Proceeding of the IEEE
International Conference on Robotics and Biomimetics
Dali, China, December 2019

Positional Self-attention Based Hierarchical Image Captioning

Qianxia Ma, Jingyan Song and Tao Zhang
Department of Automation
Tsinghua University
Beijing, China
mqx15@mails.tsinghua.edu.cn, {jysong, taozhang}@mail.tsinghua.edu.cn

***Abstract*— Intelligent robotic systems process images captured by camera devices to get high-level semantic concepts. Current image captioning approaches use convolutional neural networks as encoder and recurrent neural networks as decoder. Various modalities are treated differently and parallel computation is not fully allowed due to the chain architecture. To address these issues, we present a parallel hierarchical neural network based on encoder-decoder architecture to generate descriptions for images. Only convolutional neural networks are adopted here, and the hierarchical framework takes less steps to capture distant dependencies. Masked positional self-attention mechanism is utilized in the decoder to improve the performance. The fixed sized windows make it possible for parallel computation. This generative model uses unified architecture for different modalities, reaching a BLEU-1 score (the higher the better) over 0.7 with a higher training speed.**

***Index Terms*— computer vision; natural language processing; multi-modal machine learning;**

I. INTRODUCTION

More recently, end-to-end learning has shown its superiority and cross-domain artificial intelligence algorithm has become a research hotspot. The advent of we-media era makes social platforms rooted in our daily life. Information propagates through microblogging service such as Weibo, Twitter, Facebook every day in diverse modalities. Texts, images, videos and voice messages are used, which poses challenges to multimodal machine learning (MMML). Translation is the process of transferring one modal information into another, and is called the top five challenges in multimodal machine learning with representation, alignment, fusion and co-learning[1]. As a typical example, image caption aims at automatically translating an image into textual descriptions using machine learning, fusing computer vision and natural language processing. Image caption undoubtedly has been one of the most elusive tasks for decades. In 2016, image caption became a hotspot on CVPR.

As a visual-to-language (V2L) problem, image caption is a fusion of computer vision, natural language processing and machine learning. Given one or a series of pictures, the model could generate corresponding caption for each image automatically. A good caption needs to contain central or dominant objects, their attributes and scene. There seems no swear on such a job for human beings, but it is indeed an elusive task in artificial intelligence, calling for a large quantity of cross-domain effort. The principle of image caption is similar to scene understanding. The model needs not only the ability of detecting the objects and capturing their attributes in the image, but also expression ability to organize the words. Furthermore, one image contains abundant information. A practical model has to make to-the-point selections.

Encoder-decoder model is a widely-used end-to-end architecture in image caption. Most of these models share the following features: the encoder part uses convolutional neural networks to extracte features, and the decoder part utilize recurrent neural networks to generate sequences. m-RNN model[2] and Neural Image Caption (NIC)[3], for example, are both typical models of this kind. Although GRU, LSTM and other variants of recurrent neural networks capture more dependencies of distant words, the long-time memory ability is highly limited, especially when it comes to long sequence. Besides, in training phase, the sequence generation process follows a strict order, making it a time-consuming operation.

In this paper, we propose a novel image caption model based on encoder-decoder architecture. Instead of using different types of neural networks to process different modalities, we use convolutional neural networks both in the encoder and the decoder part. Thus, parrallel computation is realized when generating the captions in training. In order to catch more precise dependencies of distant elements, we propose a masked positional self-attention mechanism.

Our contributions. Firstly, we present a new end-to-end image caption model, using only convolutional neural networks both in encoder and decoder, which makes parallel computation possible in training. Secondly, a new attention mechanism, masked positional self-attention, is introduced in our model to improve the performance. Last but not the least, our model is a unified architecture to process different modalities, providing a new idea for further multimodal machine learning applications in robotic systems.

II. RELATED WORK

As one of the typical representatives of multimodal problems, image caption only enjoys a history of around one decade. Nevertheless, in such a short period, image caption models have seen rapid development and several significant

978-1-7281-6321-5/19//$31.00 ©2019 IEEE

changes. Early researches tended to decompose the issue into subproblems in separate domains, which was replaced by end-to-end learning by degrees to boom the training speed.

Before the problem was put forward formally, related works have been done. In 2009, Li et. al. applied multi instance learning in regions of interest (ROI) problems, which was the state-of-the-art locating method[4]. Early image caption methods could be classified into two main categories. One would be object detection and attribute recognition based methods[5], [6], [7]. Triplet, usually contains object, action and scene, was frequently used as the representation[8], [9]. The other one uses the captions of similar image retrieval with minor adjustments[10], [11]. Summarization based methods[12] is similar to retrieval-based approaches.

Both of these two conventional types have shortcomings. The first type is based on object detection. Thus, it has no access to abstract objects. And in order to return grammatically right sentences, phrase fusion, parsing and language models are of significant importance. Additionally, basic grammar is preferred in such models, which leads to simple sentence structures. The second type highly depends on the pre-existing descriptions from the database, and in most occasions is lacking in creativity and richness. Besides, captions are descriptive and compact representations, neither of the above-mentioned methods.

The development of image caption has rocketed, thanks to Deep Neural Networks (DNNs). Kiros et al. first utilized neural networks to address this issue in 2014[13]. In the same year, they improved the multimodal model for retrieval ranking and description generation[14]. Mao et al. adapted the architecture and replaced the forward propagation network by recurrent neural network[2]. Since then, many attempts have found that deeper neural networks or change of constructions may boost the performance. For instance, Gated Recurrent Units (GRU), Long Short-term Memory (LSTM)[15], bidirectional LSTM[16] and phi-LSTM[17] outperform simple Recurrent Neural Networks. And deeper convolutional neural networks[3] also improve the precision of prediction.

Encoder-decoder mechanism is the most commonly used neural machine translation architecture. A fixed-length vector is extracted from the encoder and then put into the decoder to produce the output sequence[18], [19]. The greatest limitation is that all the information flows from encoder to decoder only relies on the fixed-length vector. Machine translation suffers from performance reduction as the length of sequences increases[20]. Attention mechanism was raised to address this issue[21]. Models could focus on different part of the input and thus increase prediction precision.

Image caption has realized remarkable development with the help of semantic features of high level recently. Yao et al. studied caption framework using Long Short-Term Memory with Attributes (LSTM-A)[22]. Jiang et al. developed Guiding Network, putting attributes of images into consideration[23]. Rupprecht et al. used user responses to update the activation values. Thus the precision of the model could be raised without retraining.

III. HIERARCHICAL CONVOLUTIONAL MODEL

We propose a novel hierarchical encoder-decoder model for image caption, which only uses convolutional neural networks to process different modalities, both in encoder and decoder. Fig. 1 shows an overview of our model architecture. Details of each part are discussed below.

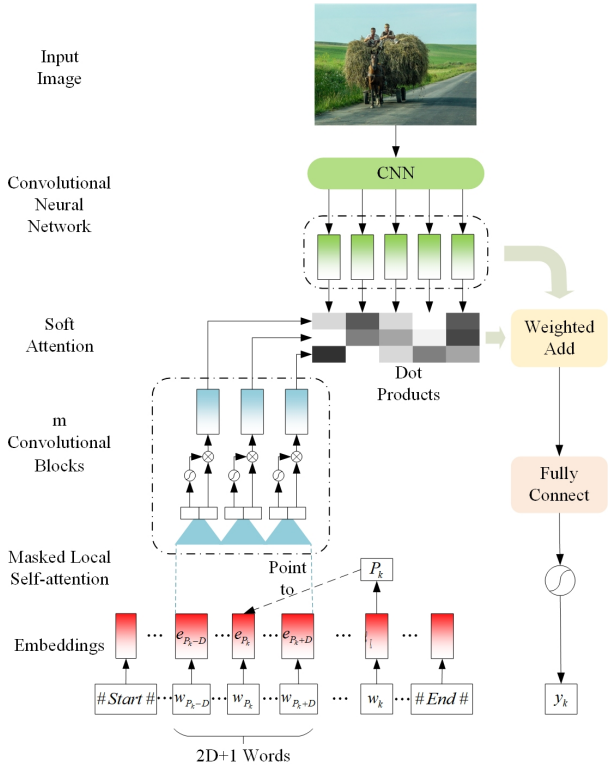

Fig. 1. Hierarchical image caption model architecture

A. Hierarchical Structure

Chain structure, Recurrent neural network for example, is the mainstream in neural network methods in natural language processing. RNNs share the same parameters at each time step, and is able to deal with variable-length sequence problems with memory capacity. Although subsequent GRU and LSTM have enhanced memory ability, such function is still limited.

In each layer of convolutional neural network, only features of several nearby elements can be created. As a result, convolutional neural networks are not as common as recurrent neural networks in sequence modelling, especially for long sequences. However, with hierarchical structure, the higher layer of convolutional neural networks can capture the dependencies of distant elements. The greatest advance of

using hierarchical structure is that less distance is sufficient to capture the relationship. Take CNN as an example, assume we use kernels of size k, we only need $\mathcal{O}(\frac{n}{k})$ to get the dependencies of words in the window of size n. But as for RNNs, $\mathcal{O}(n)$ is needed instead.

A gated convolutional machine translation network, which outperforms strong recurrent models, was first put forward by Dauphin et al. in 2016[24]. Inspired by ConvS2S model[25], which is developed by Facebook to address sequence to sequence learning, we developed a hierarchical image caption model. Instead of using RNNs when generating sequences, only convolutional neural networks are used in our model, which makes it possible to compute parallelly.

Meanwhile, to increase precision, soft attention and masked positional self-attention are introduced. In ConvS2S model, every element in the sequence is considered as part of the input when predicting the next word. According to the results of multi-head attention[26], One word has strong connections with several certain words. Other elements do not have any great influences for predictions. It would be more efficient to find the major dependencies. Inspired by local attention mechanism, we developed masked positional self-attention to work with convolutional neural network to form a parallel hierarchical model.

Our hierarchical model is based on the encoder-decoder structure. When the image features I are captured by deep convolutional neural network, a central position p_k for alignment will be found. The formula is as follows.

$$p_k = S \cdot \mathrm{sigmoid}(v_p^T \tanh(W_p h_k)) \tag{1}$$

Where S denotes the whole length of the image caption. v_p^T and W_p are parameters of the model, which are obtained from training. Since we use parallel computation, latent variable h_k uses e_k, the embedding vector of decoder's current input word.

Then a fixed-size window is utilized to assign the input of decoder. In case we may choose the words after the current time step, a mask layer is needed. To ease gradient propagation, gated linear unites (GLU) is used after each convolutional connection. According to Dauphin et al.[24]'s results, GLU also improves the performance and speeds up the computation compared to Oord et al.[27] in the context of language modelling. And residual forward connections are used between sub convolutional blocks. In this way, for the prediction of each output word, the length of input is fixed.

B. Masked Positional Self-attention

Images convey enormous information. How to pick out information of interest effectively and how to predict more precisely at each time step still remain challenging. Conventional encoder-decoder structure only uses a fixed-length multidimensional feature vector c to transmit encoded message $(s_1, s_2, \cdots, s_n)$. And then this intermediate vector is made use of to create hidden variables $(h_1, h_2, \cdots, h_m)$ in decoder. Finally, we get output from these variables.

However, an apparent drawback exists in such pattern: the model suffers from information loss, and the situation gets worse as the sequence gets longer. To solve this problem, attention mechanism was developed. The model uses a sequence of vectors $(c_1, c_2, \cdots, c_m)$ to take place of c. For LSTM-LSTM model, for example, the hidden variable of the decoder h_i at time i is computed based on the previous hidden variable h_{i-1}, the previous output y_{i-1} and c_i. and vector c_i is the weighting summation of the hidden variables of the encoder.

$$c_i = \sum_j a_{ij} s_j \tag{2}$$

$$a_{ij} = \mathrm{softmax}(\mathrm{f_a}(h_{i-1}, s_j)) \tag{3}$$

Where $\mathrm{f_a}()$ denotes the attention model.

Attention mechanism, in fact can be interpreted as an alignment algorithm. Therefore, the main task is to establish an appropriate alignment model. Given the request, attention mechanism gives the expectation from a set of key-value pairs. In most cases, the output is the weighted sum. And the weights are computed from the requests and keys. Attention models differ in the following two ways: the first one is how the weights and values combine, which leads to the classification of additive attention and multiplicative attention. The second part is the approach to get the weights, which gives rise to soft attention and hard attention. What worth mentioning is that the latent states change with the process of description generation. That is to say, the network needs to pay attention to what has been generated.

Our model uses two attention models- soft attention to align image features and texts, and positional self-attention to create more accurate captions. Firstly, soft attention lets the model focus on various part of the image at different time steps. The information loss decreases with the help of these multidimensional intermediate variables $(c_1, c_2, \cdots, c_m)$.

Besides, we developed masked positional self-attention mechanism. As shown in Fig. 2, for each sentence, we add two special tags #Start# and #End# at the beginning and end, separately. The sequence of predicted words $(w_0, w_1, \cdots, w_{k-1})$ is embedded to sequence $(e_0, e_1, \cdots, e_{k-1})$ to form the input of the decoder. We then uses self-attention mechanism to locate the position p_k of the most related word w_{p_k} when predicting word w_k. With p_k as the central position, we choose a window $[p_k - D, p_k + D]$ as the input of multi-layer convolutional network. The probabilistic formula of the output is as follows.

$$p(y_k | y_1, \cdots, y_{k-1}, I) = \mathrm{softmax}(W_o h_k^L + b_o) \tag{4}$$

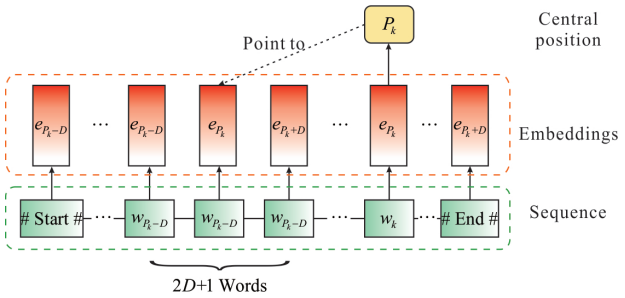

Fig. 2. Masked positonal self-attention

Where L is the total number of convolutional blacks in decoder, W_o and b_o are the parameters of convolutional network.

IV. EXPERIMENTS

A. Datasets

We did experiments both on Flickr8k and MSCOCO Caption datasets. Flickr8k Dataset is mainly for validation at early stages. Out of MSCOCO Caption dataset[29] (2017 version), 113287 images are selected randomly for training and 5000 for testing. For each image, at least 5 captions are provided. And the total number of captions is over 1.5 million.

B. Precision

As for search algorithm, normal max search and beam search are used. And for better comparison, we chose $k = 3, 5, 7$, three different situations of beam search. The BLEU scores are computed against all the reference sentences (at least five for each image). According to our result, for most images, our model gives descent descriptions with no or slight grammatical mistakes. Fig. 3 gives two examples of our model's predictions for images from two datasets. Fig. 3 (a) is from Flickr8k and Fig. 3 (b) is from MSCOCO Caption 2017. Just like the returned sentences shown in Fig. 3, all the search algorithms give correct and proper captions, although may differ a little bit in some details.

We verify the precision of our hierarchical model quantitatively as well. Table 1 lists BLEU values (BLEU-1 to BLEU-4) of baselines and our model. LSTM baseline refers to Karpathy's model[28], which uses convolutional neural networks and bidirectional recurrent neural networks to capture inter-modal correspondences between visual data and language. All the evaluation values are computed on MSCOCO Caption dataset. Our hierarchical model outperforms LSTM baseline in all four BLEU scores. And the BLEU-1 scores of beam search ($k = 3$ and $k = 5$) are over 0.7, reaching close to state-of-the-art of CNN-LSTM architecture image caption models but with less training time.

Additionally, according to Table 1, none of these search algorithms has made far ahead on BLEU values compared

Normal Max search: A skateboarder in the air with his skateboard
Beam Search, k=3: A skateboarder in the middle of a trick jumping.
Beam Search, k=5: A skateboarder doing a trick on a ramp.
Beam Search, k=7: A skateboarder doing a trick on a ramp.

(a) An example from Flickr8k

Normal Max search: A group of elephants walking across a dirt field.
Beam Search, k=3: A group of elephants standing next to each other.
Beam Search, k=5: A couple of elephants standing next to each other.
Beam Search, k=7: A couple of elephants standing next to each other.

(b) An example from MSCOCO

Fig. 3. Two examples of hierarchical image caption model

to each other. Some may get higher BLEU-1 score but not as good as other search algorithms in BLEU-2 or BLEU-3. So we keep all of them here. To be more concise, from the predictions of Fig. 3 (b), we could find more details about the background using normal max search. But in Fig. 3(a), beam search tells detailed actions- the skateboarder is in the middle of a trick. We don't go deep here to keep the paper reasonably concise.

C. Further discussion

Nevertheless, some attempts failed. Fig. 4 gives a collection of different failing cases. Major wrong descriptive words are marked in red. We categorize them into the following six types.

1) Occlusions: Some images contain partly sheltered objects. In this example (Fig. 4 (a)), a woman is standing on the street with her back to the camera. Besides, she was half sheltered in the shadow. The model recognizes her as a young boy. And the action is also mistaken.

2) Out-of-focus blur: Incorrectly focused background usually comes with fast moving objects, such as sports scenes and portrait. Neural networks may detect the blurry color patches as other similar objects, which are far away from the topic.

3) Images with multiple objects: Images with multiple objects are significantly hard to focus. In Fig. 4 (c), prediction 1 and prediction 2 are both correct descriptions. The primary deficiency lies in the lack of globality. Better captions would be fundamental abstracts, since only one sentence is allowed for each input image.

TABLE I

COMPARISON OF BLEU SCORES ON MSCOCO CAPTION DATASET

Model		BLEU-1	BLEU-2	BLEU-3	BLEU-4
Baselines	LSTM baseline[a]	0.625	0.450	0.321	0.230
	NIC[b]	-	-	-	0.277
Our hierarchical model	Normal Max Search	0.697	0.517	0.362	0.251
	Beam Search (k = 3)	0.703	0.501	0.369	0.254
	Beam Search (k = 5)	0.705	0.513	0.378	0.248
	Beam Search (k = 7)	0.692	0.518	0.371	0.266

[a]The image caption model presented by Karpathy and et al.[28]

[b]Neural image caption model presented by Oriol Vinyals and et al.[3]

Prediction 1: A little boy in a blue shirt is standing on a slide.
Prediction 2: A young boy is playing with a soccer ball outside .

(a) Occlusions

Prediction 1: A man is riding his bicycle and is on a dirt path and red and People carrying yellow flowers.
Prediction 2: A man on a race car race down a grassy track .

(b) Out-of-focus blur

Prediction 1: A bunch of bananas sitting on a table.
Prediction 2: A pile of oranges sitting on top of a wooden table.

(c) Images with multiple objects

Prediction 1: A group of people sitting around a table in a room.
Prediction 2: A group of people playing a video game in a living room.

(d) Too much attention to partial information

Prediction 1: A dog is jumping over a swing at a park.
Prediction 2: A dog is jumping over a swing at a park.

(e) Out of vocabulary

Prediction 1: A bowl of fruit is sitting on a table.
Prediction 2: A pile of oranges sitting on top of a wooden table.

(f) Synecdoche

Fig. 4. Examples of hierarchical image caption model

4) Too much attention to certain partial information: Obviously, prediction 1 is more rational than prediction 2 for Fig. 4 (d). The latter one has paid too much attention to the screens in the top right corner of the photo.

5) Out of vocabulary: Some objects could not be described properly. To be more precisely, the model describes the obstacle as a swing. It happens when the object has never appeared before in the training dataset. When more diverse images are put into the experiment and the vocabulary is enriched, such situation can be greatly improved.

6) Synecdoche: If there are a number of different kinds of fruits in the picture, we may not depict each type clearly. Instead, we only use the word FRUIT to represent all the objects.

V. CONCLUSION

Based on encoder-decoder architecture, our model unifies the neural networks used in image processing and natural language generation part. It reaches a BLEU-1 score over 0.7 with a faster training speed compared with CNN-LSTM architecture models. Our model could give inspiration to further study in multi-modal machine learning. Admittedly, there are still several issues need to be addressed to improve the performance.

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2017.

[2] J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," *Computer Science*, 2014.

[3] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," *2015 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)*, pp. 3156–3164, 2015. [Online]. Available: ¡Go to ISI¿://WOS:000387959203020

[4] Y. F. Li, J. T. Kwok, I. W. Tsang, and Z. H. Zhou, *A Convex Method for Locating Regions of Interest with Multi-instance Learning*. Springer Berlin Heidelberg, 2009.

[5] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, "Babytalk: understanding and generating simple image descriptions," *IEEE Trans Pattern Anal Mach Intell*, vol. 35, no. 12, pp. 2891–903, 2013. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/22848128

[6] Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos, "Corpus-guided sentence generation of natural images," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, Conference Proceedings, pp. 444–454.

[7] V. Ordonez, G. Kulkarni, and T. L. Berg, "Im2text: describing images using 1 million captioned photographs," in *International Conference on Neural Information Processing Systems*, 2011, Conference Proceedings, pp. 1143–1151.

[8] A. Farhadi, M. Hejrati, M. A. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth, "Every picture tells a story: generating sentences from images," in *European Conference on Computer Vision*, 2010, Conference Proceedings, pp. 15–29.

[9] S. Li, G. Kulkarni, T. L. Berg, A. C. Berg, and Y. Choi, "Composing simple image descriptions using web-scale n-grams," in *Proceedings of the Fifteenth Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 2011, Conference Proceedings, pp. 220–228.

[10] P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi, "Collective generation of natural image descriptions," in *Meeting of the Association for Computational Linguistics: Long Papers*, 2012.

[11] A. Aker and R. J. Gaizauskas, "Generating image descriptions using dependency relational patterns," in *Meeting of the Association for Computational Linguistics*, 2010.

[12] Y. Feng and M. Lapata, "How many words is a picture worth? automatic caption generation for news images," in *Acl, Meeting of the Association for Computational Linguistics, July, Uppsala, Sweden*, 2011.

[13] R. Kiros, R. Salakhutdinov, and R. Zemel, "Multimodal neural language models," in *International Conference on International Conference on Machine Learning*, 2014.

[14] R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *Computer Science*, 2014.

[15] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[16] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional lstms," in *ACM on Multimedia Conference*, 2016, Conference Proceedings, pp. 988–997.

[17] Y. H. Tan and C. S. Chan, "phi-lstm: A phrase-based hierarchical lstm model for image captioning," in *Asian Conference on Computer Vision*, 2016.

[18] K. Cho, B. V. Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *Computer Science*, 2014.

[19] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 3104–3112. [Online]. Available: http://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf

[20] K. Cho, B. V. Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," *Computer Science*, 2014.

[21] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *Computer Science*, 2014.

[22] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei, "Boosting image captioning with attributes," in *IEEE International Conference on Computer Vision*, 2016, Conference Proceedings, pp. 4904–4912.

[23] W. Jiang, L. Ma, X. Chen, H. Zhang, and W. Liu, "Learning to guide decoding for image captioning," *arXiv preprint arXiv:1804.00887*, 2018.

[24] Y. N. Dauphin, A. Fan, M. Auli, and D. Grangier, "Language modeling with gated convolutional networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 933–941.

[25] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1243–1252.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[27] A. Van den Oord, N. Kalchbrenner, L. Espeholt, O. Vinyals, A. Graves *et al.*, "Conditional image generation with pixelcnn decoders," in *Advances in neural information processing systems*, 2016, pp. 4790–4798.

[28] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *IEEE Trans Pattern Anal Mach Intell*, vol. 39, no. 4, pp. 664–676, 2017. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/27514036

[29] X. Chen, H. Fang, T. Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft coco captions: Data collection and evaluation server," *Computer Science*, 2015.