Two-Headed Monster And Crossed Co-Attention Networks

Yaoyiran Li
Language Technology Lab, TAL
University of Cambridge
yl711@cam.ac.uk

Jing Jiang
Living Analytics Research Centre
Singapore Management University
jingjiang@smu.edu.sg

Abstract

This paper investigates a new co-attention mechanism in neural transduction models for machine translation tasks. We propose a paradigm, termed Two-Headed Monster (THM), which consists of two symmetric encoder modules and one decoder module connected with co-attention. As a specific and concrete implementation of THM, Crossed Co-Attention Networks (CCNs) are designed based on the Transformer model. We test CCNs on WMT 2014 EN-DE and WMT 2016 EN-FI translation tasks and show both advantages and disadvantages of the proposed method. Our model outperforms the strong Transformer baseline by 0.51 (big) and 0.74 (base) BLEU points on EN-DE and by 0.17 (big) and 0.47 (base) BLEU points on EN-FI but the epoch time increases by circa 75%.

1 Introduction

Attention has emerged as a prominent mechanism extensively adopted in neural modules in a wide range of research problems (Das et al., 2017; Hermann et al., 2015; Rocktäschel et al., 2015; Santos et al., 2016; Xu and Saenko, 2016; Yang et al., 2016; Yin et al., 2016; Zhu et al., 2016; Xu et al., 2015; Chorowski et al., 2015) such as VQA, reading comprehension, textual entailment, image captioning and speech recognition. Its remarkable success is also embodied in machine translation tasks (Bahdanau et al., 2014; Vaswani et al., 2017).

This work proposes an end-to-end co-attentional neural structure named Crossed Co-Attention Networks (CCNs) to address machine translation, a typical sequence-to-sequence NLP task. We customize the transformer (Vaswani et al., 2017) featured by non-local operations (Wang et al., 2018) with two input branches and tailor the transformer's multi-head attention mechanism to the needs of information exchange between these two parallel branches.

A higher-level and more abstract paradigm generalized from CCNs is denoted as "Two-Headed Monster" (THM), representing a broader class of neural structures benefiting from two parallel neural channels that would be intertwined with each other through co-attention mechanism as illustrated in Fig. 1.

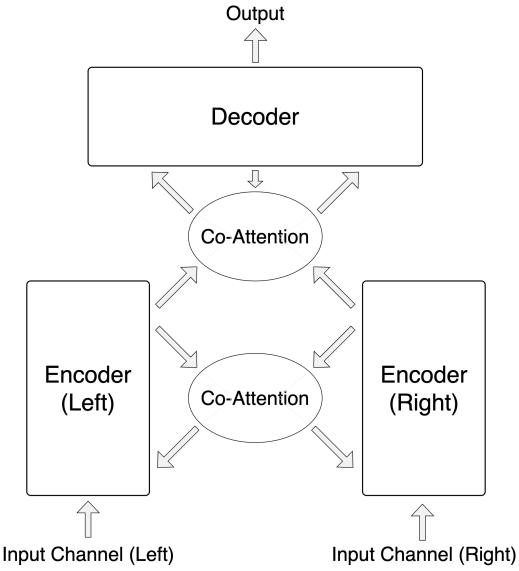

Figure 1: Two-Headed Monster.

Needless to say, co-attention is widely adopted in multi-modal scenarios (Lu et al., 2016a; Yu et al., 2017; Tay et al., 2018; Xiong et al., 2016; Lu et al., 2016b), the basic idea of which is to make two feature maps from different domains to attend to each other and thus output summarized representations for each domain. In this work, we emphasize a parallel and symmetric manifold operating on two input channels and possessing two output channels but do not assume that the two channels of input must be disparate. Our co-attention mechanism is designed in a "Transformer" style, and to the best of our knowledge, our proposed Crossed Co-Attention Network is one of

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop, pages 8–15
December 4 - 7, 2020. ©2020 Association for Computational Linguistics

the first implementations of co-attention on transformer model (Lu et al., 2019; Tan and Bansal, 2019) and we are the first to apply transformer-based co-attention on machine translation tasks. In particular, we apply our model on the popular WMT machine translation tasks where two input channels are in one same domain. Our code also leverages half-precision floating-point format (FP16) training and synchronous distributed training for inter-GPU communication (we do not discard gradients calculated by "stragglers") which dramatically accelerate our training procedure (Ott et al., 2018; Micikevicius et al., 2018).

2 Model Architecture

In this section, we first define co-attention as a generic concept, following Wang et al. (2018)'s definition of non-local operation. After that, we propose an end-to-end neural architecture based on the transformer to address machine translation tasks where the model takes input from two channels. In particular, we design a Crossed Co-Attention Mechanism to make our model capable of attending to two parallel information flows simultaneously in both the encoding and the decoding stages. Our co-attention mechanism is naively realized by a crossed connection of Value (V), Key (K) and Query (Q) gates of a regular multi-head attention module, so we term our model Crossed Co-Attention Networks.

2.1 Generic Co-Attention

In this section, we first review a non-local operation and bridge it to the dot-product attention that is widely used in self-attention modules and then formulate the co-attention mechanism in a generic way. A non-local operation is defined as a building block in deep neural networks that captures long-range dependencies where every response is computed as a linear combination of all features in the input feature map (Wang et al., 2018). Suppose the input feature maps are $V = [v_1, v_2, \dots, v_n]^T \in \mathbb{R}^{n\times d}$, $K = [k_1, k_2, \dots, k_n]^T \in \mathbb{R}^{n\times d}$ and $Q = [q_1, q_2, \dots, q_n]^T \in \mathbb{R}^{n\times d}$ and the output feature map $Y = [y_1, y_2, ..., y_n]^T \in \mathbb{R}^{n\times d}$ is of the same size as the input. Then a generic non-local operation is formulated as follows:

$$y_i = \frac{1}{C(q_i, K)} \sum_j f(q_i, k_j) g(v_j). \quad (1)$$

We basically follow the definition of no-local operation by Wang et al. (2018) where $f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a pairwise function ("$\times$" is Cartesian product), $g : \mathbb{R}^d \to \mathbb{R}^d$ is a unary function and $C : \mathbb{R}^d \times \mathbb{R}^{n\times d} \to \mathbb{R}$ calculates a normalizer, but dispense with the assumption that $V = K = Q$. However, if we assume $f(q_i, k_j) = e^{(q_i^T W^Q)\cdot(k_j^T W^K)^T}$, $g(v_i) = v_i^T W^V$, the normalizer $C(q_i, K) = \sum_j f(q_i, k_j)$ and $V = K = Q$, then the non-local operation degrades to the multi-head self-attention as is described in (Vaswani et al., 2017) (formula 2 describes only one attention head):

$$Y = \text{softmax}(QW^Q(KW^K)^T)VW^V. \quad (2)$$

Considering two input channels, denoted as 'left' and 'right', we present the following operation as a definition of a generic co-attention where $\alpha(\cdot), \beta(\cdot)$ define if the input to V,K and Q input gates are from 'left' or 'right' encoder branches, or 'decoder' branch respectively, directing the connections of co-attention.

$$y_i^{\text{left}} = \frac{1}{C^{\text{left}}(q_i^{\alpha(Q)}, K^{\alpha(K)})} \cdot \sum_j f^{\text{left}}(q_i^{\alpha(Q)}, k_j^{\alpha(K)}) g^{\text{left}}(v_j^{\alpha(V)}), \quad (3)$$

$$y_i^{\text{right}} = \frac{1}{C^{\text{right}}(q_i^{\beta(Q)}, K^{\beta(K)})} \cdot \sum_j f^{\text{right}}(q_i^{\beta(Q)}, k_j^{\beta(K)}) g^{\text{right}}(v_j^{\beta(V)}). \quad (4)$$

Under the umbrella of Two-Headed Monster (THM), as in Fig. 1, for the co-attention serving encoders, $\alpha(\cdot), \beta(\cdot)$ take value from {'left', 'right'} and for encoder-decoder co-attention, they take value from {'left', 'right', 'decoder'}. Note that when $\alpha(\cdot)$ = 'left', $\beta(\cdot)$ = 'right' the co-attention degrades to two self-attention modules in Transformer encoders. Another example of $\alpha(\cdot), \beta(\cdot)$ is a crossed connection which we will introduce in Section 2.2 as illustrated in Fig. 2. In its encoder's co-attention, $\alpha(V)$ = $\alpha(K)$ = 'left' and $\alpha(Q)$ = 'right', $\beta(V)$ = $\beta(K)$ = 'right' and $\beta(Q)$ = 'left'. For the encoder-decoder co-attention, however, $\alpha(V)$ = 'left', $\alpha(K)$ = 'right', $\beta(V)$ = 'right', $\beta(K)$ = 'left' and $\alpha(Q)$ = $\beta(Q)$ = 'decoder'.

2.2 Crossed Co-Attention Networks

Our implementation of CCN, based on Transformer, consists of two symmetrical branches, working in parallel. Different from previously

Figure 2: Crossed Co-Attention Networks.

known co-attentions such as (Xiong et al., 2017; Lu et al., 2016a), our design is built through connecting two multiplicative attention modules (Vaswani et al., 2017), each containing three gates, i.e., V, K and Q. The information flows from two input channels and then interact with and benefit from each other via crossed connections. Suppose the input fed into the left branch is X_{left}, and the right branch X_{right}. In our encoder, the left branch takes input from X_{left} as V and K and takes the input X_{right} as Q. The right branch, however, takes the input X_{left} as Q and X_{right} as V and K (each way of connection corresponds to a choice of $\alpha(\cdot), \beta(\cdot)$). This design is, in a sense, meant for the two branches to relatively keep the information in their own domains. A special case is, if $g(v_i) = v_i$, then the response y_i will be in the row space of V. Because when an attention takes input V from its own branch, the output responses will, by and large, carry the information of the branch. In machine translation tasks, the two encoder branches take in one same input sequence, but in order to reduce the redundancy of two parallel branches, we apply dropout and input corruption on input embeddings for two branches respectively. While our model shares BPE embeddings (Sennrich et al., 2015) globally, we randomly swap two sub-word tokens in the input matrices at a probability of 0.5.

In the encoder-decoder co-attention layers, the multi-head attention on two decoder branches uses the output from two encoder branches as V and K alternatively while taking in the output of the self-attention layer in the decoder as Q. The output of the two branches in the decoder is processed through concatenation, linear transformation and then fed into a feed-forward network. The self-attention layer in the decoder is also used for reading shifted output embedding into the network. We adopt the same input masking and sinusoidal position encoding as the Transformer, which will not be expanded here.

Although we connect the attention gates the way we describe above, there are plenty of other choices of connections for the co-attention when $\alpha(\cdot), \beta(\cdot)$ vary. We find that some different selections may produce similarly good results on some NMT datasets but this work, as a preliminary investigation, only demonstrates the case specified above in Fig. 2 but leaves the study of how different ways of connections cater to different tasks for our future work.

Model	Dataset	Epoch Time (s)	BLEU	Number of Parameters	Batch Size
Transformer-Base	WMT2014 EN-DE	**684.52**	27.21	61,364,224	6,528
THM / CCN-Base	WMT2014 EN-DE	1090.65	**27.95**	114,928,640	6,528
Transformer-Base	WMT2016 EN-FI	**232.97**	16.12	55,883,776	6,528
THM / CCN-Base	WMT2016 EN-FI	410.79	**16.59**	109,448,192	6,528
Transformer-Big	WMT2014 EN-DE	**1982.63**	28.13	210,808,832	2,176
THM / CCN-Big	WMT2014 EN-DE	3611.53	**28.64**	424,892,416	2,176
Transformer-Big	WMT2016 EN-FI	**726.51**	16.21	199,847,936	2,176
THM / CCN-Big	WMT2016 EN-FI	1387.22	**16.38**	413,931,520	2,176

Table 1: Comparisons between our proposed method and Transformer baseline on WMT 2014 EN-DE and WMT 2016 EN-FI.

3 Experiments

3.1 Setup

We demonstrate our model on WMT 2014 EN-DE and WMT 2016 EN-FI machine translation tasks. For convenience, in this section, we do not differentiate between the notion of THM and CCN (which is an implementation of THM). The raw input data is pre-processed with length filtering as previous work (Ott et al., 2018). Our final dataset consists of $4,575,637$ training examples, $3,000$ validation examples and $3,003$ test examples for EN-DE, and $2,073,194$ training examples, $1,500$ validation examples and $3,000$ test examples for EN-FI. Considering the scale of the training sets, we adopt shared BPE dictionaries of size $33,712$ for EN-DE and $23,008$ for EN-FI. Our CCNs are established with 6 encoder and decoder blocks and a hidden state of size 512 for base models and with also 6 such blocks but a hidden state of $1,024$ neurons for big models. That exactly corresponds to the settings of the Transformer paper (Vaswani et al., 2017). We train our models on an NVIDIA DGX-1 GPU server with 4 TESLA V100-16GB GPUs. In order to make full use of the computational resources, FP16 computation is adopted, and we use a batch size of $6,528$ tokens/GPU for base models and $2,176$ for big models (both Transformer and THM). We adopt the Sequence-to-Sequence Toolkit FairSeq (Ott et al., 2019) released by the Facebook AI Research for our Transformer baseline[1], upon which our THM code is built as well. We train all base models for around one day and big models for around two days. For model selection, we strictly choose the model that achieves the highest BLEU on Dev set.

[1]https://github.com/pytorch/fairseq

3.2 Experimental Results

Main Results: Our experiments demonstrate the efficiency of our proposed crossed co-attention mechanism, which significantly improves the BLEU scores of machine translation as illustrated in Table 1. Besides, the co-attention mechanism has, by and large, reduced training, validation and test loss from the first training epoch compared with the transformer baselines as shown in Fig. 3,4,5,6. However, since the number of parameters doubles, the epoch time also increases by roughly $60\% \sim 80\%$.

Capability of Model Selection: In addition to the BLEU, loss and time efficiency, we also find that the THM/CCN models demonstrate a better capability of selecting good models with Dev set from all models derived in all training epochs. As is shown in Table 2, for THM/CCN, the models that achieved the highest BLEU on Dev set are also high-ranking on the Test set. In 75% of the cases, THM will select TOP 3 models, and in all cases, it will select TOP 10 models whereas Transformer can only select TOP 10 models in 50% of the cases.

Performance across Languages: We test our proposed method on two language pairs, EN-DE and EN-FI, and the improved BLEU scores and the capability of model selection on both base and big models demonstrate the effectiveness of our proposed method.

4 Related Work

Attention: Multi-head self-attention has demonstrated its capacity in neural transduction models (Vaswani et al., 2017), language model pre-training (Devlin et al., 2018; Radford et al., 2018) and speech synthesis (Yang et al., 2019c). While the novel attention mechanism, eschewing re-

	THM / CCN	Transformer
TOP 1	**25%**	0
TOP 3	**75%**	0
TOP 5	**100%**	0
TOP 10	**100%**	50%

Table 2: This table evaluates if the models selected by the Dev set are also better than others on the test set. Here we provide the percentage of selected models that rank TOP 1, TOP 3, TOP 5 or TOP 10 among all models derived from all training epochs.

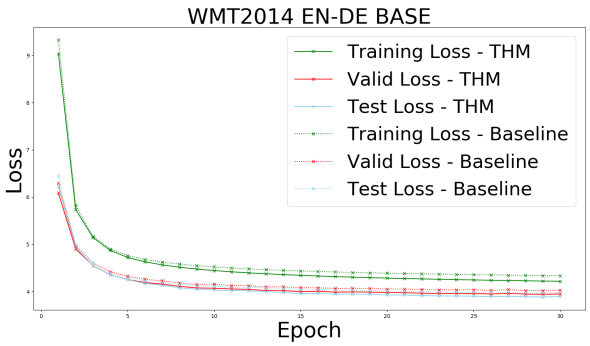

Figure 3: Loss vs Epoch for THM-base and Transformer-base on EN-DE

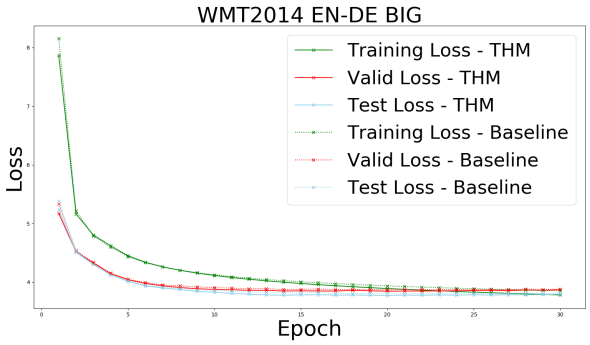

Figure 4: Loss vs Epoch for THM-big and Transformer-big on EN-DE

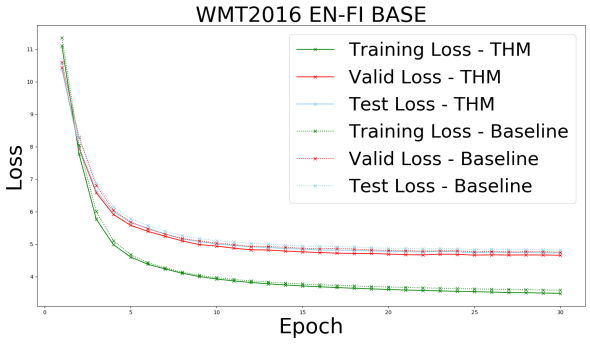

Figure 5: Loss vs Epoch for THM-base and Transformer-base on EN-FI

currence, is famous for modeling global dependencies and considered faster than recurrent lay-

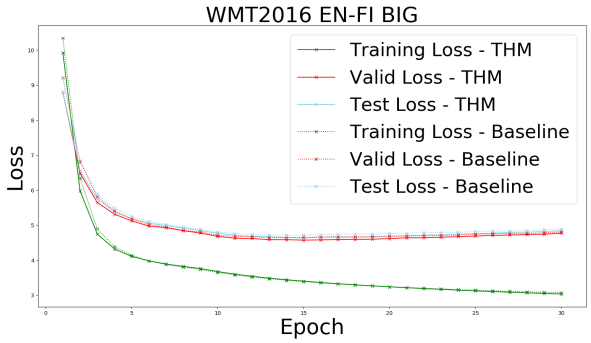

Figure 6: Loss vs Epoch for THM-big and Transformer-big on EN-FI

ers (Vaswani et al., 2017), recent work points out that it may tend to overlook neighboring information (Yang et al., 2019a; Xu et al., 2019). It is found that applying an adaptive attention span could be conducive to character level language modeling tasks (Sukhbaatar et al., 2019). Yang et al. (2018a) propose to model localness for self-attention, which would be conducive to capturing local information by learning a Gaussian bias predicting the region of local attention. Other work indicates that adding convolution layers would ameliorate the aforementioned issue (Yang et al., 2018b, 2019b). Multi-head attention can also be used in multi-modal scenarios when V, K and Q gates take in data from different domains. Helcl et al. (2018) adds an attention layer on top of the encoder-decoder layer with K and V being CNN-extracted image features. Lu et al. (2019) and Tan and Bansal (2019) use co-attention on BERT to learn multi-modal representations for vision-and-language tasks.

Machine Translation: Some recent advances in machine translation aim to find more efficient model architecture based on the Transformer: Hao et al. (2019) add an additional recurrence encoder to model recurrence for Transformer; So et al. (2019) demonstrate the power of neural architecture search and find that the found evolved transformer architecture outperforms human-designed ones; Wu et al. (2019) propose dynamic convolutions that would be more efficient and simpler compared with self-attention. Lample and Conneau (2019) and Liu et al. (2020) investigate how language pretraining can improve machine translation performance. Zhu et al. (2020) studies how to incorporate BERT into machine translation tasks. Other work shows that training on 128 GPUs can significantly boost the experimental results and shorten

the training time (Ott et al., 2018). A novel research direction is semi- or un-supervised machine translation aimed at addressing low-resource languages where parallel data is usually unavailable (Cheng, 2019; Artetxe et al., 2017; Lample et al., 2017).

5 Conclusion and Future Work

We propose Two-Headed Monster (THM) neural structures consisting of two parallel encoders and a decoder connected with each other using co-attention, which is the core module for information flow in THM. First, we formulate the co-attention in a general sense as a non-local operation. Then we design Crossed Co-attention Networks (CCNs) based on our defined co-attention under the umbrella of THM. Our implementation of CCN adopts a specific way of connections of attention gates and improve the machine translation tasks by $0.17 \sim 0.74$ BLEU points and enhance the capability of model selection. However, the time efficiency is reduced since the number of parameters increases.

While we investigate the advantages and disadvantages of our proposed model on machine translation, we still think that the work is at a preliminary stage and our future work will focus on two aspects: (1) adopt input from two different domains and test our model on tasks such as multi-modal machine translation; (2) investigate how different ways of co-attention connections will influence the model performance on different tasks.

Acknowledgments

We thank all anonymous reviewers for their valuable comments and suggestions on our paper. We also thank Dr. Sonit Singh who serves as our mentor in AACL-IJCNLP SRW.

References

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2017. Unsupervised neural machine translation. *arXiv preprint arXiv:1710.11041*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yong Cheng. 2019. Semi-supervised learning for neural machine translation. In *Joint Training for Neural Machine Translation*, pages 25–40. Springer.

Jan K Chorowski, Dzmitry Bahdanau, Dmitriy Serdyuk, Kyunghyun Cho, and Yoshua Bengio. 2015. Attention-based models for speech recognition. In *Advances in neural information processing systems*, pages 577–585.

Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. 2017. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jie Hao, Xing Wang, Baosong Yang, Longyue Wang, Jinfeng Zhang, and Zhaopeng Tu. 2019. Modeling recurrence for transformer. *arXiv preprint arXiv:1904.03092*.

Jindřich Helcl, Jindřich Libovickỳ, and Dušan Variš. 2018. Cuni system for the wmt18 multimodal translation task. *arXiv preprint arXiv:1811.04697*.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *Advances in neural information processing systems*, pages 1693–1701.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *arXiv preprint arXiv:2001.08210*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016a. Hierarchical question-image co-attention for visual question answering. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 289–297. Curran Associates, Inc.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016b. Hierarchical question-image co-attention for visual question answering. In *Advances In Neural Information Processing Systems*, pages 289–297.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, and Hao Wu. 2018. Mixed precision training. In *International Conference on Learning Representations*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Myle Ott, Sergey Edunov, David Grangier, and Michael Auli. 2018. Scaling neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 1–9.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning. Technical report, Technical report, OpenAI.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomáš Kočiskỳ, and Phil Blunsom. 2015. Reasoning about entailment with neural attention. *arXiv preprint arXiv:1509.06664*.

Cicero dos Santos, Ming Tan, Bing Xiang, and Bowen Zhou. 2016. Attentive pooling networks. *arXiv preprint arXiv:1602.03609*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

David R So, Chen Liang, and Quoc V Le. 2019. The evolved transformer. *arXiv preprint arXiv:1901.11117*.

Sainbayar Sukhbaatar, Edouard Grave, Piotr Bojanowski, and Armand Joulin. 2019. Adaptive attention span in transformers. *arXiv preprint arXiv:1905.07799*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Yi Tay, Anh Tuan Luu, and Siu Cheung Hui. 2018. Multi-pointer co-attention networks for recommendation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2309–2318. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. *CVPR*.

Felix Wu, Angela Fan, Alexei Baevski, Yann N Dauphin, and Michael Auli. 2019. Pay less attention with lightweight and dynamic convolutions. *arXiv preprint arXiv:1901.10430*.

Caiming Xiong, Victor Zhong, and Richard Socher. 2016. Dynamic coattention networks for question answering. *arXiv preprint arXiv:1611.01604*.

Caiming Xiong, Victor Zhong, and Richard Sochern. 2017. Dynamic coattention networks for question answering. In *International Conference on Learning Representations*.

Huijuan Xu and Kate Saenko. 2016. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057.

Mingzhou Xu, Derek F Wong, Baosong Yang, Yue Zhang, and Lidia S Chao. 2019. Leveraging local and global patterns for self-attention networks. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3069–3075.

Baosong Yang, Jian Li, Derek F Wong, Lidia S Chao, Xing Wang, and Zhaopeng Tu. 2019a. Context-aware self-attention networks. *arXiv preprint arXiv:1902.05766*.

Baosong Yang, Zhaopeng Tu, Derek F Wong, Fandong Meng, Lidia S Chao, and Tong Zhang. 2018a. Modeling localness for self-attention networks. *arXiv preprint arXiv:1810.10182*.

Baosong Yang, Longyue Wang, Derek Wong, Lidia S Chao, and Zhaopeng Tu. 2019b. Convolutional self-attention networks. *arXiv preprint arXiv:1904.03107*.

Shan Yang, Heng Lu, Shiying Kang, Lei Xie, and Dong Yu. 2019c. Enhancing hybrid self-attention structure with relative-position-aware bias for speech synthesis. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6910–6914. IEEE.

Zhilin Yang, Jake Zhao, Bhuwan Dhingra, Kaiming He, William W Cohen, Ruslan Salakhutdinov, and Yann LeCun. 2018b. Glomo: Unsupervisedly learned relational graphs as transferable representations. *arXiv preprint arXiv:1806.05662*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. Abcnn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for Computational Linguistics*, 4:259–272.

Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 1821–1830.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.