TEXT EMBEDDING BANK MODULE FOR DETAILED IMAGE PARAGRAPH CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image paragraph captioning is the task of automatically generating multiple sentences for describing images in grain-fined and coherent text. Existing typical deep learning-based models for image captioning consist of an image encoder to extract visual features and a language model decoder, which has shown promising results in single high-level sentence generation. However, only the word-level scalar guiding signal is available when the image encoder is optimized to extract visual features. The inconsistency between the parallel extraction of visual features and sequential text supervision limits its success when the length of the generated text is long (more than 50 words). In this paper, we propose a new module, called the Text Embedding Bank (TEB) module, to address the problem for image paragraph captioning. This module uses the paragraph vector model to learn fixed-length feature representations from a variable-length paragraph. We refer to the fixed-length feature as the TEB. This TEB module plays two roles to benefit paragraph captioning performance. First, it acts as a form of global and coherent deep supervision to regularize visual feature extraction in the image encoder. Second, it acts as a distributed memory to provide features of the whole paragraph to the language model, which alleviating the long-term dependency problem. Adding this module to two existing state-of-the-art methods achieves a new state-of-the-art result by a large margin on the paragraph captioning Visual Genome dataset.

1 INTRODUCTION

Automatically generating a natural language description for visual content like image or video is an emerging interdisciplinary task. This task involves computer vision, natural language processing and artificial intelligence. Thanks to the advent of large datasets Lin et al. (2014); Young et al. (2014); Krishna et al. (2017b), many recent works Mao et al. (2014); You et al. (2016) have shown promising results in generating a single high-level scene for images and videos. However, the coarse, scene-level descriptions that these models produce cannot meet real-world applications such as video retrieval, automatic medical report generation Greenspan et al. (2016); Wang et al. (2017; 2018); Li et al. (2018a), blind navigation and automatic video subtitling which capture fine-grained entities and have a coherent and logically detailed description.

To tackle this challenge, a relatively new task called paragraph captioning is emerging. Paragraph captioning is the task of generating coherent and logically detailed descriptions by capturing the fine-grained entities of the image or video. A few works Krause et al. (2017); Liang et al. (2017); Melas-Kyriazi et al. (2018) have pushed the performance to new heights with the main paragraph captioning dataset, the Visual Genome corpus, a dataset introduced by Krause et al. (2017).

Compared with the performance of single-sentence caption generating models, the performance paragraph-length caption generating models is lower by a large margin. Paragraph captioning for images and videos is challenging due to the requirement of both fine-grained image understanding and long-term language reasoning. To overcome these challenges, we propose the TEB module, a module that is easy to integrate with existing image captioning models. This module maps varied-length paragraphs to a fixed-length vector which we call TEB. Each unique vector in the TEB has distance meaning and indexed by the order of the word in the vocabulary. The TEB has a distributed memory. This is illustrated in detail in section 3. Existing deep learning based models typically consist of an image encoder to extract visual features in parallel with a RNN language model decoder

to generate the sentences word by word sequentially. In the training stage, only a tiny partial scalar guiding information from the word level loss is available to optimize the image encoding training. This results in an insufficient fine-grained and coherent image visual feature extraction. The TEB module, which holds the whole paragraph in a distributed memory model, can provide global supervision to better regularize the image encoder in the training stage. The RNNs are known to have a long-term dependency problem because of vanishing and exploding gradients which make it unable to meet long-term language reasoning. Since the TEB module has distributed memory and can provide ordering, it is better with long-term language reasoning.

We integrated our TEB module with the state-of-the-art methods on the only available paragraph captioning dataset, the Visual Genome corpus, and achieved new state-of-the-art by a large margin.

2 Related Works

2.1 Standard Caption or Dense Caption

This image to text problem is a classic problem in computer vision and NLP. The first work to use deep neural networks to solve this problem was the Neural Image Caption (NIC) in Vinyals et al. (2015), which uses a pre-trained CNN as the visual model and a RNN as the language model. The visual model extracts visual features which are fed to the first time step of the RNN. The language model takes visual features from the visual model at the first time step and predicts the first word, before feeding the predicted word into the next time step and so on. At each time step, the difference between the predicted word and the ground truth word is optimized by softmax with cross entropy loss. This work can only predict one short simple sentence for each natural image. The performance of this one sentence caption task is improved in Xu et al. (2015) by introducing an attention mechanism which focuses on related regions when generating a word per time step in the RNN model. In order to give a description for every object in an image, DenseCapJohnson et al. (2016) proposed a fully convolutional localization network which upgraded the region proposal network from Faster R-CNNRen et al. (2015) to localize the salient regions. The RNN model then takes the corresponding visual features for each localized region to generate a sentence. However, simply joining all of the generated sentences together doesn't produce a coherent paragraph as there are semantic relationships between sentences, which is a shortcoming of DenseCap.

Similarly, dense video captioning, a task which gives each event a description in a video, was first explored in Krishna et al. (2017a) by a variant of the existing proposal module and using 3D features. Later it was further improved by jointly localizing and describing eventsLi et al. (2018b).

Recently, the RNN/LSTM language model was replaced by a CNN in Aneja et al. (2017); Wang & Chan (2018) with comparable performance and the potential for parallel computing, which is a drawback of sequential models. In the inference process, however, this CNN model also need to be computed sequentially. Since computation cost is a big issue for video captioning, Chen et al. (2018) introduced a new method to find the useful frames which cut redundant information and reduce computation cost.

2.2 Paragraph Captioning

Standard captioning generates single high-level sentence. Dense captioning generates a description for each salient object in an incoherent way. Paragraph captioning, however, overcomes the weaknesses of the previous two tasks by generating fine-grained and coherent natural language descriptions, like a story. To meet long-term language reasoning and the requirement of multiple topics in multiple sentences, a hierarchical recurrent neural network architectureLi et al. (2015); Lin et al. (2015); Yu et al. (2016); Krause et al. (2017); Jing et al. (2017) is widely used in paragraph captioning. For example, Yu et al. (2016) generate multiple sentences for video captioning by capturing strong temporal dependencies. Krause et al. (2017) uses a hierarchical recurrent network to build relationships between sentences. Regional features are passed to a sentence RNN to generate topic vectors with a halting distribution to control the ending of new topic generation. The generated topic vectors are then consumed by a word RNN to generate sentences. In this way, this hierarchical RNN and DenseCap offer two ways of generating new topics, which is essential for multiple sentence generation. The IU Chest X-ray dataset is used for automatic report generation on this unstructured reportJing et al. (2017) by using co-attention and the hierarchical LSTM. The Diversity model

Melas-Kyriazi et al. (2018) improves sentence diversity by introducing a repetitive penalty in the sequence-level training. However, all of these methods suffer from the fact that only a tiny partial scalar from the word level loss can be used as guiding information to optimize the image encoding in training. Our TEB module can overcome this and provides an alternative for the hierarchical recurrent neural network architecture. With our TEB module, only one level recurrent neural network is enough to generate multiple sentences with multiple find-grained topics.

2.3 LONG-TERM DEPENDENCY

GANs have proved to improve real text generation in Zhang et al. (2017). SeqGAN Yu et al. (2017) is proposed to deal with the sequential and discrete property of text for text generation. LeakGAN Guo et al. (2017) solves the sparse signal from generator problem by leaking feature from the generator to the discriminator for long sentence generation. MaskGan Fedus et al. (2018) introduced a way to fill in the blank with GAN. Similarly, Wu et al. (2019) use long-term feature banks for detailed video understanding.

3 APPROACH

The proposed TEB module improves paragraph captioning by describing the rich content of a given image. Figure 1 shows an example of how the TEB module can be integrated with an existing typical image captioning pipeline.

3.1 LEARNING VECTOR REPRESENTATION OF WORDS

The paragraph vector is based on word vectors. A word vector is the concept of using a distributed vector representation of words. The basic idea is to predict a word given the other words in a context. The framework is shown in Figure 2.

In this framework, each word is mapped to a unique vector which is a column of a matrix $\boldsymbol{W}$. The column is indexed by the order of the word in the vocabulary. The features to predict the next word are the sum or concatenation of the vectors.

To express this in a mathematical equation, let $\boldsymbol{w}_1, \boldsymbol{w}_2, \boldsymbol{w}_3, ..., \boldsymbol{w}_T$ represent the vectors of a sequence of training words. The objective function of the framework is to maximize the average log probability

$$\frac{1}{T}\sum_{t=k}^{T-k} \log p(\boldsymbol{w}_t \mid \boldsymbol{w}_{t-k}, ..., \boldsymbol{w}_{t+k}) \tag{1}$$

Typically, a multi-class classifier such as softmax is used for the prediction task. So, we have

$$p(\boldsymbol{w}_t \mid \boldsymbol{w}_{t-k}, ..., \boldsymbol{w}_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}} \tag{2}$$

where y_i is the un-normalized log-probability for each output word i, which is computed as

$$y = b + Uh(\boldsymbol{w}_{t-k}, ..., \boldsymbol{w}_{t+k}; \boldsymbol{W}) \tag{3}$$

This framework is implemented in a neural network and trained using stochastic gradient descent through back-propagation Rumelhart et al. (1988). This type of model is the well known neural language model Bengio et al..

Compared to existing image captioning models, which only using recurrent neural networks, after training converges, this framework can map words with similar meaning to a similar position in the vector space. For example, "wind" and "beautiful" are far away from each other in the vector space, while "beautiful" and "pretty" are closer. Additionally, the distance between each unique word vector also carries meaning. This means that it can be used for analogy questions answering in a simple vector algebra manipulation: "waiter" - "man" + "women" = "waitress". This makes it easy to learn a linear matrix, such as a fully connected layer, to translate between visual features and these word vectors.

3.2 PARAGRAPH VECTOR: A DISTRIBUTED MEMORY MODEL

Inspired by the word vector framework which can capture the semantics as a result of a prediction task, The paragraph vector also contributes the prediction of the next word. In this paragraph vector framework (See Figure 3), similarly to the word vector framework, each word is still mapped to a unique vector which is a column of a matrix $\boldsymbol{W}$, while each paragraph is mapped to a unique vector which is a column of a matrix $\boldsymbol{D}$. Then both the word vector and paragraph vector are fused (either sum or concatenated) as features to predict the next word. We use concatenation in our implementation.

The paragraph vector can be treated as a super word (or the topic of the paragraph) which acts as memory of the missing information from the current context. Hence, this framework is known as a distributed memory model. This property can compensate the recurrent neural network for its lack of generating logical connections between sentences or paragraphs.

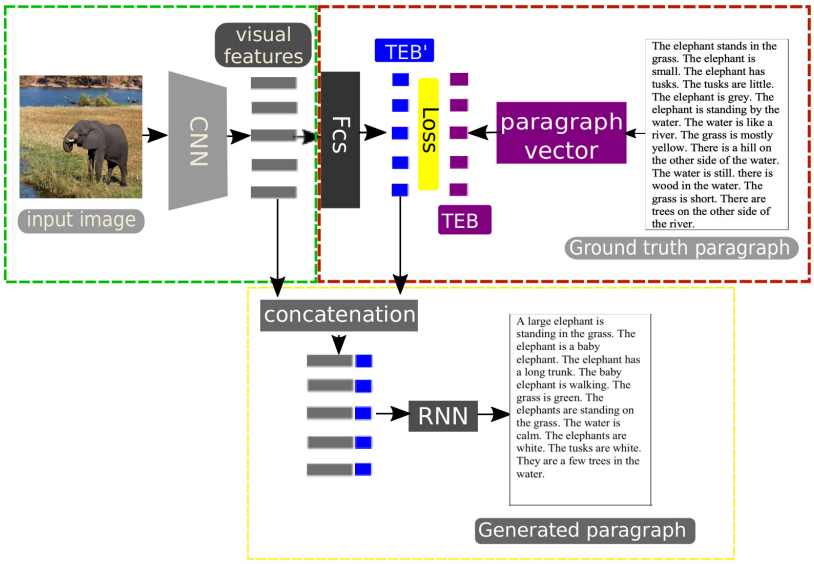

Figure 1: Integration of the paragraph vector framework as a TEB module to an existing deep learning based image captioning model. There are three interconnected components divided into three dashed rectangular boxes. In the green box on the top left, the image encoder extracts visual features through a CNN model. In the yellow box on the bottom, a RNN based language model decoder is used to generate paragraphs. Existing deep learning based models only contain these two components. The red box on the top right box is the TEB module: In the training stage, for an image, paragraph pair, the varied-length paragraph is mapped to a fixed-length vector which is called TEB through the paragraph vector framework. The visual features from the image encoder are converted to the predicted TEB (called TEB') through several fully connected layers. The TEB' is supervised by the TEB through an L1 loss, which acts as global deep supervision to regularize the visual feature extraction for the image encoder. The visual features and TEB' are concatenated and feed into the RNN as input. The generated paragraph is supervised by the ground truth paragraph through a word-level loss. In the inference stage, the TEB is not available and the TEB' acts as the TEB to provide the features of the whole paragraph to alleviate the long-term dependency problem for the language model.

3.3 INTEGRATION OF THE PARAGRAPH VECTOR AS A TEB MODULE FOR IMAGE PARAGRAPH CAPTIONING

The integration of the paragraph vector as a TEB module for image paragraph captioning is illustrated in Figure 1.

Table 1: Our result compared with prior results on Visual Genome dataset
Notations: Models 1-4 for Krause et al. (2017) are Template, Flat w/o object detector,Flat and Hierarchical individually.
Models 1-2 for Liang et al. (2017) are w/o discriminator and with discriminator individually.
Models 1-4 for Melas-Kyriazi et al. (2018) are XE training, w/o rep. penalty, XE training, w/ rep. penalty, SCST training, w/o rep. penalty and SCST training, w/ rep. penalty individually.

Methods	METEOR	CIDEr	BLEU-1	BLEU-2	BLUE-3	BLEU-4
Krause et al. (2017)[1]	14.31	12.15	37.47	21.02	12.30	7.38
Krause et al. (2017)[2]	12.82	11.06	34.04	19.95	12.20	7.71
Krause et al. (2017)[3]	13.54	11.14	37.30	21.70	13.07	8.07
Krause et al. (2017)[4]	15.95	13.52	41.90	24.11	14.23	8.69
Liang et al. (2017)[1]	16.57	15.07	41.86	24.33	14.56	8.99
Liang et al. (2017)[2]	17.12	16.87	41.99	24.86	14.89	9.03
Melas-Kyriazi et al. (2018)[1]	13.66	12.89	32.78	19.00	11.40	6.89
Melas-Kyriazi et al. (2018)[2]	15.17	22.68	35.68	22.40	14.04	8.70
Melas-Kyriazi et al. (2018)[3]	13.63	13.77	29.67	16.45	9.74	5.88
Melas-Kyriazi et al. (2018)[4]	17.86	30.63	43.54	27.44	17.33	10.58
Ours (Transformer)	15.45	23.38	41.49	23.38	11.96	6.00
Ours (Transformer + TEB)	15.88	24.84	41.86	24.64	13.97	6.40
Ours (Diversity + TEB)	**18.36**	**32.53**	**45.24**	**28.44**	**17.93**	**10.98**

4 IMPLEMENTATION

4.1 TEB MODULE

For the paragraph vector frameworkLe & Mikolov (2014), we adapted the implementation of Lau & Baldwin (2016). The hyperparameters are as follows: The vector size (TEB size) is 512, the sliding window size is 50, the sampling threshold is $1e-5$, the negative size is 5. The paragraph vector model is trained for 1000 epochs before performing the inference to generate the TEB. Regardless of the dimension size of the visual features from the image encoder, the visual features are converted to the same dimension of the TEB by several fully connected layers. In the concatenation of the TEB' and visual features, a weight of 0.1 is applied to the TEB'.

4.2 INTEGRATING TEB ON DIVERSITY MODEL MELAS-KYRIAZI ET AL. (2018)

We integrate our TEB module with the Diversity model Melas-Kyriazi et al. (2018) which is the current state of art model on the Visual Genome dataset. We used the model architecture and the entire training procedure from the Diversity model Melas-Kyriazi et al. (2018) except the TEB module for a fair comparison. This model uses the Bottom-Up and Top-Down model Anderson et al. (2018) as its backbone. Self-critical sequence training (SCST) and repetitive training are also used.

4.3 INTEGRATING TEB ON TRANSFORMER MODEL

We also integrate the TEB module with a transformer model. The transformer model is adapted from the Bottom-Up and Top-Down model Anderson et al. (2018) with the following modification: The LSTM-based language model is replaced by the transformer modelVaswani et al. (2017). We used both cross-entropy and SCST training, without the repetition penalty, and beam search instead o greedy search.

5 RESULTS

Table 1 shows the quantitative results. We have three models. The "Diversity + TEB" model is the Diversity model Melas-Kyriazi et al. (2018) with SCST trainingRennie et al. (2017), repetition

penalty and TEB module. The "Transformer" model is Replacing the LSTM model with Transformer Vaswani et al. (2017) in the Bottom-UP and Top-Down modelAnderson et al. (2018). The "Transformer + TEB" is the "Transformer" model with TEB module. The TEB module improve both baseline model by a large margin and our model "Diversity + TEB' achieve state of the art result on the visual genome result.

Figure 2 shows the qualitative results between DiversityMelas-Kyriazi et al. (2018) and our DiversityMelas-Kyriazi et al. (2018) with TEB module. Our result generates richer topics and more detailed information.

Input Image	Ours (Diversity + TEB)	Diversity	Ground truth
	A bunch of boats are sitting on a beach. The water is calm and blue. There are a lot of boats in the water. The boats are white. There is a yellow umbrella on the boat. The sky is blue. The clouds are white and white.	A bunch of boats are on a pier. There are a large white boat in the water. There is a large blue and white boat on the water.	This is an image of a harbor. The harbor has many small boats in it. The water is blue. The water is reflecting the sky. The sky is partly cloudy. The clouds are white and fluffy. The sky is light blue. There are white buoys on the dock with small cloth sails in them. The sails are light brown and white.
	A man is standing in a of a man. The man is wearing a white shirt. The man has a black shirt on. The man is holding a hot dog. The sandwich is wearing a black. The men are wearing a blue shirt. There are people standing behind the man. There is a man in a blue shirt standing in the background. There are trees behind the man.	A man is standing in a white basket. He is wearing a black shirt and a black hat. The man is holding a hot dog in his hand. There is a man in a black shirt standing behind the man.	There is a man. The man is wearing a yellow shirt. The man is standing at a park. There are more people in the park. There are people sitting under the threes. There are people walking in the paths. The man is holding a sandwich. The sandwich is a hot dog. The sandwich has a sausage. The sandwich has onions.
	A large elephant is standing in the grass. The elephant is a baby elephant. The elephant has a long trunk. The baby elephant is walking. The grass is green. The elephants are standing on the grass. The water is calm. The elephants are white. The tusks are white. They are a few trees in the water.	A large elephant is standing in the water. The elephant is walking in the water. There is a large body of water behind the elephant. There are a small rock behind the elephant.	The elephant stands in the grass. The elephant is small. The elephant has tusks. The tusks are little. The elephant is grey. The elephant is standing by the water. The water is like a river. The grass is mostly yellow. There is a hill on the other side of the water. The water is still. there is wood in the water. The grass is short. There are trees on the other side of the river.
	A white toilet is in the bathroom. The toilet is white. The lid is white. There is a white toilet in the toilet. The toilet lid is up. The floor is made of white. The tiles are white. There is a white wall behind the toilet.	A white toilet is sitting on the ground. There is a white toilet in the toilet. There is a toilet in front of the toilet.	The toilet lid is up. The toilet bowl is cleaning. The toilet is a very light beige color. There's a white bar between the toilet lid and the toilet seat. The toilet is encased in a cubby space. The water in the toilet is low. The floor around the toilet is made of tiles. There are wires on the bottom left side of the toilet bowl.

Figure 2: Qualitative result comparison of paragraph outputs of our model (Diversity with TEB) and the baseline Diversity model Melas-Kyriazi et al. (2018)

6 Conclusion

In this paper, we propose the Text Embedding Bank (TEB) module for visual paragraph captioning, a task which requires capturing fine-grained entities in the image to generate a detailed and coherent paragraph, like a story. Our TEB module provides global and parallel deep supervision and distributed memory for find-grained image understanding and long-term language reasoning. Integrating the TEB module to existing state-of-the-art methods achieves new state-of-the-art results by a large margin.

REFERENCES

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, 2018.

Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. Convolutional image captioning. *arXiv preprint arXiv:1711.09151*, 2017.

Yoshua Bengio, Holger Schwenk, Jean-Sébastien Senécal, and Fréderic Morin. Gauvain, jean-luc. neural probabilistic language models. *Innovations in Machine Learning*, pp. 137–186.

Yangyu Chen, Shuhui Wang, Weigang Zhang, and Qingming Huang. Less is more: Picking informative frames for video captioning. *arXiv preprint arXiv:1803.01457*, 2018.

William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: Better text generation via filling in the _. *arXiv preprint arXiv:1801.07736*, 2018.

Hayit Greenspan, Bram Van Ginneken, and Ronald M Summers. Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique. *IEEE Transactions on Medical Imaging*, 35(5):1153–1159, 2016.

Jiaxian Guo, Sidi Lu, Han Cai, Weinan Zhang, Yong Yu, and Jun Wang. Long text generation via adversarial training with leaked information. *arXiv preprint arXiv:1709.08624*, 2017.

Baoyu Jing, Pengtao Xie, and Eric Xing. On the automatic generation of medical imaging reports. *arXiv preprint arXiv:1711.08195*, 2017.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4565–4574, 2016.

Jonathan Krause, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. A hierarchical approach for generating descriptive image paragraphs. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 3337–3345. IEEE, 2017.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and J Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, pp. 6, 2017a.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017b.

Jey Han Lau and Timothy Baldwin. An empirical evaluation of doc2vec with practical insights into document embedding generation. *arXiv preprint arXiv:1607.05368*, 2016.

Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pp. 1188–1196, 2014.

Christy Y Li, Xiaodan Liang, Zhiting Hu, and Eric P Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. *arXiv preprint arXiv:1805.08298*, 2018a.

Jiwei Li, Minh-Thang Luong, and Dan Jurafsky. A hierarchical neural autoencoder for paragraphs and documents. *arXiv preprint arXiv:1506.01057*, 2015.

Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7492–7500, 2018b.

Xiaodan Liang, Zhiting Hu, Hao Zhang, Chuang Gan, and Eric P Xing. Recurrent topic-transition gan for visual paragraph generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3362–3371, 2017.

Rui Lin, Shujie Liu, Muyun Yang, Mu Li, Ming Zhou, and Sheng Li. Hierarchical recurrent neural network for document modeling. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 899–907, 2015.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

Luke Melas-Kyriazi, Alexander Rush, and George Han. Training for diversity in image paragraph captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 757–761, 2018.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7008–7024, 2017.

David E Rumelhart, Geoffrey E Hinton, Ronald J Williams, et al. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Qingzhong Wang and Antoni B Chan. Cnn+ cnn: Convolutional decoders for image captioning. *arXiv preprint arXiv:1805.09019*, 2018.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 3462–3471. IEEE, 2017.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, and Ronald M Summers. Tienet: Text-image embedding network for common thorax disease classification and reporting in chest x-rays. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9049–9058, 2018.

Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 284–293, 2019.

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

Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4651–4659, 2016.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4584–4593, 2016.

Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, pp. 2852–2858, 2017.

Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. *arXiv preprint arXiv:1706.03850*, 2017.