Journal of Machine Learning Research 23 (2023) 1-10 Submitted 4/17; Revised -/-; Published -/-

Assessment of Automated Image Captioning Models

Advait Shah AR4SHAH@UWATERLOO.CA
Department of Management Sciences
University of Waterloo
Waterloo, ON N2L 3G1, Canada

Editor: Pascal Poupart

Abstract

Automatic image captioning, which involves describing the contents of an image, is a challenging problem with many applications in various research fields. Recently, there have been significant advances in image captioning methods owing to the breakthroughs in deep learning. This project aims to build a deep learning image captioning model based on attention mechanism described by Xu et al. (2015). We have used Flickr8k dataset for training and measured its performance on the test dataset with a BLEU score. Then we also compared and contrasted our model with other popular deep learning models on image captioning such as Unified VLP by Zhou et al. (2020) and sGPN by Zhong et al. (2020).

Keywords: Image Captioning, Multimodal Deep Learning, Text Generation, Neural Networks, Attention

1 Introduction

Automatically generating captions for an image is a task close to the heart of scene understanding — one of the primary goals of computer vision. Not only must caption generation models be able to solve the computer vision challenges of determining what objects are in an image, but they must also be powerful enough to capture and express their relationships in natural language. For this reason, caption generation has long been seen as a difficult problem. As shown by Ghandi et al. (2022), applications of automatic image captioning include human-computer interaction, medical image captioning and automatic medical prescription, quality control in industry, traffic data analysis, and especially assistive technologies for visually impaired individuals. Given the many challenges and obstacles of a visually impaired lifestyle, finding a means to ease these problems can be valuable, and may improve the life quality of visually impaired individuals.

Over the recent years, there have been significant advances in image captioning methods owing to the breakthroughs in deep learning. But still, as discussed in Ghandi et al. (2022), there are many opportunities to further improve these models to make them more useful. Through this project, we will build an image captioning model with one of the multimodal deep learning techniques, as described in the 'Show, attend and tell' paper by Xu et al. (2015) which consists of convolutional and recurrent neural networks with visual attention mechanism. To showcase the unique visual attention mechanism of this paper, we will show through visualization how the model is able to automatically learn to fix its gaze on salient objects while generating the corresponding words in the output sequence. We will train the

model with Flickr8k dataset and evaluate the model's performance with its BLEU score on the test dataset. We will then also compare our model performance with benchmarks from other state of the art image captioning models mentioned in the recent research papers.

2 Techniques to tackle the problem

Image Captioning is the task of describing the content of an image in words. This task lies at the intersection of computer vision and natural language processing. Most image captioning systems use a CNN based encoder coupled with RNN based decoder framework, where an input image is encoded into an intermediate representation of the information in the image, and then decoded into a descriptive text sequence. The most popular benchmarks are nocaps and COCO, and models are typically evaluated according to a BLEU or CIDER metric.

Despite significant results, these methods usually give general and vague captions for images and do not describe image contents appropriately since all information is compressed into a single vector. This causes problems with learning the information at the beginning of the sequence and the deeper relations between image contents. Many new methods have been proposed to solve these problems, most of them having the encoder-decoder structure as their core component. In addition to these methods, other methods, such as dense captioning by Johnson et al. (2016), have been proposed to solve the image captioning problem. RNN with visual attention mechanism-based decoder model described in Xu et al. (2015) is computationally expensive but can also generate captions with high accuracy.

Through this project, we are implementing the attention mechanism based image captioning model presented by Xu et al. (2015). The selection of this model for our image captioning task would enable us to evaluate empirically how attention mechanism helps in improving image captioning task performance when employed with image captioning model built upon deep CNN model as encoder and LSTM based RNN model as decoder. Figure 1 gives the brief idea behind how attention mechanism is employed in the RNN decoder phase. Due to computational resources related constraints, we will use Flickr8K dataset instead of large dataset such as MS COCO to train our model, and we will go ahead with the train- validation- test split as 6000-1000-1000 images known as Karpathy split from Karpathy and Fei-Fei (2015) through the Flickr8k captioned dataset json file. And as we have 5 captions per image, we have 40000 captions in total for this dataset. This would allow us to compare our model performance with that of the state-of-the-art models trained and tested on the same dataset.

By examining models that came before attention-based models, we notice that they first learn detectors for various visual concepts utilizing a framework called multi-instance learning. Then, they employ a language model trained on captions to the detector outputs, followed by rescoring from a joint imagetext embedding space. In contrast to these models, attention framework does not utilize object detectors directly but instead learns latent alignments from scratch. This capability permits the model to surpass the limits of "objectness" and learn to attend to abstract concepts.

Now, we will see in detail how this attention-based model as shown in Figure 1 works. As described in Xu et al. (2015), firstly, in the encoder part, we use a convolutional neural network in order to extract a set of feature vectors which we refer to as annotation vectors.

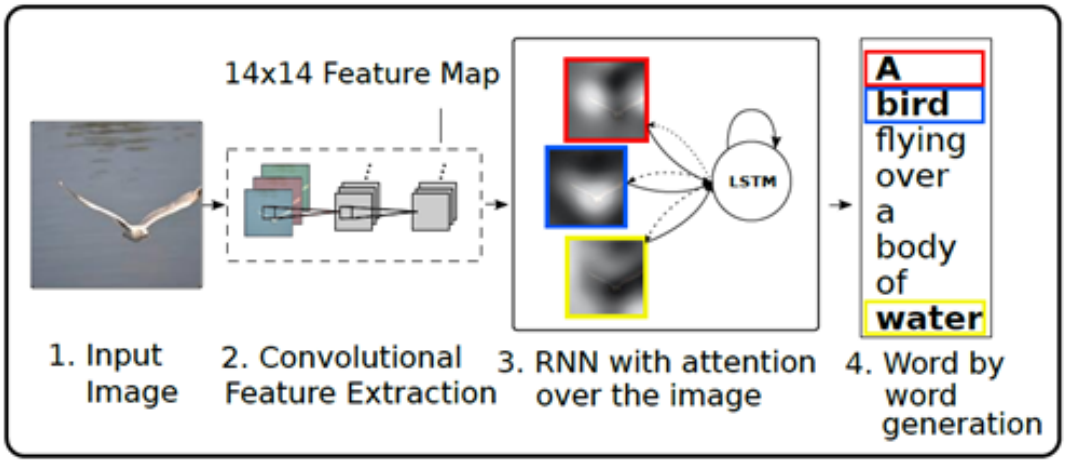

Figure 1: Working of Image captioning model with visual attention based RNN decoder (best viewed in colour), From Xu et al. (2015)

The extractor produces L vectors, each of which is a D-dimensional representation corresponding to a part of the image. In order to obtain a correspondence between the feature vectors and portions of the 2-D image, we extract features from a lower convolutional layer unlike previous work which instead used a fully connected layer. This allows the decoder to selectively focus on certain parts of an image by weighting a subset of all the feature vectors. In the decoder part, we use a long short-term memory (LSTM) network from Hochreiter and Schmidhuber (1997) that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words.

Paper from Xu et al. (2015) suggests two ways to implement attention mechanism: 1. Deterministic "Soft" attention and 2. Stochastic "Hard" Attention. "Soft" attention assigns a weight to each element in the input sequence, based on its relevance to the output. These weights are computed using a softmax function and can be interpreted as a probability distribution over the input elements. The output is then a weighted sum of the input elements, where the weights are given by the softmax probabilities. Soft attention is deterministic because the weights are computed using a fixed formula based on the input and the current state of the network.

Stochastic "hard" attention, on the other hand, is a variant of the attention mechanism that randomly selects a single input element to attend to. Instead of computing a probability distribution over the input elements, hard attention samples one element from the input sequence based on a probability distribution. The output is then the selected element, rather than a weighted sum of all elements. Hard attention is stochastic because it involves a random selection process. Figure 2 below illustrates these mechanisms visually.

While soft attention is differentiable, which allows for efficient backpropagation during training, hard attention is not differentiable and requires the use of reinforcement learning

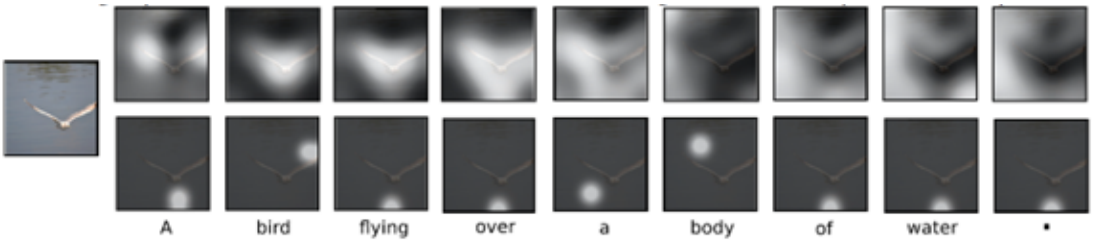

Figure 2: Visualization of (top) "soft" and (bottom) "hard" attention for each generated word. Here both models generated the same captions, From Xu et al. (2015)

techniques to learn the probability distribution. Hard attention can be useful when the input sequence is very long or when only a small number of elements in the sequence are relevant to the output. However, the stochasticity of hard attention can make it more difficult to train and can lead to instability in the network.

With the recent development in the transformer models for the Natural language generation, they are sometimes preferred over traditional RNN and LSTM techniques in the decoder phase of Image Captioning models. This is due to the fact that they can capture long-term dependencies in sequences and can generate high-quality captions due to unique positional encoding layer and multi-head attention layer. They are also less complex than RNN models and require less computational power. Liu et al. (2021a) introduce CaPtion TransformeR (CPTR), which takes sequentialized raw images as input to the Transformer. As an encoder-decoder framework, CPRT is a full Transformer network that replaces the commonly used CNN in the encoder part with the Transformer encoder. A purely Transformer-based architecture, PureT, is designed by Wang et al. (2022). In PureT, SwinTransformer from Liu et al. (2021b) replaces Faster-RCNN, and the architecture features a refining encoder and decoder. Furthermore, BraIN framework of Wang and Cook (2020) uses Bi-directional Generative Adversarial Network (GAN) mechanism to train image captioning model.

Recently, graph neural networks-based techniques are increasingly becoming popular for image captioning. One technique suggested by Zhong et al. (2020), as shown in Figure 3, takes a scene graph extracted from an input image, and decomposes the graph into a set of sub-graphs. It designs a sub-graph proposal network (sGPN) that learns to identify meaningful sub-graphs, which are further decoded by an attention-based LSTM for generating sentences and grounding sentence tokens into sub-graph nodes (image regions). By leveraging sub-graphs, this model enables accurate, diverse, grounded and controllable image captioning.

Also, in some recent works, large-scale model is pre-trained on a dataset with an enormous amount of data by self-supervised learning. The pre-trained model is then generalized to various downstream tasks. One widely used pre-trained model is CLIP (Contrastive Language-Image Pre-Training) from Mokady et al. (2021). CLIP is designed to provide a shared representation for both image and text prompts. Xia et al. (2021) present Cross-modal Generative Pre-Training for Image Captioning (XGPT), which uses a cross-modal encoder-decoder architecture and is directly optimized for generation tasks.

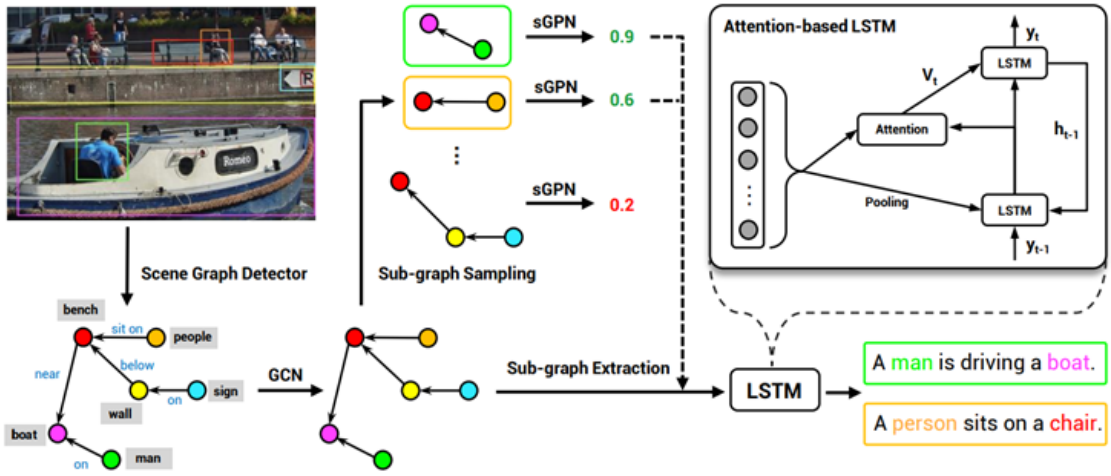

Figure 3: Working of Image captioning model with sub-graph based encoder (best viewed in colour), From Zhong et al. (2020)

In recent works on image captioning dating from 2018 to 2022, methods based on the attention mechanism and graphs have been used more frequently, as shown by Ghandi et al. (2022). Overall, the choice of model depends on the trade-off between complexity, performance, and ease of use for a particular application.

3 Empirical evaluation

To create the annotations used by our decoder, we used He et al. (2016) presented ResNet-101 pretrained on ImageNet without finetuning in our encoder phase. And we removed last 2 layers i.e. pooling and fully connected layers from it, as we need convolutions layer vectors as input to our attention mechanism.

Our soft attention model was trained with stochastic gradient descent using adaptive learning rates. For the Flickr8k dataset, Adam algorithm from Kingma and Ba (2014) is used for SGD. Other hyperparameters such as Learning rates were selected based on the paper source code and other code references from GitHub such as Xu (2015) and Rahman (2020).

Furthermore, we had used beam size of 7. Beam size, or beam width, is a parameter in the beam search algorithm which determines how many of the best partial solutions to evaluate. In an LSTM model, beam size limits the number of candidates to take as input for the decoder. A beam size of 1 is a best-first search - only the most probable candidate is chosen as input for the decoder. A beam size of k will decode and evaluate the top k candidates. A large beam size means a more extensive search - not only the single best candidate is evaluated.

On our Flickr8k dataset, our soft attention model took about 6 hours to train on an NVIDIA GeForce GTX 1650 GPU. Due to computing resources limitations, we trained it for 10 epochs, and the best BLEU-4 score was 12.36. Note that BLEU-4 uses the geometric

Figure 4: Train Loss vs. Epoch

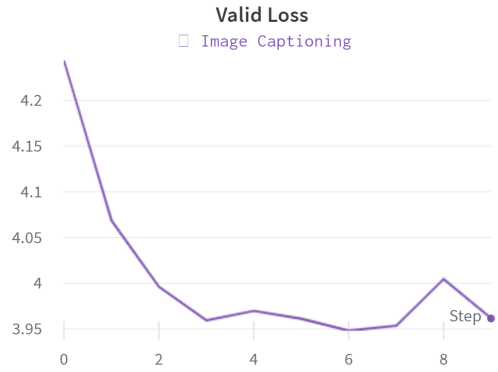

Figure 5: Validation Loss vs. Epoch

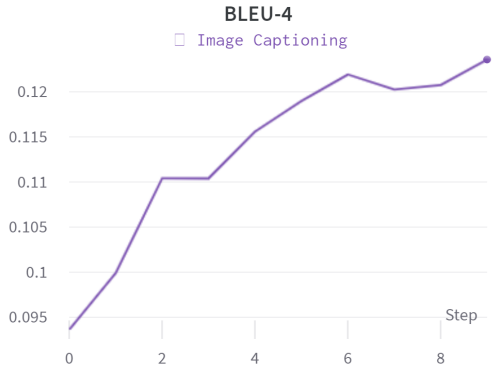

Figure 6: BLEU-4 Score vs. Epoch

Table 1: BLEU-4 metrics compared to other methods

Dataset	Model	BLEU-4 Metric
Flickr8k	Log Bilinear (Kingma and Ba (2014))	17.7
	Soft-Attention (Xu et al. (2015))	19.5
	Hard-Attention (Xu et al. (2015))	21.3
	Ours (Soft-Attention)	12.3
MS COCO	Soft-Attention (Xu et al. (2015))	24.3
	Hard-Attention (Xu et al. (2015))	25.0
	sGPN (Zhong et al. (2020))	36.4
	Unified VLP (Zhou et al. (2020))	39.5
	OSCAR (Li et al. (2020))	42.9
	BLIP-2 (Li et al. (2023))	43.7

average of unigram, bigram, trigram and 4-gram precisions. From our train, validation losses and test BLEU-4 scores vs. Epoch curves presented in Figure 4, Figure 5 and Figure 6, we can say that we could have achieved same level performance (BLEU-4 of 19.5) as

shown in the original paper if we could have trained our model for about 50 epochs. As our model uses ResNet-101 pre-trained model in encoder, we can actually expect even higher performance than the original model which uses pre-trained Oxford VGGnet of Simonyan and Zisserman (2014) in the encoder. Table 1 gives the comparison of BLEU-4 scores of different deep learning image captioning models on Flickr8K and MS Coco datasets. We can see that when attention mechanism is coupled with latest transformer and graph based architectures, it has significantly worked better, compared to traditional CNN and RNN based architectures in encoder and decoder.

Figure 7 shows how our model worked for a sample image captioned from our test dataset, and our attention mechanism seemed to perform satisfactorily well. However, as we can seen from its visualization, it made an error by giving us the caption that "a little boy is sitting on a bed", which actually should be "a little boy is standing near a bed". Also, we observed that as Flickr8k dataset consists more of humans and animals like dogs, it gave satisfactory captions for images involving these animals, but failed to give good captions for images involving other objects and structures.

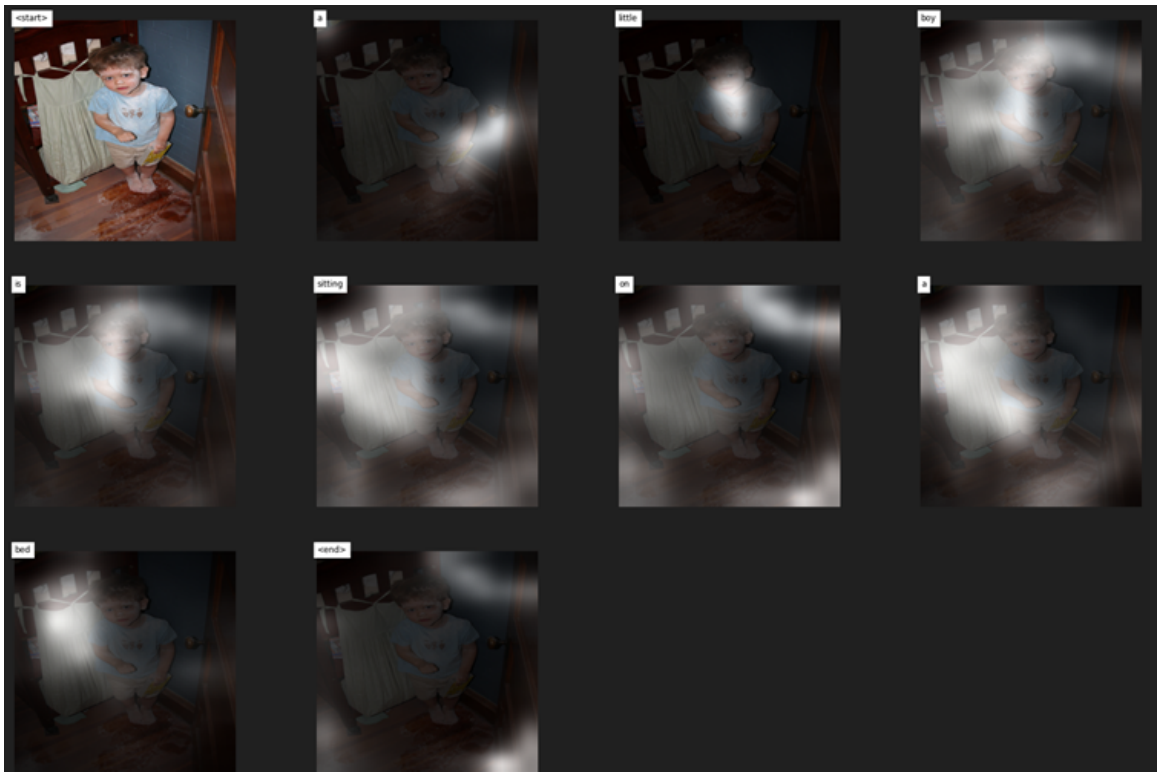

Figure 7: Example of model attending to the correct object (white indicates the attended regions, left top corner of the image indicates the corresponding word) on a test dataset image

4 Conclusion

With this paper, we could experiment and empirically assess the researchers suggested attention mechanism for image captioning. We saw that attention mechanism improved image captioning results compared to just employing 'show and tell' mechanism of Vinyals et al. (2015). We also compared this method with other latest methods used in the image captioning domain. We saw that BLEU-4 scores achieved through tranformers, large language and graph based architectures in encoder, decoder such as sGPN, OSCAR and BLIP-2 were even much higher than using just CNN and RNN with attention mechanism. Despite the numerous methods and solutions presented for the image captioning problem, there are still some major problems and challenges and that makes it increasingly developing research area. The generated captions still need to be higher in quality and are far from human-generated captions. Also, the datasets cannot cover the infinite real world. The evaluation metrics still need to be improved and are still not ideal for evaluating the precise performance of the models, as shown by Ghandi et al. (2022).

The performance of the supervised methods relies significantly on the quality of the datasets. However, datasets can not cover the real world regardless of how massive they are, and the applicability of supervised methods is limited to the set of objects the detector is trained to distinguish. On the other hand, datasets with image-caption pairs inevitably contain more examples of a specific situation (one example being: "man riding a skateboard"). These examples in the training data falsely bias the model towards generating more captions similar to those examples rather than including actual detected objects, as shown by Ghandi et al. (2022). The supervised paradigm overly relies on the language priors, which can lead to the object-hallucination phenomenon as well, as shown by Li et al. (2022). Considering the issues mentioned and gaps, unsupervised learning, and unpaired setting are of great potential. Also, the graph-based approach is expected to become even more popular in the near future. Transformers in combination with vision-language pre-training methods are also very likely to become standard practice, as shown by Ghandi et al. (2022).

Vision-Language Pre-Training (VLP) methods are frequently used in recent works and have shown promising performance. VLP methods and Transformers are likely to be inseparable components of models in the future of image captioning. VLP methods have been used to resolve some of the flaws with supervised methods and object detector-based designs, per Ghandi et al. (2022).

Furthermore, per Ghandi et al. (2022), more research is necessary for developing visual assistants intended for individuals with visual impairments. Such an assistant requires specific features to be incorporated, which makes it distinct from other applications of image captioning. The best models introduced in the previous research studies do not perform well as visual assistants and do not account for the unique requirements and needs of visually impaired people. An appropriate caption for a visually impaired person should prioritize the most critical aspects of the image and then describe other notable details. Consequently, a caption suitable for the necessities of visually impaired individuals contains denser and more detailed information compared to captions generated by traditional methods and models.

References

Taraneh Ghandi, Hamidreza Pourreza, and Hamidreza Mahyar. Deep learning approaches on image captioning: A review. *arXiv preprint arXiv:2201.12944*, 2022.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

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*, pages 4565–4574, 2016.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020.

Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17990–17999, 2022.

Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804*, 2021a.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021b.

Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

Abdur Rahman. GitHub - arkalim/PyTorch: PyTorch Codes — github.com. `https://github.com/arkalim/PyTorch`, 2020. [Accessed 16-Apr-2023].

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

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*, pages 3156–3164, 2015.

Yiyu Wang, Jungang Xu, and Yingfei Sun. End-to-end transformer based model for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36-3, pages 2585–2594, 2022.

Yuhui Wang and Diane Cook. Brain: A bidirectional generative adversarial networks for image captions. In *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence*, pages 1–6, 2020.

Qiaolin Xia, Haoyang Huang, Nan Duan, Dongdong Zhang, Lei Ji, Zhifang Sui, Edward Cui, Taroon Bharti, and Ming Zhou. Xgpt: Cross-modal generative pre-training for image captioning. In *Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10*, pages 786–797. Springer, 2021.

Kelvin Xu. GitHub - kelvinxu/arctic-captions — github.com. `https://github.com/kelvinxu/arctic-captions`, 2015. [Accessed 16-Apr-2023].

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 Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul 2015. PMLR. URL `https://proceedings.mlr.press/v37/xuc15.html`.

Yiwu Zhong, Liwei Wang, Jianshu Chen, Dong Yu, and Yin Li. Comprehensive image captioning via scene graph decomposition. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 211–229. Springer, 2020.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34-07, pages 13041–13049, 2020.