Attention Mechanism with Energy-Friendly Operations

Anonymous ACL submission

Abstract

Attention mechanism has become the dominant module in natural language processing models. It is computationally intensive and depends on massive power-hungry multiplications. In this paper, we rethink variants of attention mechanism from the energy consumption aspects. After reaching the conclusion that the energy costs of several energy-friendly operations are far less than their multiplication counterparts, we build a novel attention model by completely replacing multiplications with either selective operations or additions. Empirical results on three machine translation tasks demonstrate that the proposed method, against the vanilla one, achieves comparable accuracy while only consumes a half of energy. Our code will be released upon the acceptance.

1 Introduction

Attention mechanism (ATT, Bahdanau et al., 2015; Vaswani et al., 2017; Kovaleva et al., 2019) has demonstrated huge success in a variety of natural language processing tasks (Kitaev and Klein, 2018; Tan et al., 2018; Devlin et al., 2019). The module learns hidden representations of a sequence by serving each word as a query to attend to all keys in the target sentence, then softly assembling their values. It is a de-facto standard to achieve this via performing linear projections and dot products on representations of queries and keys (Vaswani et al., 2017), resulting in large amount of multiplications. In spite of its promising quality, such kind of paradigm may be not the preferred solution from the energy consumption aspect (Horowitz, 2014; Raffel et al., 2020). How to build a high energy-efficient ATT still remains a great challenge.

Our work starts from in-depth investigations on approaches in ATT context with respect to model compression (Hinton et al., 2015; Jiao et al., 2020) and complexity optimization (Raganato et al., 2020; Tay et al., 2021; Beltagy et al., 2020). These approaches can potentially alleviate the problem of high energy consumption in ATT. Nevertheless, intentions of all these methods are not exactly from the energy-friendly perspective, thus overlooking the origin of energy consumed, i.e., basic arithmetic operations in electric equipments. Most of these approaches remain massive multiplications, which costs far more energy than its addition counterpart (Table 1, You et al., 2020).

Operation (FP32)	**ASIC**	**FPGA**
Addition	0.9	0.4
Multiplication	3.7	18.8

Table 1: Energy cost (pJ) of addition/multiplication operation on ASIC/FPGA hardware. 1 $pJ = 10^{-12} Joule$. Multiplication requires far more energy than addition.

To this end, we propose to approach this problem from a new direction – replacing massive multiplications in ATT with cheaper operations. Concretely, we propose a novel energy-efficient attention mechanism (E-ATT). It equips binarized selective operations instead of linear projections over input hidden states, and measures attentive scores using L_1 distance rather than dot-product. Consequently, E-ATT completely abandons multiplications to reach the goal of energy cost reduction.

We examine our method with Transformer model (Vaswani et al., 2017), and conduct experiments over three machine translation tasks. Compared with conventional ATT, our E-ATT can save over 99% energy during alignment procedure, and yield acceptable translation qualities across language pairs. Extensive analyses also demonstrate that E-ATT can functionally model semantic alignments without using multiplications.

2 Preliminary

Conventional Attention Mechanism Given input representations $\mathbf{X} \in \mathcal{R}^{l_1 \times d}$ and $\mathbf{Y} \in \mathcal{R}^{l_2 \times d}$ with l_1, l_2 being sequence length, and d is the in-

Model	# add	# mul	$\Delta(\%)$
Vanilla	$2ld^2 + l^2d$	$2ld^2 + l^2d$	100.00
Dense	$ld^2 + l^2d$	$ld^2 + l^2d$	51.10
RandInit	0	0	0.00
E-ATT	$2ld\rho + l^2d$	0	0.44

Table 2: Calls of addition (add) / multiplication (mul), and energy consumption ratio to derive attention scores in vanilla ATT (Vaswani et al., 2017), Dense (Tay et al., 2021), RandInit (Tay et al., 2021), and our model. Results are conducted over TRANSFORMER-*Base* setting. l, d and ρ are sequential length, model size, and the ratio of nonzero values after binarization, respectively.

put dimensionality. Note l_1 and l_2 may be equal for self-attention pattern, and represent lengths of target and source sequence in cross-attention. Conventional ATT first projects the inputs into three representations with multi-head property[1]:

$$\mathbf{Q} = \mathbf{X}\mathbf{W}_Q, \quad [\mathbf{K}; \mathbf{V}] = \mathbf{Y}[\mathbf{W}_K; \mathbf{W}_V], \quad (1)$$

$$\hat{\mathbf{Q}} = \textit{SplitHead}(\mathbf{Q}) \in \mathcal{R}^{h \times l_1 \times k_h}, \quad (2)$$

$$\hat{\mathbf{K}}; \hat{\mathbf{V}} = \textit{SplitHead}([\mathbf{K}, \mathbf{V}]) \in \mathcal{R}^{h \times l_2 \times k_h}, \quad (3)$$

where $\mathbf{W}_Q, \mathbf{W}_K \in \mathcal{R}^{d \times k}, \mathbf{W}_W \in \mathcal{R}^{d \times d}$ are parameters, k and k_h are the dimensionalities for linear transition and each head, $\mathbf{Q}$, $\mathbf{K}$ and $\mathbf{V}$ are query, key and value representations, respectively. For the m-th head, the attention alignment is calculated by obtaining the logits with dot-product multiplication, following by softmax activation:

$$\mathbf{A}^m_{ij} \propto \exp\left(\frac{\hat{\mathbf{Q}}^m_i \hat{\mathbf{K}}^{m\top}_j}{\sqrt{k_h}}\right) \in \mathcal{R}^{l_1 \times l_2}. \quad (4)$$

Then, the output is derived by multiplying attention weights with value representation $\hat{\mathbf{V}}$, concatenating heads and additional linear projection:

$$\hat{\mathbf{O}}^m = \mathbf{A}^m \hat{\mathbf{V}}^m \in \mathcal{R}^{l_1 \times h_k}, \quad (5)$$

$$\mathbf{O} = \textit{ConcatHead}(\hat{\mathbf{O}}) \in \mathcal{R}^{l_1 \times d}, \quad (6)$$

As seen, matrix multiplications are massively exploited into conventional ATT.

Related Work Several related approaches potentially alleviate the power-hungry drawback of ATT. One direction relies on model compression by pruning redundant parameters (Denton et al., 2014; Wang et al., 2016; Zhuang et al., 2018) or distilling the learned knowledge from a large model to a smaller one (Hinton et al., 2015; Yim et al., 2017), which still maintains multiplicative operations. Another direction aims at reducing the computational complexity of obtaining alignment logits, e.g. linearly projecting input (Dense, Tay et al., 2021), or randomly initializing and training (RandInit, Tay et al., 2021). Interestingly, Tay et al. (2021) point out that their approach fails to be employed into cross-attention networks. We think the main reason stems from the mechanism of these modules, where attention logits are obtained by using the linear transition of target representation, or randomly initialized parametric matrix. They can not strongly contribute aligned information across languages. To give a full comparison of energy consumption of those approaches, we conduct the number of additive and multiplicative operations across modules, as well as estimate the ratio of energy cost in Table 2. As seen, vanilla ATT (Vaswani et al., 2017) involves the most multiplicative operations, and requires the most energy than other methods. These ATT designs still suffer from considerable energy consumption.

3 Energy-Efficient Attention Mechanism

In this section, we describe E-ATT by pertinently reducing the multiplicative operations of ATT, including selective operation and L_1 distance.

3.1 Feature Selection with Discreteness

As the linear transitions of queries and keys (Equation 1∼3) involve massive multiplications within conventional ATT, we propose to modify them with binarized quantization (Liu et al., 2018; Qin et al., 2020). Concretely, the inputs $\mathbf{X} \in \mathcal{R}^{l_1 \times d}$ and $\mathbf{Y} \in \mathcal{R}^{l_2 \times d}$ are turned into discrete value with a threshold function $f(\cdot)$:

$$f(x) = \begin{cases} 1 & x > \tau, \\ 0 & \text{otherwise}, \end{cases} \quad (7)$$

where τ and d are threshold and hidden size, respectively. The derived representations $\tilde{\mathbf{X}} = f(\mathbf{X})$ and $\tilde{\mathbf{Y}} = f(\mathbf{Y})$ thus contain discrete representations composing of zeros and ones. Since this procedure is undifferentiable, we need to predefine a pattern of gradient calculation for $\mathbf{X}$ when receiving back-propagated gradient $\mathbf{Z}$. Inspired by recent work (Wu et al., 2018), here we use a modified Gaussian function during back-propagation:

$$\nabla \mathbf{X} = \sqrt{\frac{2}{\pi}} e^{-2(\mathbf{Z}-\tau)^2}, \quad (8)$$

[1]For simplicity we omit the bias term in related equations.

and the same procedure is applied for $\mathbf{Y}$. Then given parameters $\mathbf{W}_Q, \mathbf{W}_K \in \mathcal{R}^{d\times k}$, we derive query and key representations $\mathbf{Q}, \mathbf{K}$ by applying masked selection function:

$$\tilde{\mathbf{Q}} = g(\tilde{\mathbf{X}}, \tilde{\mathbf{W}_Q}) \in \mathcal{R}^{l_1\times d\times k}, \quad (9)$$

$$\tilde{\mathbf{K}} = g(\tilde{\mathbf{Y}}, \tilde{\mathbf{W}_K}) \in \mathcal{R}^{l_2\times d\times k}, \quad (10)$$

$$\mathbf{Q} = \sum_{i=1}^{d} \tilde{\mathbf{Q}}_{\cdot,i,\cdot}; \quad \mathbf{K} = \sum_{i=1}^{d} \tilde{\mathbf{K}}_{\cdot,i,\cdot}, \quad (11)$$

where $\tilde{\mathbf{W}_Q} \in \mathcal{R}^{l_1\times d\times k}$ and $\tilde{\mathbf{W}_K} \in \mathcal{R}^{l_2\times d\times k}$ are derived by tiling $\mathbf{W}_Q, \mathbf{W}_K$ with l_1 and l_2 times, respectively; and $g(\cdot,\cdot)$ represents indexed feature selection defined as follows:

$$g(\mathbf{U}, \mathbf{P}) = \begin{cases} \mathbf{U}_{i,j,\cdot} & \mathbf{P}_{i,j} = 1, \\ \mathbf{0} & \text{otherwise.} \end{cases} \quad (12)$$

3.2 Pairwise L_1 Distance

We further propose to use pairwise L_1 distance for measurements between $\mathbf{Q}$ and $\mathbf{K}$, which does not require any multiplication. Attention score calculation in Equation 4 is then modified as:

$$\mathbf{A}^m_{ij} \propto \exp(-\frac{||\hat{\mathbf{Q}}^m_i - \hat{\mathbf{K}}^m_j||_1}{\sqrt{k_h}}), \quad (13)$$

where $||\cdot||_1$ denotes the L_1 norm of inputted vector.

4 Experiments

4.1 Dataset and Model Setting

We choose three machine translation tasks, i.e. IWSLT'15 English-Vietnamese (En-Vi), WMT'14 English-German (En-De) and WMT'17 Chinese-English (Zh-En), to evaluate the effectiveness of our approach. We follow the setting of TRANSFORMER-*Base* (Vaswani et al., 2017) for all involved tasks, with model hidden size d as 512, and the number of layers in encoder and decoder as 6.

4.2 Experimental Results

As shown in Table 4, vanilla model achieves best performance over all translation tasks. However, replacing conventional attention networks with E-ATT does not lead to significant performance drop, with small decrease of 0.15∼0.78 BLEU score. Besides, after referring the statistics from Table 1 and 2, our E-ATT module only takes 0.44% energy of conventional ATT. These results reveal that, E-ATT can achieve comparative translation quality to baseline attention module, and more importantly, highly reduce the energy consumption in attention score calculation.

Model	**BLEU** (%)
Vanilla	**28.12**
Replace with discrete selection	27.51
Replace with L_1 distance	28.05
E-ATT	27.45

Table 3: Model performance with component replacements over En-Vi dev set. Using L_1 distance as similarity measurement slightly harms model performance.

4.3 Ablation Study

We further conduct ablation experiments on En-Vi task. As seen in Table 3, using discrete feature selection instead of linear transition does slight harm to performance, with 0.61 BLEU score decrease. Besides, replacing dot-product attention with L_1 distance does not significantly affect model performance, with only 0.07 BLEU score drop against baseline. We conclude that, L_1 distance can measure the similarity of vectorized representations and give modest performance compared to baseline.

5 Analyses

5.1 Hybrid Attention Networks

We first collect a series of experiments involving hybrids of attention networks among vanilla ATT, Dense, RandInit, and E-ATT module in Table 5. As shown, the conventional attention network performs the best among all models. Our module performs well when served as either self-attention or cross-attention modules. Besides, for all cases applying Dense/RandInit as cross-attention modules, models perform significantly worse, identical with the findings in Tay et al. (2021). On the contrary, our E-ATT module can give better performance with marginal performance drop comparing with baseline, indicating that E-ATT module is capable of providing adequate semantic alignments across languages for translation. Besides, it is encouraging to see that our method works compatibly with other modules with marginal performance drop.

5.2 Knowledge Distillation

To compare our model with knowledge distillation approaches (Hinton et al., 2015; Tang et al., 2019), we further conduct experiments on the various dimensionalities inside ATT. As shown in Figure 1, we simulate the energy consumption of each model with modified dimensionality d, and conduct the relationship between corresponding energy cost and

Attention mechanism	En-Vi	En-De	Zh-En	Energy (%)
Vanilla	30.26 ± 0.07	27.60 ± 0.04	24.28 ± 0.08	100.00
E-ATT	29.48 ± 0.08	27.45 ± 0.04	24.23 ± 0.06	**0.44**

Table 4: Averaged BLEU scores (%) upon test set on IWSLT'15 En-Vi, WMT'14 En-De and WMT'17 Zh-En tasks over 5 independent runs. E-ATT gives comparable results against conventional ATT, reducing the energy cost at 99.56% in alignment procedure. Since the energy cost for a specific module is difficult to be empirically evaluated, we report the theoretical values following the common practice (Chen et al., 2020; You et al., 2020).

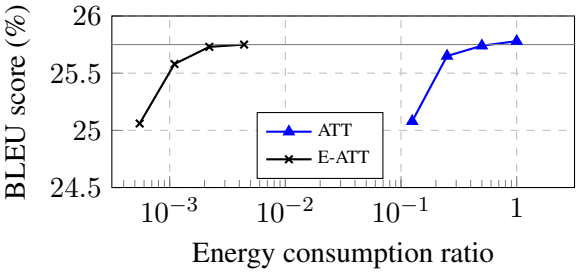

Figure 1: Performance of attention procedure with respect to the ratio of energy consumption on En-De task. We regard the energy consumption of ATT baseline as 1, and accumulatively halve the dimensionality of model till untrainable (from 512 to 64). E-ATT requires far less energy to meet up the baseline performance.

-	Vanilla	Dense	RandInit	E-ATT
Vanilla	**28.12**	19.92	19.31	27.72
Dense	27.48	19.43	19.21	27.60
RandInit	27.36	18.98	18.83	27.48
E-ATT	28.08	19.85	19.67	27.45

Table 5: BLEU score (%) of different model hybrids with modifying self-attention (horizontal) and cross-attention (vertical) network upon En-Vi dev set. E-ATT can achieve good performance when applied as cross-attention modules, whereas Dense or RandInit can not.

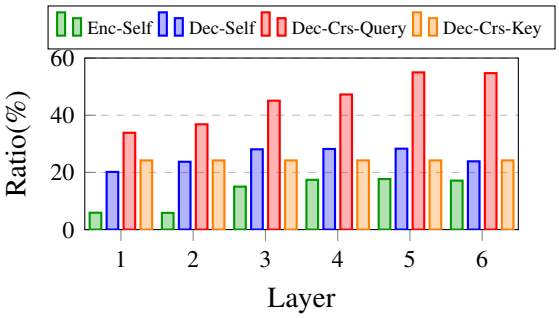

Figure 2: Ratio of nonzero values in the representations of E-ATT. Enc-Self: encoder self-attention ; Dec-Self: decoder self-attention; Dec-Crs-Query/Key: query/key representation for decoder cross-attention. Query representations in cross-attention are the most active.

performance. As seen, by accumulatively halving d from 512, both ATT and E-ATT significantly loses the performance. Besides, E-ATT saves around 99.45% energy compared to baseline to achieve comparable performance against ATT. This proves the advantage of our E-ATT, where it gives a better trade-off between model performance and energy consumption than knowledge distillation methods.

5.3 Binarization Statistics

We further collect the ratio of nonzero values ρ for each attention module in Figure 2, we can see that it increases with the number of encoder layers, denoting that more information is arranged into attentive calculation at higher layer of source side. However, for decoder E-ATT, the ratio meets its peak at middle layers, revealing that decoder E-ATT tends to focus on target semantics at the middle term of semantic processing. Interestingly, ratio in the query of cross-attention modules, which align source and target semantics, is higher for the layer closer to output. As the binarized key representation of each cross-attention module is equivalent, higher ratio of nonzero values in query representation means that, E-ATT at higher decoder layer provides more information for cross-lingual alignments, thus enrich the information for translation.

6 Conclusion

In this paper, we empirically investigate the high energy-consumption problem in ATT. We argue that the alignment modeling procedure can be completely achieved by additions other than multiplications, thus to reduce the energy costs. Experimental results demonstrate that our approach is able to save around 50% energy comparing with baseline, whereas yielding considerable quality. Extensive analyses suggest that: 1) Binarized representations marginally harm the feature extraction procedure; 2) L_1 distance can be efficiently exploited to measure alignment among queries and keys; and 3) Binarization ratio of E-ATT can give more intuitive and detailed description for model interpretability.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *International Conference on Learning Representations (ICLR)*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.

Hanting Chen, Yunhe Wang, Chunjing Xu, Boxin Shi, Chao Xu, Qi Tian, and Chang Xu. 2020. Addernet: Do we really need multiplications in deep learning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1468–1477.

Emily Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. 2014. Exploiting Linear Structure within Convolutional Networks for Efficient Evaluation. In *Neural Information Processing Systems (NIPS)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling The Knowledge in A Neural Network. *arXiv preprint arXiv:1503.02531*.

Mark Horowitz. 2014. 1.1 Computing's Energy Problem (and What We Can Do About It). In *IEEE International Solid-State Circuits Conference Digest of Technical Papers (ISSCC)*.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2020. TinyBERT: Distilling BERT for Natural Language Understanding. In *Findings of the Association for Computational Linguistics: EMNLP*.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with A Self-Attentive Encoder. In *Association for Computational Linguistics (ACL)*.

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the Dark Secrets of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.

Xuan Liu, Di Cao, and Kai Yu. 2018. Binarized LSTM Language Model. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*.

Haotong Qin, Ruihao Gong, Xianglong Liu, Xiao Bai, Jingkuan Song, and Nicu Sebe. 2020. Binary Neural Networks: A Survey. *Pattern Recognition*, 105:107281.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21:140:1–140:67.

Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2020. Fixed Encoder Self-Attention Patterns in Transformer-Based Machine Translation. In *Findings of the Association for Computational Linguistics: EMNLP*.

Zhixing Tan, Mingxuan Wang, Jun Xie, Yidong Chen, and Xiaodong Shi. 2018. Deep Semantic Role Labeling with Self-attention. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*.

Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. 2019. Distilling Task-Specific Knowledge from BERT into Simple Neural Networks. *arXiv preprint arXiv:1903.12136*.

Yi Tay, Dara Bahri, Donald Metzler, Da-Cheng Juan, Zhe Zhao, and Che Zheng. 2021. Synthesizer: Rethinking Self-Attention for Transformer Models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS)*.

Yunhe Wang, Chang Xu, Shan You, Dacheng Tao, and Chao Xu. 2016. CNNpack: Packing Convolutional Neural Networks in the Frequency Domain. In *Advances in Neural Information Processing Systems (NIPS)*.

Yujie Wu, Lei Deng, Guoqi Li, Jun Zhu, and Luping Shi. 2018. Spatio-Temporal Backpropagation for Training High-performance Spiking Neural Networks. *Frontiers in Neuroscience*, 12:331.

Junho Yim, Donggyu Joo, Jihoon Bae, and Junmo Kim. 2017. A Gift from Knowledge Distillation: Fast Optimization, Network Minimization and Transfer Learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haoran You, Xiaohan Chen, Yongan Zhang, Chaojian Li, Sicheng Li, Zihao Liu, Zhangyang Wang, and Yingyan Lin. 2020. ShiftAddNet: A Hardware-Inspired Deep Network. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhuangwei Zhuang, Mingkui Tan, Bohan Zhuang, Jing Liu, Yong Guo, Qingyao Wu, Junzhou Huang, and Jinhui Zhu. 2018. Discrimination-Aware Channel Pruning for Deep Neural Networks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

A Dataset Preprocessing

In this paper we evaluate our approach with three widely used machine translation datasets: IWSLT'15 English -Vietnamese (En-Vi), WMT'14 English - German (En-De) and WMT'17 Chinese - English (Zh-En). All datasets are segmented into subwords by byte-pair encoding (BPE, Sennrich et al., 2016) with 32k merge operations. Specially, for the former two tasks, we apply joint BPE for both source and target languages. All datasets are modified into truecase format with mosesdecoder by training truecase models upon train set.

Dataset	Train	Dev	Test
En-Vi	13.3K	1,553	1,268
En-De	4.50M	3,000	3,003
Zh-En	20.6M	2,002	2,001

Table 1: Dataset statistics. Each cell represents the number of examples. K: thousand, M: million.

B Experimental Setting

We apply TRANSFORMER-*Base* (Vaswani et al., 2017) setting for all experiments. The model dimensionality is 512, and 6 layers are engaged in both encoder and decoder side. The inner-connection dimensionality for feedforward block is 2,048, and the number of heads in multi-head attention networks is 8. We share the source embedding, target embedding and target softmax projection weight for En-Vi task, and share the latter two matrices for En-De. We modify the learning rate schedule as: $lr = 0.001 \cdot \min\left(\frac{t}{8000}, 1, (\frac{20000}{t})^{0.5}\right)$, where t denotes the current step. Across all tasks, we determine the threshold τ as 1.0.

For both baseline and our model, En-Vi, En-De and Zh-En tasks take 50k, 150k and 200k updates, and each batch contains 4,096, 32,768 and 32,768 source tokens. The dropout ratio is determined as 0.3, 0.1 and 0.1, respectively. All experiments are conducted over 4 NVIDIA V100 GPUs. For each task, we choose the best model over dev set, defining beam size as 4, 4, 10 and decoding alpha as 1.5, 0.6 and 1.5, respectively.

C Case Study

We visualize the averaged attention values over one case from WMT'17 Zh-En dev set. As seen, our model can give good aligned information, where

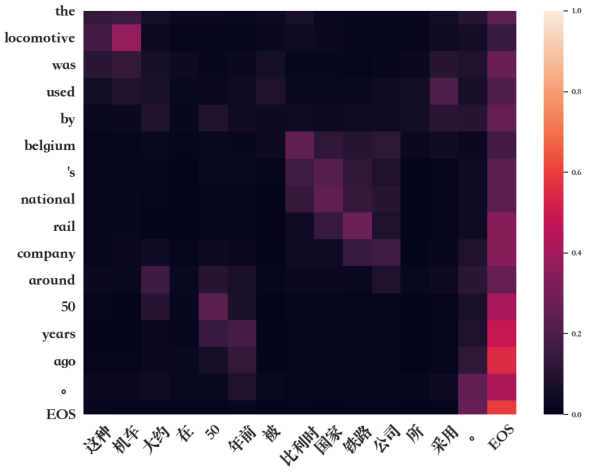

Figure 1: Case study from WMT'17 Zh-En dev set.

preposition phrase "around 50 years ago" is arranged at the end of sentence in English, while its aligned phrase is at the front in Chinese.

References

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. In *ACL*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems (NIPS)*.