Entropy– and Distance-Regularized Attention Improves Low-Resource Neural Machine Translation

Ali Araabi a.araabi@uva.nl
Vlad Niculae v.niculae@uva.nl
Christof Monz c.monz@uva.nl
Informatics Institute, University of Amsterdam, Amsterdam, The Netherlands

Abstract

Transformer-based models in Neural Machine Translation (NMT) rely heavily on multi-head attention for capturing dependencies within and across source and target sequences. In Transformers, attention mechanisms dynamically determine which parts of the sentence to focus on in the encoder and decoder through self-attention and cross-attention. Our experiments show that high-resource NMT systems often exhibit a specific peaked attention distribution, indicating a focus on key elements. However, in low-resource NMT, attention tends to be dispersed throughout the sentence, lacking the focus demonstrated by high-resource models. To tackle this issue, we present EaDRA (Entropy– and Distance-Regularized Attention), which introduces an inductive bias to prioritize essential elements and guide the attention mechanism accordingly. Extensive experiments using EaDRA on diverse low-resource language pairs demonstrate significant improvements in translation quality, while incurring negligible computational cost.

1 Introduction

Neural networks have revolutionized Machine Translation (MT), as evidenced by the significant progress made in recent years (Sutskever et al., 2014). The Transformer architecture (Vaswani et al., 2017) has garnered substantial attention and achieved remarkable advancements across various downstream tasks (Devlin et al., 2019; Liu et al., 2020; Brown et al., 2020), including its application to Neural Machine Translation (NMT). However, the performance of the Transformer architecture heavily relies on the effectiveness and reliability of its attention mechanism.

Our observations from well-performing models suggest that attention should prioritize important elements, resulting in a peaked distribution of attention weights. By emphasizing crucial information, the attention mechanism enables more accurate predictions. This selective attention allows the model to effectively capture and utilize relevant information, leading to improved performance. Therefore, optimizing the attention mechanism is critical for harnessing the full potential of the Transformer architecture and enhancing its performance across tasks, including NMT. However, achieving focused attention behavior poses a significant challenge for NMT systems Raganato et al. (2020), especially in low-resource settings. Our preliminary experiments show that as the amount of available training data decreases, NMT systems tend to exhibit a lack of the desired focused attention behavior. In such low-resource scenarios, where training data is limited, the attention distribution becomes more dispersed and less selective. Consequently, the model's ability to effectively capture and utilize crucial information is hindered, leading to reduced translation performance. Therefore, as the amount of available data diminishes, it becomes crucial to develop techniques that can guide the attention mechanism towards relevant and informative elements of the source sentence. In order to address this issue, prior research has suggested hard-coded or fixed attention patterns for self-attention heads to improve translation qual-

Proceedings of the 16th Conference of the Association for Machine Translation in the Americas,
Chicago, USA, September 30 - October 2, 2024. Volume 1: Research Papers

ity (Raganato et al., 2020; You et al., 2020). However, the complexity and diversity of language necessitate the consideration of varied attention patterns for different sentences in the context of translation. By constraining the attention weights to fixed values, the model may encounter difficulties in accommodating diverse sentence structures and capturing long-range dependencies with accuracy.

Consequently, this constraint may result in diminished performance (You et al., 2020), particularly for sentences that do not align harmoniously with the predetermined attention patterns.

In this paper, after identifying a significant difference in the entropies of attention heads between high-resource and low-resource trained models (Section 5.1), we introduce an inductive bias through the proposition of entropy and distance regularization (Section 3.3). Our approach aims to induce selective attention by regularizing the distance and entropy in the distribution of attention heads. Specifically, we introduce a novel term into the loss function to guide the learning process, which encourages the low-resource NMT model to emulate the patterns observed in the attention of higher-resource models. This additional bias is incorporated to improve the low-resource NMT model's capability to capture intricate language patterns and enhance translation performance. Experimental results demonstrate the effectiveness of our approach and underscore the importance of inductive bias in narrowing the performance disparity between low– and high-resource NMT systems.

2 Related work

Prior work has explored various approaches to improve low-resource performance by leveraging high-resource language pairs. This includes initializing model parameters from a large-scale trained model (Zoph et al., 2016), as well as techniques such as Multilingual Neural Machine Translation (Aharoni et al., 2019), cross-lingual knowledge distillation (Tan et al., 2019; Saleh et al., 2020) and large pre-trained models that aim for universal language understanding (Liu et al., 2020; Tang et al., 2020; Brown et al., 2020; Touvron et al., 2023). While these methods have significantly improved low-resource NMT, they rely on the availability of a large amount of additional data. However, it is crucial to explore techniques that facilitate the more efficient utilization of the model. Inductive bias plays a fundamental role in machine learning as it allows for the incorporation of prior knowledge or assumptions into learning systems (Mitchell, 1980). Different regularization techniques and architectural choices can introduce specific biases to shape the behavior of models. For example, regularization biases models towards relying less on a few influential features, Convolutional Neural Networks bias models to capture local relationships between input, and attention mechanisms (Bahdanau et al., 2015; Vaswani et al., 2017) bias models to capture long-range dependencies. Additionally, in the context of attention mechanisms, specific biases can be introduced to shape the behavior of models and improve their performance. Lin et al. (2018) encourage the attention to pay more focus on the content words rather than functions words. In the context of summarization, Aralikatte et al. (2021) propose an attention mechanism that proactively generates tokens in the decoder that are similar or topical to the input. Niculae and Blondel (2017) introduce an attention mechanism that is encouraged to assign similar attention weights to consecutive words. Structured attention networks (Kim et al., 2017) incorporate graphical models to generalize simple attention, while the training time significantly ($5\times$) increases. More similar to our motivation, LP-SparseMAP (Niculae and Martins, 2020) models attention distance between consecutive words for a classification task by introducing trainable parameters, but its scalability to large-scale experiments is limited. In contrast, our approach, based on applying a regularizer, is faster, less complex, and can be efficiently executed on GPUs, making it scalable for large-scale training and fine-tuning setups.

The closest work to our method is Fixed-attention (Raganato et al., 2020), which enforces fixed (untrainable) attention patterns. However, they focus solely on encoder self-attention, overlooking the importance of cross-attention heads in neural machine translation (Voita et al., 2019; You et al., 2020). Similarly, You et al. (2020) introduce Hard-Coded Gaussian Attention that replaces the attention distribution computation, i.e., scaled dot product of queries and keys, with a fixed Gaussian distribution, leading to a negative impact on translation quality. Given the concept of entropy that has been used in machine translation (Montahaei et al., 2019),

in the next section, we propose our method that can be applied to all different attention components, i.e., encoder self-attention, decoder self-attention, and encoder-decoder (cross) attention, while consistently yielding significant improvements across various experimental setups.

3 Methodology

3.1 Entropy

Entropy, a fundamental concept in information theory, has found various applications in the field of NLP (Pimentel et al., 2021; Vanmassenhove et al., 2021). One prominent area where the concept of entropy has been utilized is in language modeling and generation (Han et al., 2018; Meister et al., 2020). By quantifying the degree of uncertainty or unpredictability of a language model's output, entropy serves as a measure of the model's confidence or information content (Shannon, 1948). Given a probability vector $\boldsymbol{a} \in \mathbb{R}^n$, i.e., whose entries are non-negative and sum to 1, the Shannon entropy is defined as:

$$H(\boldsymbol{a}) \coloneqq -\sum_{i=1}^{n} a_i \log_2(a_i) \tag{1}$$

In order to capture the shape of attention distributions (more peaked versus more flat) we use the Shannon entropy, defined in Equation 1. For a discrete distribution of dimension n, Shannon entropy takes values between 0 and $\log_2 n$, with $H(\boldsymbol{a}) = 0$ when $\boldsymbol{a}$ is a peaked one-hot vector, and $H(\boldsymbol{a}) = \log_2 n$ when $\boldsymbol{a} = (1/n, \ldots, 1/n)$. where we define the length-normalized entropy:

$$H_N(\boldsymbol{a}) \coloneqq \frac{1}{\log_2 n} H(\boldsymbol{a}) \tag{2}$$

in order to remove unwanted effects induced by varying sentence lengths, by ensuring the output of H_N falls within the range of 0 to 1.

3.2 Entropy penalties

We propose a method to replicate the desirable behavior observed in higher-resource models by introducing an inductive bias to the attention mechanism in lower-resource models, encouraging a focused behavior to guide the attention mechanism towards more important information. In Transformers, there are multiple attention heads that allow the model to capture diverse and fine-grained relationships within the input sequence: enc (self-attention in the encoder), dec (self-attention in the decoder), and x (encoder-decoder or cross attention). Each attention mechanism computes the attention distribution for each word in the input sentence x. More specifically, when translating a sentence pair x, y, the attention heads of a Transformer model compute several attention distributions:

$$\text{Attention}(x, h, t) = \sum_{i=1}^{n} \boldsymbol{a}_{i,h,t} \cdot V_{i,h,t} \tag{3}$$

where $V_{i,h,t}$ is the value matrix and $\boldsymbol{a}_{i,h,t}$ is the attention distribution at word i calculated at head h, for attention type $t \in \{\text{enc}, \text{dec}, \text{x}\}$. $\boldsymbol{a}_{i,h,t}$ is a probability vector of length n_{src} when $t \in \{\text{enc}, \text{x}\}$ and of length n_{tgt} when $t = \text{dec}$.

To encourage peaked attention and nudge attention heads toward selecting the important information, we apply an entropy-minimizing penalty on all attention distributions:

$$R_{\text{peak}} \coloneqq \sum_{i,h,t} H_N(\boldsymbol{a}_{i,h,t}), \tag{4}$$

By itself, this regularizer can force attention heads to trivial solutions, e.g., where all mass is concentrated on a token in a sentence. To mitigate this we invoke another inductive bias based on a desirable property observed in high-performing models: even though individual attention heads are peaked, the attention distribution averaged over the entire sentence:

$$\bar{\boldsymbol{a}}_{h,t} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{a}_{i,h,t} \tag{5}$$

should be flat. We therefore propose an additional sentence-level entropy-maximizing penalty:

$$R_{\text{sent}} \coloneqq -\sum_{h,t} H_N(\bar{\boldsymbol{a}}_{h,t}) \tag{6}$$

3.3 EaDRA (Entropy- and Distance-Regularized Attention)

In this section, we propose a distance-based method that goes beyond simply minimizing attention entropies. This method not only reduces entropy and enhances attention concentration, but also induces a preference for attending to adjacent tokens, motivated by the significance of proximity-based attention in NMT tasks (Raganato et al., 2020).

To develop the intuition, we focus on a single attention head (hence temporarily dropping the h, t

indices). Imagine for a moment our attention was a hard selection mechanism, e.g., $(\boldsymbol{e}_i)_{j_i} = 1$, indicating that attention at token i selects only token j_i. If neighboring words and contiguous phrases are highly relevant to each other, we would expect the total distance between consecutive selections

$$R_{\text{dist}} \stackrel{\sim}{=} \sum_{i=1}^{n-1} d(j_i, j_{i+1}) \tag{7}$$

to be rather small, where d is a discrete, one-dimensional distance function,[1] and $\stackrel{\sim}{=}$ is used since this is just an intuition and not yet a usable definition: since our attention is soft and not hard, we cannot directly measure this total distance suggested by Equation 7. Instead, we relax the definition by considering expectation over j_i and j_{i+1}, interpreted as random variables with marginal distributions $\boldsymbol{a}_i$, respectively $\boldsymbol{a}_{i+1}$. We can then penalize the total *expected distance*:

$$\begin{aligned} R_{\text{dist}} &:= \sum_{i=1}^{n-1} \mathbb{E}_{j_i} \mathbb{E}_{j_{i+1}} [d(j_i, j_{i+1})] \\ &= \sum_{i-1}^{n-1} \boldsymbol{a}_i^\top \boldsymbol{D} \boldsymbol{a}_{i+1}, \end{aligned} \tag{8}$$

where $\boldsymbol{D}$ is the distance matrix defined by $(\boldsymbol{D})_{st} = d(s,t)$ for our chosen distance function d. This matrix can be precomputed and the quadratic form in Equation 8 is fast to evaluate on GPUs, although we remark, since d is symmetric, that $\boldsymbol{D}$ is a Toeplitz matrix and therefore R_{dist} could be computed via fast discrete Fourier transform.

Putting together all terms, our objective for a given training sentence pair (x, y) minimizes:

$$\begin{aligned} L(x,y) = &\sum_{i=1}^{n} -\log p(y_i \mid x, y_{1:i-1}) \\ &+ \alpha_{\text{peak}} R_{\text{peak}} + \alpha_{\text{sent}} R_{\text{sent}} + \alpha_{\text{dist}} R_{\text{dist}}. \end{aligned} \tag{9}$$

Here, α parameters control the relative impact of the various penalties. We call this method EaDRA (Entropy- and Distance-Regularized Attention), the distance-based and entropy-based regularizers. Unlike fixed diagonal patterns in attention, EaDRA allows for more flexibility in achieving a peaky attention distribution.

[1] We use the absolute distance, $d(s,t) = |s - t|$, but arbitrary functions may be used instead.

4 Experimental setup

4.1 Data setup

In our preliminary experiments, we use a dataset comprising 4 million German-English training samples from WMT14, which includes Europarl, Common Crawl, and News Commentary.

Code	Dataset	#Sents
Ex. LR		
Be-En	TED Qi et al. (2018)	4.5k
Gl-En	TED Qi et al. (2018)	10k
De-En	WMT14	50k
Sk-En	TED Qi et al. (2018)	55k
LR		
Ko-En	Jungyeul Park et al. (2016)	90k
Kk-En	WMT19	91k
En-De	WMT14	100k
Vi-En	IWSLT15 (Cettolo et al., 2012)	133k
En-De	IWSLT14 (Cettolo et al., 2012)	160k
Tr-En	WMT17	207k
Ja-En	IWSLT17 (Cettolo et al., 2012)	223k
En-De	WMT14	250k

Table 1: Details of extremely low-resource (Ex. LR) and low-resource (LR) datasets in our experiments.

To simulate the low-resource scenario in a controlled setting, we randomly choose subsets of 50k, 100k, 250k and 1m samples. We evaluate on the newstest2014 test set. Additionally, we conduct experiments on two sets of language pairs (Table 1), one representing low-resource scenarios and the other representing extremely low-resource scenarios.

All datasets, except Japanese-English, are preprocessed by applying punctuation normalization, tokenization (Koehn et al., 2007), limiting the length of the sentences to 200 tokens and removing sentence pairs with a source/target length ratio exceeding 1.5, following previous work (Ng et al., 2019). Then, we use BPE (Sennrich et al., 2016) to split the data with BPE parameter selection with respect to the data size (Araabi and Monz, 2020).

For the Japanese-English language pair, we use SentencePiece with a shared vocabulary size of 16k, as it has been widely recognized for its effectiveness in handling Japanese text (Kudo and Richardson,

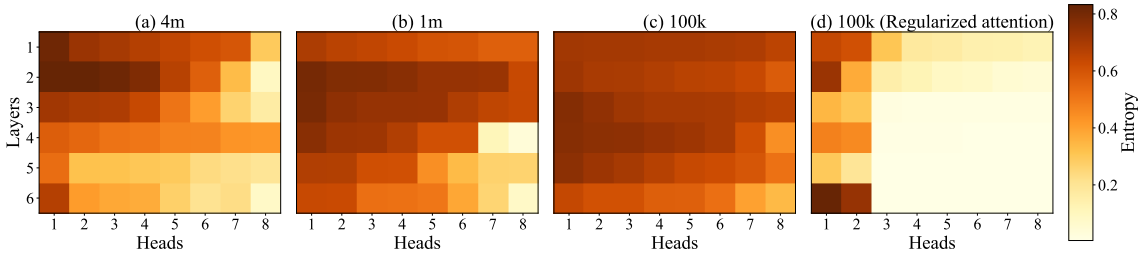

Figure 1: Entropy distribution of the encoder self-attention in a Transformer with 6 layers and 8 attention heads. (a), (b), and (c) are models trained on 4m, 1m, and 100k samples, respectively. (d) is the model trained on 100k samples after applying EaDRA. All training sets are random samples from WMT14 En-De. Entropy values are sorted within each layer to highlight the contrasting patterns.

2018). In order to evaluate the models, for Belarusian (Be), Galician (Gl), Slovak (Sk), Korean (Ko), [1] Kazakh (Kk), WMT German (De), Vietnamese (Vi), Turkish (Tr), and Japanese (Ja) we use their own official test sets. For IWSLT German (De), following (Raganato et al., 2020) we use the concatenation of the IWLST 2014 dev sets (tst2010–2012, dev2010, dev2012).

Model	#sent	min	ave	max	BLEU
T.base	4m	0.08	0.42	0.83	28.1
	1m	0.04	0.58	0.79	24.1
	100k	0.34	0.66	0.77	13.5
$\text{EaDRA}_{\text{enc+dec}}$	100k	0.01	0.18	0.82	16.2

Table 2: Statistics of entropy values over all encoder self-attention heads of models with different sample sizes from WMT14 English-German, trained on Transformer-base. $\text{EaDRA}_{\text{enc+dec}}$ denotes EaDRA applied on the self-attention in the encoder and decoder of Transfomer.

4.2 Model Configuration

We adopt the Transformer-base (denoted by T.base) architecture with its original hyperparameters (Vaswani et al., 2017) as our baseline model, upon which our proposed modifications are built. In addition, we consider the Fixed-attention method (Raganato et al., 2020) as the most closely related baseline approach. Our experiments are conducted using the Fairseq library (Ott et al., 2019). We evaluate the translation quality using sacreBLEU (Post, 2018) as evaluation metric.[2] All experiments can be completed within a few hours using a single GPU with the model parameters ranging from 49m to 65m.

5 Results

In this section, we start with a comprehensive analysis of multi-head attention entropy across various data setups. Subsequently, we demonstrate the striking effectiveness of EaDRA when compared to both the Transformer model and the most closely related approach, Fixed Attention. Additionally, we delve into the influence of EaDRA's hyperparameters. Moreover, we present results involving large pre-trained fine-tuning, a method widely recognized as a strong baseline.

5.1 Analysis of entropy in multi-head attention

The limitations in low-resource NMT performance can be attributed to the inherent difficulties associated with training models using limited data (Koehn and Knowles, 2017). However, the impact of this data scarcity on the multi-head attention mechanism remains unclear. Building on the observation of dispersion of weights in attentions (Voita et al., 2019; Correia et al., 2019), in this section we aim to analyze and compare the weight distribution of multi-head attention in NMT models across different data regimes. For this purpose, entropy serves as a useful measure by providing valuable insights into the

[1] https://github.com/jungyeul/korean-parallel-corpora/tree/master/korean-english-news-v1

[2] sacreBLEU signature:
nrefs:1—case:lc—eff:no—tok:13a—smooth:exp—version:2.0.0

peakedness of the attention distribution. We conduct preliminary experiments to investigate the entropy of attention heads and gain insights into their behavior. Figure 1 (a-c) illustrates the entropy of encoder self-attention heads for models trained on different sample sizes: 4m, 1m, and 100k. A clear trend is observed where the entropy of attentions decreases as the amount of training data increases. Therefore, the models trained with smaller data sizes face challenges in learning focused attention distributions. Based on this observation, we hypothesize that this trend of decreasing entropy with larger training samples will continue, and with a substantial amount of data, ideally, the entropy will approach zero. Figure 1 (d) illustrates the entropies of the encoder self-attention after the application of our method, showing a significant decrease in entropy. This decrease indicates a higher level of peakedness or concentration in the attention distribution. In order to compare EaDRA's entropy patterns with those of Fixed-attn, it is essential to note that Fixed-attn primarily utilizes attention heads characterized by fixed diagonal or tridiagonal-like patterns. As a result, the entropy for three of these heads reaches zero, while the remaining heads consistently maintain an entropy close to zero, forming a consistent value irrespective of the dataset size or input characteristics.

Table 2 presents the statistics of Figure 1. We observe a substantial difference in the average and minimum entropy values across all attention heads between the higher-resource models and low-resource one. Therefore, EaDRA results in a significant decrease in entropy of attention weights, resulting in a more peaked distribution of attention weights similar to what can be achieved with a large amount of training data. However, it is crucial to contextualize these findings by considering that a fair comparison, as exemplified by the performance of EaDRA compared to the T.base trained on 100k samples, demonstrates the efficacy of our approach under more controlled conditions, where both are trained on a similar number of sentences. Additionally, it is worth noting that the improvement observed in row 4 is a direct consequence of our precise parameter tuning for EaDRA.

5.2 EaDRA in multi-head attention components

While Raganato et al. (2020) only focus on the encoder self-attention, EaDRA is applicable to all attention components. We empirically demonstrate this through our experiments, which involve the encoder self-attention, decoder self-attention, and cross-attention. The performance of EaDRA on various components and their combinations is presented in Table 3. The results demonstrate that EaDRA consistently leads to substantial improvements across all cases, with the encoder and decoder combination (enc+dec) yielding the highest performance on lower-resource setups.

model	50k	100k	250k	1m
T.base	6.2	13.5	19.9	24.1
Fixed-attn	9.3	13.1	19.0	20.4
EaDRA$_{enc}$	9.4	15.2	20.2	24.4
EaDRA$_{dec}$	8.1	15.2	20.0	24.4
EaDRA$_{x}$	8.2	14.1	20.0	24.5
EaDRA$_{dec+x}$	8.2	14.8	20.0	24.4
EaDRA$_{enc+x}$	9.0	15.6	**20.6**	24.6
EaDRA$_{enc+dec}$	**9.7**	**16.2**	20.2	24.1
EaDRA$_{enc+dec+x}$	9.6	16.1	20.4	**24.7**

Table 3: Results of applying EaDRA to encoder self-attention (enc), decoder self-attention (dec), and cross-attention (x) on 50k, 100k, 250k, and 1m random samples from WMT14 English-German. BLEU scores are reported on newstest2014. Fixed-attn refers to our reimplementation of the Fixed-attention method (Raganato et al., 2020)

However, the cross-attention component does not benefit substantially from EaDRA, compared to the other components and combinations. We speculate that this observation may be attributed to the inherent differences in word ordering between the source and target languages, where EaDRA might discourage some specific reorderings. Moreover, EaDRA consistently outperforms Fixed-attention in all experimental settings and Fixed-attention fails to exhibit any improvement over the vanilla Transformer, except for the smallest training set with 50k samples. Notably, as the amount of training data decreases, the degree of performance degradation in Fixed-attention also diminishes. In addition, we conduct experiments with applying Fixed-

attention to other attention components (decoder self-attention and cross-attention) and their combination, observing a notable decline in translation quality. This observation aligns with the results of hard-coded attention (You et al., 2020), which revealed that hard-coded encoder and decoder attention adversely affect translation quality, and hard-coded cross-attention leads to a more significant decrease in BLEU score, potentially due to its higher importance in the translation process (Voita et al., 2019; Gheini et al., 2021). Nevertheless, due to EaDRA's focus on biasing attentions without imposing strict constraints, it exhibits flexibility that allows for improvements even in case of cross-attention.

To further explore the impact of EaDRA in achieving focused attention, we perform a set of experiments in low-resource settings across various translation tasks. The results are summarized in Table 4, clearly demonstrating the significant improvements achieved by EaDRA. Specifically, our analysis focus on individual attention components as well as the combined encoder and decoder attention components ($\text{EaDRA}_{\text{enc+dec}}$), which consistently outperformed other combinations in smaller samples from WMT14 En-De, as shown in Table 3.

Interesting observations arise in the context of extremely low-resource scenarios, specifically for Belarusian and Galician datasets, with training sample sizes of only 4.5k and 10k, respectively. Surprisingly, in these cases, Fixed-attention outperforms EaDRA. We suspect that this superiority of fixed attention patterns in extreme scenarios can be attributed to the model's limited capacity to effectively learn attention distributions with such a small amount of training data, even when biased towards selective attention. The fact that the performance degradation is mitigated as the training size decreases and Fixed-attention only exhibits improvement on the smallest dataset, see Table 3, further supports this hypothesis. Also, this observation aligns with the findings of Araabi and Monz (2020) in extremely low-resource settings, which demonstrate that in the presence of limited data, having more than two attention heads leads to a significant performance drop, potentially as the model struggles to learn attention patterns.

[1]For more details, see Appendix A.

5.3 Hyper-Parameters

We tune the hyperparameters (α_{peak}, α_{sent}, and α_{dist}) for every attention components separately, such that once the optimal value of a hyper-parameter has been determined, it remains fixed and we sweep over the next one. [1] We conducted additional experiments to investigate the influence of the number of attention heads used in EaDRA. Figure 2 depicts the relationship between the BLEU score and the number of attention heads employed in EaDRA (enc), showing that around 6 attention heads appear to be an optimal choice. This pattern was consistent across the experiments conducted for decoder self-attention and cross-attention, indicating that 6 heads yield favorable results for all attention components.

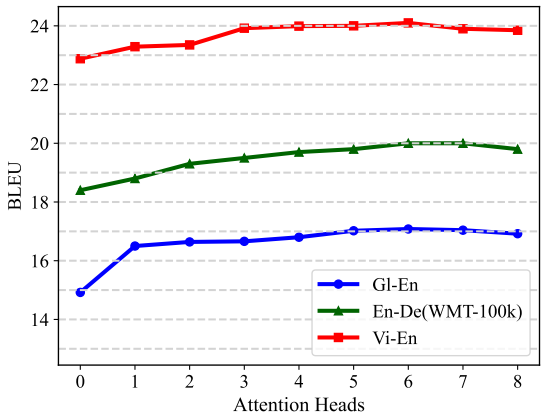

Figure 2: Effect of EaDRA with involving different number of encoder attention heads on validation BLEU score. Similar patterns are observed for other language pairs and attention components.

Initially, we conducted hyperparameter tuning on a subset of the data, specifically 100k training samples from the WMT14 English-German dataset. This process resulted in a substantial improvement of +3.1 BLEU points over Fixed-attention. Encouraged by these promising results, we proceeded to apply the same hyperparameter settings to the 50k, 250k, and 1m datasets, which led to substantial improvements across all scenarios. These findings demonstrate the effectiveness of the optimized hyperparameter values across different tasks, eliminating the need for fine-tuning on each individual task.

model	Be-En 4.5k	Gl-En 10k	Sk-En 55k	Ko-En 90k	Vi-En 133k	De-En 160K	Tr-En 207k	Ja-En 223k
T.base	5.0	13.1	22.8	6.5	25.6	32.0	16.2	10.6
Fixed-attn	5.5	**18.8**	25.4	8.1	25.3	32.4	17.0	11.7
$\text{EaDRA}_{\text{enc}}$	5.2	15.3	**26.3**	7.8	27.3	32.8	17.4	**11.8**
$\text{EaDRA}_{\text{dec}}$	5.2	16.1	25.9	**8.3**	**27.6**	**33.0**	**17.6**	11.0
EaDRA_{x}	5.3	15.6	25.7	7.6	26.7	32.7	16.6	11.1
$\text{EaDRA}_{\text{enc+dec}}$	**5.8**	16.7	25.3	8.1	27.3	32.9	17.3	11.3

Table 4: Comparing EaDRA applied to single attention components and also combination of encoder and decoder self-attention with Fixed-attention and Transformer-base for low-resource language pairs.

However, it is important to note that a more thorough hyperparameter sweep for each language pair in Table 4 produced slightly different optimal parameter values, which resulted in slight further improvements.

5.4 Large pre-trained fine-tuning

Large pre-trained models, such as mBART (Liu et al., 2020), have become an integral part of many natural language processing tasks, as they capture a vast amount of knowledge from extensive training on massive datasets. Modifying or fine-tuning such models while preserving their learned representations is a challenging task, requiring careful consideration of the model's complex architecture, attention mechanisms, and overall behavior. Therefore, it is imperative to develop methods that can leverage the existing strengths of pre-trained models while pushing for further improvements.

model	Ko-En 90k	Kk-En 91k	Vi-En 133k	Tr-En 207k	Ja-En 223k
mBART-FT	**16.0**	17.2	36.0	22.8	16.3
Fixed-attn	15.1	16.8	35.2	21.9	15.7
$\text{EaDRA}_{\text{enc}}$	**16.0**	**18.1**	**36.4**	**23.0**	**16.5**
$\text{EaDRA}_{\text{enc+dec}}$	15.6	17.9	36.1	21.0	16.2

Table 5: Comparison of Fine-tuning mBART using Fixed-attention (Raganato et al., 2020) and EaDRA applied to encoder self-attention and also encoder and decoder self-attention components.

Table 5 shows the effectiveness of applying EaDRA on top of mBART across all language pairs except Ko-En. We use the same hyper-parameter values that were tuned for 100k samples from the WMT14 En-De dataset. However, interestingly, we found that involving only two attention heads in EaDRA yields slightly higher performance. This observation can be attributed to the fact that the attention heads in mBART already exhibit a significant degree of peakedness—perhaps thanks to the pretraining—and further regularization through EaDRA does not yield additional improvements. We observe a consistent degradation of mBART when using the Fixed-attention method. One possible explanation is that applying fixed attention patterns on top of mBART introduces limitations or constraints that hinder the model's ability to fully leverage its capacity, ultimately leading to performance degradation. This suggests that the flexibility and adaptability of mBART's attention mechanisms play a crucial role in its overall performance. Furthermore, our experiments with the two most importance fixed patterns, namely the previous and next tokens (Raganato et al., 2020), also resulted in performance degradation.

6 Discussion

By introducing regularization techniques that target distance and entropy in attention heads, we achieve substantial improvements over various language pairs. Extensive experiments demonstrate the effectiveness of these methods in low-resource NMT scenarios. The flexibility offered by EaDRA enables the NMT model to selectively allocate attention during training. Conversely, fixed and unlearnable attention patterns prove to be more beneficial in the case of extremely low-resource languages with fewer than 50k training samples. In such scenarios, fixing the attention mechanism provides a more reliable approach, as the model's capacity to learn from a small dataset is limited.

7 Conclusion

In this work, we mitigate the challenge of improving low-resource NMT by introducing a form of regularized attentions. We introduce EaDRA, which promotes focused attention by prioritizing key elements. Extensive experiments on diverse low-resource language pairs demonstrate significant improvements in translation quality, validating the effectiveness of EaDRA. Our findings highlight the importance of attention regularization techniques in enhancing NMT performance, particularly in low-resource settings. EaDRA offers a practical and scalable solution with negligible computational overhead and a few lines of code.

8 Limitations

We only focus on improving low-resource NMT. However, higher-resource settings might also gain from regularized attentions facilitated by EaDRA and it may contribute to faster convergence as well. Additionally, we demonstrate the effectiveness of our proposed method using multiple low-resource language pairs, whereas there are many other language pairs with limited data. Furthermore, the encouragement of focused attention rather than dispersed attention through EaDRA leads us to hypothesize that our method may exhibit higher generalizability to sentence perturbations. This, in turn, could result in less volatile behavior of the NMT system (Fadaee and Monz, 2020). We leave these investigations to future work.

References

Aharoni, R., Johnson, M., and Firat, O. (2019). Massively multilingual neural machine translation. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3874–3884. Association for Computational Linguistics.

Araabi, A. and Monz, C. (2020). Optimizing transformer for low-resource neural machine translation. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3429–3435. International Committee on Computational Linguistics.

Aralikatte, R., Narayan, S., Maynez, J., Rothe, S., and McDonald, R. T. (2021). Focus attention: Promoting faithfulness and diversity in summarization. In Zong, C., Xia, F., Li, W., and Navigli, R., editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6078–6095. Association for Computational Linguistics.

Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In Bengio, Y. and LeCun, Y., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Cettolo, M., Girardi, C., and Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *Conference of European Association for Machine Translation*, pages 261–268.

Correia, G. M., Niculae, V., and Martins, A. F. T. (2019). Adaptively sparse transformers. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2174–2184. Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers

for language understanding. In Burstein, J., Doran, C., and Solorio, T., editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Fadaee, M. and Monz, C. (2020). The unreasonable volatility of neural machine translation models. In Birch, A., Finch, A. M., Hayashi, H., Heafield, K., Junczys-Dowmunt, M., Konstas, I., Li, X., Neubig, G., and Oda, Y., editors, *Proceedings of the Fourth Workshop on Neural Generation and Translation, NGT@ACL 2020, Online, July 5-10, 2020*, pages 88–96. Association for Computational Linguistics.

Gheini, M., Ren, X., and May, J. (2021). Cross-attention is all you need: Adapting pretrained transformers for machine translation. In Moens, M., Huang, X., Specia, L., and Yih, S. W., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1754–1765. Association for Computational Linguistics.

Han, Y., Jiao, J., Lee, C., Weissman, T., Wu, Y., and Yu, T. (2018). Entropy rate estimation for markov chains with large state space. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9803–9814.

Kim, Y., Denton, C., Hoang, L., and Rush, A. M. (2017). Structured attention networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In Carroll, J. A., van den Bosch, A., and Zaenen, A., editors, *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*. The Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In Luong, T., Birch, A., Neubig, G., and Finch, A. M., editors, *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39. Association for Computational Linguistics.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In Blanco, E. and Lu, W., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018: System Demonstrations, Brussels, Belgium, October 31 - November 4, 2018*, pages 66–71. Association for Computational Linguistics.

Lin, J., Sun, X., Ren, X., Li, M., and Su, Q. (2018). Learning when to concentrate or divert attention: Self-adaptive attention temperature for neural machine translation. In Riloff, E., Chiang, D., Hockenmaier, J., and Tsujii, J., editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2985–2990. Association for Computational Linguistics.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:726–742.

Meister, C., Salesky, E., and Cotterell, R. (2020). Generalized entropy regularization or: There's nothing special about label smoothing. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6870–6886. Association for Computational Linguistics.

Mitchell, T. M. (1980). *The need for biases in learning generalizations*. Citeseer.

Montahaei, E., Alihosseini, D., and Baghshah, M. S. (2019). Jointly measuring diversity and quality in text generation models. *CoRR*, abs/1904.03971.

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., and Edunov, S. (2019). Facebook fair's WMT19 news

translation task submission. In Bojar, O., Chatterjee, R., Federmann, C., Fishel, M., Graham, Y., Haddow, B., Huck, M., Jimeno-Yepes, A., Koehn, P., Martins, A., Monz, C., Negri, M., Névéol, A., Neves, M. L., Post, M., Turchi, M., and Verspoor, K., editors, *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*, pages 314–319. Association for Computational Linguistics.

Niculae, V. and Blondel, M. (2017). A regularized framework for sparse and structured neural attention. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3338–3348.

Niculae, V. and Martins, A. F. T. (2020). Lp-sparsemap: Differentiable relaxed optimization for sparse structured prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7348–7359. PMLR.

Ott, M., Edunov, S., Baevski, A., Fan, A., Gross, S., Ng, N., Grangier, D., and Auli, M. (2019). Fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 48–53.

Park, J., Hong, J., and Cha, J. (2016). Korean language resources for everyone. In Park, J. C. and Chung, J., editors, *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation, PACLIC 30, Seoul, Korea, October 28 - October 30, 2016*. ACL.

Pimentel, T., Meister, C., Teufel, S., and Cotterell, R. (2021). On homophony and rényi entropy. In Moens, M., Huang, X., Specia, L., and Yih, S. W., editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8284–8293. Association for Computational Linguistics.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 186–191.

Qi, Y., Sachan, D. S., Felix, M., Padmanabhan, S., and Neubig, G. (2018). When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 529–535.

Raganato, A., Scherrer, Y., and Tiedemann, J. (2020). Fixed encoder self-attention patterns in transformer-based machine translation. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 556–568. Association for Computational Linguistics.

Saleh, F., Buntine, W. L., and Haffari, G. (2020). Collective wisdom: Improving low-resource neural machine translation using adaptive knowledge distillation. In Scott, D., Bel, N., and Zong, C., editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 3413–3421. International Committee on Computational Linguistics.

Sennrich, R., Haddow, B., and Birch, A. (2016). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Tan, X., Ren, Y., He, D., Qin, T., Zhao, Z., and Liu, T. (2019). Multilingual neural machine translation with

knowledge distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tang, Y., Tran, C., Li, X., Chen, P., Goyal, N., Chaudhary, V., Gu, J., and Fan, A. (2020). Multilingual translation with extensible multilingual pretraining and finetuning. *CoRR*, abs/2008.00401.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023). Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Vanmassenhove, E., Shterionov, D. S., and Gwilliam, M. (2021). Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In Merlo, P., Tiedemann, J., and Tsarfaty, R., editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, pages 2203–2213. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., von Luxburg, U., Bengio, S., Wallach, H. M., Fergus, R., Vishwanathan, S. V. N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Voita, E., Talbot, D., Moiseev, F., Sennrich, R., and Titov, I. (2019). Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Korhonen, A., Traum, D. R., and Màrquez, L., editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5797–5808. Association for Computational Linguistics.

You, W., Sun, S., and Iyyer, M. (2020). Hard-coded gaussian attention for neural machine translation. In Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J. R., editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7689–7700. Association for Computational Linguistics.

Zoph, B., Yuret, D., May, J., and Knight, K. (2016). Transfer learning for low-resource neural machine translation. In Su, J., Carreras, X., and Duh, K., editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 1568–1575. The Association for Computational Linguistics.

9 Appendices

A Optimal hyperparameter values

The optimal values for the hyperparameters of several models discussed in the paper are presented in Table 6. Interestingly, our preliminary experiments indicate that α_{dist} appears to render α_{peak} redundant. As for the remaining models not listed in the table, we adopt the same hyperparameter values as those used for WMT En-De (100k) experiments. Furthermore, for experiments with applying EaDRA to combinations of attention heads, we do not perform additional hyperparameter tuning.

Dataset	α_{dist}	α_{sent}
EaDRA$_{\text{enc}}$		
WMT En-De (100k)	0.02	0.8
Be-En	0.02	1.2
Gl-En	0.02	0.8
Sk-En	0.02	0.8
Ko-En	0.01	0.4
Vi-En	0.02	0.6
Tr-En	0.04	1.2
Ja-En	0.02	0.8
EaDRA$_{\text{x}}$		
WMT En-De (100k)	0.1	8
Be-En	0.1	8
Gl-En	0.15	5
Sk-En	0.1	8
Ko-En	0.05	8
Vi-En	0.2	8
Tr-En	0.1	10
Ja-En	0.1	10
EaDRA$_{\text{dec}}$		
WMT En-De (100k)	2	0.8
Be-En	1	0.9
Gl-En	1	0.8
Sk-En	2	0.8
Ko-En	0.5	0.8
Vi-En	4	1
Tr-En	3	1.5
Ja-En	2	0.8

Table 6: Hyperparameters of EaDRA$_{\text{enc}}$, EaDRA$_{\text{x}}$, and EaDRA$_{\text{dec}}$ for the models presented in the paper.

B Ablation study

To gain deeper insights into the individual contributions of the proposed regularization terms, we conducted an ablation study focusing on the English-German language pair, utilizing a training set of 100k samples from WMT. The study specifically aimed to isolate the effects of the distance and sentence regularization terms. Table 7 demonstrates that employing only the distance regularization term resulted in attention heads converging to trivial solutions, leading to a concentration of attention on a single token within a sentence. While this induced a reduction in entropy, it adversely impacted overall performance. Conversely, exclusive reliance on the sentence regularization term led to an overly uniform attention distribution, manifesting as a diagonal attention pattern across the sentence.

These findings emphasize the necessity of striking a balance between the two regularization terms. The combination of both distance and sentence regularization proves instrumental in achieving the desired focused attention distribution, thus reinforcing the efficacy of our proposed approach in low-resource NMT scenarios. It is worth noting that while EaDRA$_{\text{enc+dec}}$ was used for this ablation study, it is conceivable that alternative configurations would have produced similar results.

Method	α_{dist}	α_{sent}	BLEU
T.base	0.00	0.00	13.5
EaDRA	0.02	0.80	16.2
EaDRA w/o α_{dist}	0.00	0.80	15.2
EaDRA w/o α_{dist}	0.02	0.00	0.7

Table 7: Ablation study results for English-German task with 100k training samples from WMT14. EaDRA$_{\text{enc+dec}}$ is used for this experiment.

C Convergence Speed Analysis

Given that EaDRA introduces a term into the loss function, it is imperative to assess its convergence speed. In Figure 3, we present the validation scores for two systems trained with 100k English-German samples from WMT14 on the same GPU.

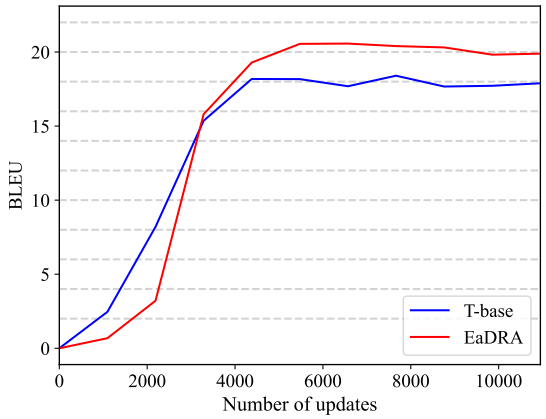

Figure 3: Convergence speed comparison on validation scores of EaDRA and T-base models trained on 100k English-German samples from WMT14.

The results demonstrate that EaDRA sustains a convergence speed comparable to the baseline. This observation underscores the efficiency of EaDRA in terms of convergence, further solidifying its viability in practical applications. This suggests that the incorporation of EaDRA does not come at the cost of prolonged training times, making it a practical choice for low-resource NMT tasks