma_0$ was found) separately for each batch size (1024 and 128) with pre-training, applying the found values to experiments without pre-training.

c) Results and Discussion: The first segment of Table IV shows results for reference vision transformer models ranging from 87.10% accuracy for the pure attention-based ViT-S-16 [9] to 91.70% accuracy for the convolution attention transformer model Swin-T [10], the second table segment presents baselines and experimental results for Swin-T transformer models where we chose the vanilla architecture [10] to resimulate a baseline for our experiments. Omitting stochastic depth [60] causes a severe loss of performance with pre-training.

TABLE V: Language model investigation for automatic speech recognition results on the Librispeech task using standard encoder-decoder transformer models. The 960h training dataset is used, see also Table I.

| Approach | absolute WER (%) | LM-induced WER reduction (%) |
|----------|------------------|-----------------------------|
| clean other | clean other | clean other | clean other |
| dev | test | dev | test |

Baseline ([18], resim.), no LM 3.92 9.00 4.47 9.23 — — — —
+ LM (train transcripts only) 3.92 8.44 4.36 8.97 3.08 2.46 1.07 1.15
+ LM (additional data, cf. Tab. I) 3.73 8.52 4.40 8.93 0.19 0.42 0.07 0.28
Relaxed cross attention, no LM 3.95 9.33 4.28 9.45 — — — —
+ LM (train transcripts only) 3.91 9.26 4.23 9.30 0.04 0.07 0.05 0.15
+ LM (additional data, cf. Tab. I) 3.44 7.74 3.58 8.35 0.51 1.59 0.70 1.10

but clearly helps when training from scratch. For the dense relative localization loss $\mathcal{L}_{\text{drloc}}$ [61], we confirm performance gains with and especially without pre-training. Smooth focus helps for the small batch size using pre-training and performs remarkably good for a large batch size when training from scratch. Without pre-training we observe that relaxed self-attention doesn’t help. This might be due to the limited number of training epochs and a slower convergence caused by the additional relaxed self-attention regularization, similar to the effect of stochastic depth in the resimulated baseline.

When applying relaxed attention after pre-training, however, relaxed self-attention alone slightly outperforms the baseline but achieves even higher accuracies when used with matched inference (88.73% vs. 88.53%) and (89.39% vs. 89.16%) for the large and small batch sizes, respectively. Matched inference turned out to be advantageous on this task in most cases, thus we continue to report based thereon. Also, we note that the combination with stochastic depth seems to be beneficial for relaxed self-attention. Our new fuzzy relaxation with matched inference turns out to be useful only on smaller batch sizes after pre-training, achieving a strong accuracy of 89.60% outperforming the baseline [10], resimulated at 89.16%. We also investigate robustness towards different initialization seeds in Section V-G.

E. Internal Language Model Suppression

In experiments where an external language model (LM) was included, we used the common shallow fusion method [32] for LM fusion to combine the output token probability vector $P_\ell$ (cf. Figure 1) for each decoding timestep $\ell$ with the same D-length LM output token probabilities $P_\ell^{(\text{LM})}$ in the logarithmic domain to gather a joint output token probability $P_\ell = \log P_\ell + \lambda \log P_\ell^{(\text{LM})}$. The external language model weight $\lambda$ is used to steer the influence of the LM during decoding and is gathered individually for each task.

As shown in Table I for automatic speech recognition, we achieved superior results with relaxed cross attention only when the transformer was combined with an external language model that is trained with large amounts of additional text-only data. This finding is in line with [18], but the authors do not provide a sound reason for such behavior. Different to hybrid ASR approaches, the output token posterior $P_\ell$ of a trained

---

⁵ImageNet-1k pre-trained Swin transformer models and fine-tuning code downloaded from https://github.com/microsoft/Swin-Transformer.
TABLE VI: Language model investigation for **automatic lip-reading** results on the **LRS3** task using standard **encoder-decoder transformer** models with pre-trained **AV-HuBERT** encoders. For fine-tuning, 433 h + 1,326 h of labeled data are used, see also Table II.

| Approach                        | absolute WER (%) | LM-induced WER reduction (%) |
|---------------------------------|------------------|------------------------------|
|                                 | dev              | test                         | dev              | test   |
| Baseline ([45], resimulated), no LM | 17.71            | 26.73                        | —                | —      |
| + LM (training transcripts only) | 17.83            | 27.22                        | -0.12            | -0.49  |
| + LM (additional data, from Tab. II) | 17.18            | 26.50                        | 0.53             | 0.23   |
| Relaxed cross attention, no LM  | 17.40            | 26.57                        | —                | —      |
| + LM (training transcripts only) | 17.48            | 26.52                        | -0.08            | 0.05   |
| + LM (additional data, from Tab. II) | **16.92**        | **25.51**                    | 0.48             | 1.01   |

Here, we investigate whether the improvement by relaxed cross attention might be due to a suppression of the **internal language model**. To accomplish this, in Table V, we measure the WER improvement achieved by using an LM when the transformer was trained with and without relaxed cross attention, respectively. Both trained transformer models are combined with two language models, one trained only from the text transcripts of the acoustic training data, and one trained with additional text-only data. Note that both, resimulated baseline results and results for the LM with additional data, are taken from Table I. We observe that for both, the baseline and the relaxed cross attention model, the improvements with the **training transcript only** LM (rows 2 and 5) vs. the no LM methods are about equally small. In contrast, when combined with the LM trained on additional data, the model trained with relaxed cross attention yields far more WER reduction as if this strong LM would be used with the baseline. In any case it exceeds an absolute reduction of 0.5% (nowhere reached with the baseline), and for the (noisy) other condition it is more than 1% absolute WER reduction if relaxed cross attention is employed. For the automatic lip-reading task we observe similar behavior in Table VI. Here the integration of the training transcripts only LM is even harmful for the baseline model (row 2), while for the relaxed cross attention approach, WERs remain roughly the same compared to the relaxed cross attention-trained model without LM (row 4 vs. 5). In combination with the strong LM, both baseline and relaxed cross attention models take profit on the dev set, while on the test set, relaxed cross attention yields a more than four-fold WER reduction by LM fusion (1.01% absolute) compared to the baseline approach (0.23% absolute).

Overall, we observe that relaxed cross attention does not yet help when the LM was trained only with the text transcript data that was already exposed to the ASR transformer model during training with acoustic data. We conclude, however, that relaxed cross attention particularly helps when the LM has been trained with additional text data and seems to weaken the internal model bias, thus suppressing the influence of the internally learned (usually poor) LM. Notably, the same behavior is observed in Table III for neural machine translation.

### F. Ablation Study on Attention Dropout

Depicted as dashed boxes in Figures 2 and 3, the well-known dropout method [27] is employed to the standard encoder-decoder transformer in three different variations: Residual dropout, activation dropout, and—most relevant for our study—attention dropout, where the latter is either applied to the attention weights $G_{i}$ after the softmax layer (baseline) or to the modified attention weights $G_{i}$ after the relaxation operation (relaxed attention approaches, see equation (1)). In Table VII, we investigate how these regularization operations interfere with each other for two different tasks that incorporate attention dropout during training. Therefore, in this ablation, we removed attention dropout throughout the encoder and the decoder of the transformer model for both approaches with and without specified types of relaxed attention. Note that the employed models for the lip-reading and image recognition tasks did not use attention dropout (following the respective baseline recipes from [45] and [10]) and are thus omitted for this ablation.

We note that relaxed attention nicely combines with attention dropout [27] as in the test conditions of both tasks the combination of relaxed self-/cross attention with attention dropout yields the best results, which are also reported in the main experiments for both specific tasks in Section V. Interestingly, attention dropout did even harm the baseline performance for machine translation, as omitting it yields an 0.09 absolute increase in BLEU score, while it improves the advantageous relaxed self-attention even further. In summary, we observe that both proposed relaxed attention approaches seem to go along with other regularization approaches, such as attention dropout, providing complementary regularization to the attention layers.

| Task                  | Automatic Speech Recognition (Librispeech) | Machine Translation (IWSLT14) |
|-----------------------|-------------------------------------------|--------------------------------|
| Setting               | w/ LM 100 h training data                  | w/o LM, self-attention matched inference |
| Relaxation type (*)   | cross attention                            |                               |
| Metric                | WER (%) ↓                                  | BLEU ↑                         |
| Data subset           | dev clean other test clean other test test |
| Baseline (resimulated) | 10.62 24.19 12.06 25.56                     | 37.42                          |
| - attention dropout   | 11.02 25.24 11.89 26.88                     | 37.51                          |
| + relaxed (*) attention | 9.33 22.16 10.62 23.04                      | 37.67                          |
| - attention dropout   | 9.68 21.38 10.91 23.16                     | 37.47                          |
G. Robustness to Different Initialization Seeds

In Table VIII, we investigate the influence of different initialization seeds on our experiments. While for the main experiments in Section V we experimented on an unchanged and non-optimized seed for random number generation, here—since both of our top-performing contributions are based on the novel self-attention—we analyze the best relaxed self-attention schemes of each task w.r.t. statistical significance when using 5 different random seeds.

We note that in these experiments, we achieve significant improvement for all three sequence-based tasks including those where we claim state-of-the-art and top performance (i.e., lip-reading and machine translation). In addition, not shown here, the relaxed cross attention method yielded even better performance on all three sequence-based tasks, outperforming relaxed self-attention, but we do not formulate performance claims in this particular analysis as it implies extra computational complexity due to the requirement of a language model as well as additional unpaired text training data. For the image classification task, note that we reach a clear improvement using the non-optimized standard seed for initialization of our main experiments (see Table IV). Here, however, with additional seeds for initialization, we observe the baseline and the fuzzy relaxation approach to differ without statistical significance. We suspect this is due to non-deterministic operations in the original baseline code from [10], which might have flawed the tuning process for the relaxation coefficients for fuzzy relaxation. However, as the average accuracy with fuzzy relaxation is still higher (89.45% vs. 89.29%), we feel encouraged to further expand the relaxed self-attention approach to attention-based approaches for computer vision tasks.

VI. CONCLUSIONS

In this work we broadly explored the idea of relaxed attention for transformer architectures, a simple smoothing method of the attention weights in the attention layers. We confirmed the advantage of relaxed cross attention when combined with strong external language models and introduced relaxed self-attention, thereby providing regularization also in the transformer encoder and increasing the versatility of relaxed attention to different transformer variants. We show improvements when applying relaxed attention to automatic speech recognition, lip-reading, machine translation, and image classification. On the LRS3 lip-reading task in particular we achieve a word error rate of 26.31% (vs. the former state of the art of 26.90%) as well as a top-performing BLEU score of 37.67 on the IWSLT14 machine translation task.

REFERENCES

[1] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk et al., “Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation,” in Proc. of EMNLP, Doha, Qatar, Oct. 2014, pp. 1724–1734.
[2] K. Cho, B. van Merriënboer, D. Bahdanau, and Y. Bengio, “On the Properties of Neural Machine Translation: Encoder-Decoder Approaches,” in Proc. of SIST; Doha, Qatar, Oct. 2014, pp. 103–111.
[3] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” in Proc. of ICLR, San Diego, CA, USA, May 2015, pp. 1–18.
[4] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, “Attention-Based Models for Speech Recognition,” in Proc. of NIPS, Montréal, Canada, Dec. 2015, pp. 577–585.
[5] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, “Listen, Attend and Spell: A Neural Network for Large Vocabulary Conversational Speech Recognition,” in Proc. of ICASSP, Shanghai, China, Mar. 2016, pp. 4960–4964.
[6] D. Bahdanau, J. Chorowski, D. Serdyuk, P. Brakel, and Y. Bengio, “End-to-End Attention-Based Large Vocabulary Speech Recognition,” in Proc. of ICASSP, Shanghai, China, Mar. 2016, pp. 4945–4949.
[7] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” in Proc. of NAACL-HLT, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.
[8] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, “End-to-End Object Detection With Transformers,” in Proc. of ECCV, Glasgow, UK, Aug. 2020, p. 213–229.
[9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner et al., “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” in Proc. of ICLR, virtual, May 2021, pp. 1–21.
[10] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang et al., “Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows,” in Proc. of ICCV, virtual, Oct. 2021, pp. 10012–10022.
[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez et al., “Attention Is All You Need,” arXiv:1706.03762, Dec. 2017.
[12] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal et al., “Language Models Are Few-Shot Learners,” in Proc. of NeurIPS, vol. 33, virtual, Dec. 2020, pp. 1877–1901.
[13] J. Kaplan, S. McCandlish, T. Henighan, B. Chess, R. Child et al., “Scaling Laws for Neural Language Models,” arXiv:2001.08361, Jan. 2020.
[14] W. Fedus, B. Zoph, and N. Shazeer, “SwitchTransformers: Scaling to Trillion Parameter Models With Simple and Efficient Sparsity,” Journal of Machine Learning Research, vol. 23, pp. 2–40, 2022.
[15] P. Xu, D. Kumar, W. Yan, Y. Zitlin, C. Huang et al., “Optimizing Deeper Transformers on Small Datasets,” in Proc. of ACL, Bangkok, Thailand, Aug. 2021, pp. 2089–2102.
[16] M. Popel and O. Bojar, “Training Tips for the Transformer Model,” The Prague Bulletin of Mathematical Linguistics, vol. 110, pp. 43–70, 2018.
[17] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, “How to Train Your ViT? Data, Augmentation, and Regularization in Vision Transformers,” arXiv:2106.10270, Jun. 2021.
