Analysis of Knowledge Distillation on Image Captioning Models

Srivatsan S[1] and Shridevi S[2]

[1]*School of Computer Science and Engineering, Vellore Institute of Technology, India*
[2]*Centre for Advanced Data Science, Vellore Institute of Technology, India*

Abstract
Image Captioning involves generating a textual description of an image, in the most accurate way possible. It requires a combination of Computer Vision and Natural Language Processing techniques, which can both be enhanced individually to improve the overall performance of the model. Knowledge Distillation is a model compression technique, where a smaller network learns from the predictions of a larger network to find a more optimal convergence space. It effectively improves the performance of the smaller network, without any problems like over fitting. Typically, the performance of image captioning is measured in terms of the BLEU score and CIDER score. In this work, we have tested and recorded the performance of three different Image Captioning model architectures, in terms of a large network, small network and knowledge distilled small network to track and analyse the effects of Knowledge Distillation. The results are promising when compared with the state of art models.
Keywords:
Image Captioning, Deep Learning, Knowledge Distillation

1. Introduction

Image Captioning is the task of generating a natural language description of the scenes and objects in an image. The sentences given as output must be grammatically correct and describe the image as accurately as possible. Therefore, the image captioning models must be able to recognize/describe the objects, their context within the scene, and their relationships, and frame them into a proper sentence in the target language. The image captioning tasks involves a combination of computer vision (CV) and natural language processing (NLP) techniques, where the CV parts come into play during object detection and recognition, and embedding into a feature vector. The NLP parts involve the conversion of this feature vector into the sentence, framing the words based on the object's location, action, and features, as well as their importance in the image. The typical input and output for an Image Captioning task is seen in Figure 1. There have been tremendous advancements in the efficiency and accuracy of CV and NLP techniques; hence this has been seen as a consequential increase in the performance of image captioning models. Models like Inception and other advanced CNNs for image detection, as well as Transformers for NLP tasks, provide access to state-of-the-art models on all devices.

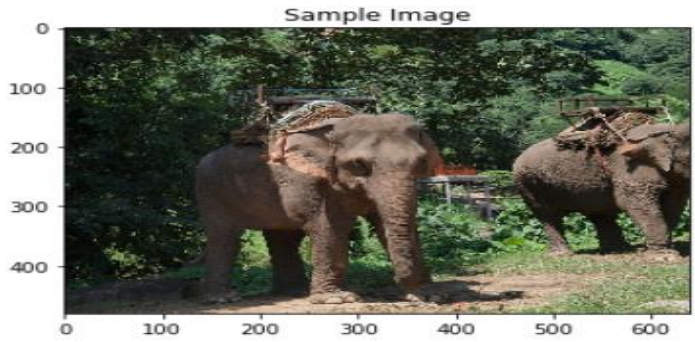

Figure 1: Sample Images with their caption

ACI'22: Workshop on Advances in Computation Intelligence, its Concepts & Applications at ISIC 2022, May 17-19, Savannah, United States
EMAIL: shridevi.s@vit.ac.in
ORCID: 0000-0002-0038-7212

In this work, we have seen the performance of 3 different models used typically for image captioning. Creating a characteristic language depiction of an image has drawn in interests as of late both in light of its significance in useful applications and on the grounds that it interfaces two significant machine learning fields: CV and NLP. Existing methodologies are either top-down, that start from an image and convert it into words, or bottom up, which concoct words portraying different parts of an image and afterward consolidate them. The general approach is seen in Figure 2 below.

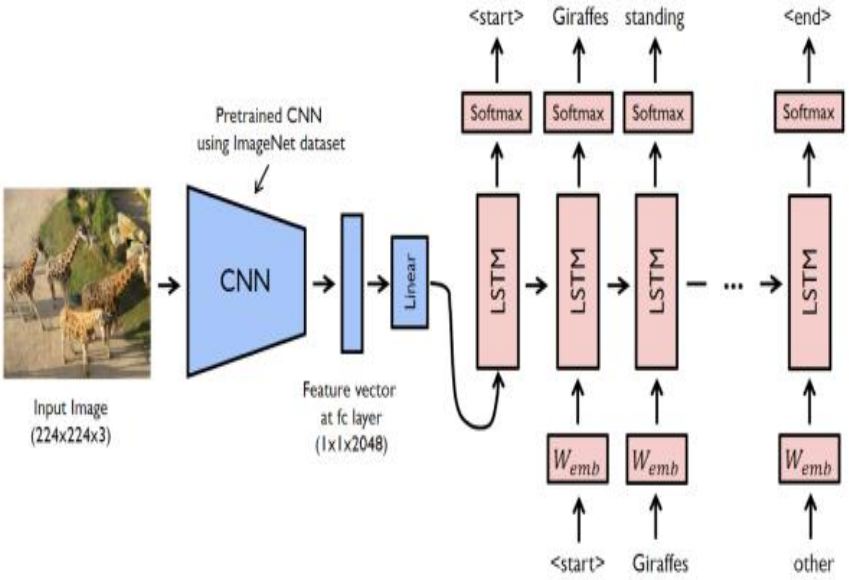

Figure 2: Image Captioning Network Architecture

The Transformer and BERT models, and their applications, exhibit that models with attention mechanisms, wherein the recurrent layers are replaced for the utilization of self-attention, offer far better performances in sequence modelling. This alternative likewise gives unique architecture modelling abilities, as the attention layer is utilized in a multi-layer structure differently. Image Captioning might need to be deployed in a number of use cases, including on edge devices or in areas with low bandwidth or low powered devices. They might not support large accurate models; hence the concept of Knowledge Distillation (KD) has been used to improve the performance of low accuracy models. By distilling the knowledge in an ensemble of models into a single model, Hinton et al [1] proposed the notion of Knowledge Distillation. In this work, we aim to compare the effects of this approach on a number of different models trained for the Image Captioning task. Multiple teacher models and student models are trained, with the student models gaining the distilled knowledge of the teacher and effectively improving its performance. The effects of KD have been analyzed on different image captioning models and performance metrics were recorded in section 4.

2. Related Work

Image Captioning has seen much advancement in recent years, in part due to the independence of the modules involved, namely CV and NLP modules. Any improvements in one of these fields can be shown to have a consequential performance improvement in the overall image captioning task too. There have also been vast improvements in model evaluations and performance metrics used, to ensure captions generated that are closer to the ground truth. The earliest works involving Image Captioning include articles like [2], and [3]. Image Captioning tasks have different criteria they can be split on, based on approach to the task as bottom-up or top-down or hybrid [4], or based on the techniques used as either classical machine learning or deep learning-based. Machine Learning would involve unsupervised learning models to detect and extract features from input data

and pass to a classifier model. Deep learning techniques allow models to be trained on large datasets for more accuracy, typically involving Convolutional Neural Networks in the image encoding process, and Recurrent Neural Networks in the decoding process. More recently, the emergence of the attention concept has led to its wide-spread use in any image recognition/ object detection tasks, due to the inherent correlation to human understanding of images. Papers like [5] can be referred to as the start of the trends in current research on image captioning, getting state-of-the-art results on multiple datasets. It was followed by [6] and [7], which used the attention mechanism in training to achieve the best results, as well as visualizing the parts focused on by the layers. More recent papers utilizing attention for image captioning include [8] [9]. [13] Uses an object relation transformer to exploit the spatial relationships between objects using geometric attention. Approaches using more modern methods like Vision Transformers, or Spatial and Semantic Graphs, have also been explored.

With respect to the decoding components of image captioning, typical approaches include greedy search and beam search. The advancements in deep learning have led to Recurrent Neural Network-based NLP models, in order to predict the highest probable sentences for the visual embedding. The LSTM based approach has been quite prevalent, with varying layers leading to better performance [10]. Attention layers have also been used in the decoder NLP model [11]. The latest and state-of-the-art approaches involve Transformers, from the paper [12]. Transformers provide inherent parallelism, which can be taken advantage of to provide faster training. Just like the success seen in object detection/classification tasks by pre-training large models like Inception V3, or YOLO and applying transfer learning to customize for the required tasks, the strategy can be replicated in these NLP tasks too, by pre-training large Transformer models like BERT and customizing later. Distillation in Image Captioning has also been explored by [14] and other papers, but in this work, a comparison of the amount of degradation in performance for each model is being done. Hence, from the above article, we can see the current state of the art in Image Captioning techniques.

3. Datasets

The datasets used for the task are the Flickr8k and MS-COCO dataset. The Flickr8k dataset consists of 8,000 photos, each of which is accompanied by five different captions that provide clear descriptions of the important items and events. An example from the dataset is shown below in Figure 3.

Figure 3: Example from Flickr8k dataset

The COCO dataset is a large-scale object detection, segmentation, and captioning dataset. Each of the almost 82000 images includes 5 captions for training Image Captioning models. For this work, only a small subset of around 5000 images has been used. An example is shown in Figure 4.

Figure 4: MS-COCO Dataset

4. Image Captioning Models

The best image captioning models involve usage of a SOTA-level object detection/recognition model, for the initial image processing purposes. The final layer of the detection model is passed through an encoder model, which captures it as an embedded feature vector. This is finally passed through the decoder model (typically RNN architecture), which outputs the probabilities for all 5001 words in pre-defined vocabulary.

In this work, the performances of 3 different image captioning techniques, with 3 models of varying architecture sizes for each technique, have been compared to analyze the impact of KD. The 3 models used are:

1. ResNet50 + Beam Search
2. Inception V3 + Encoder-Decoder with Attention
3. EffNetB0 + Transformer Encoder-Decoder

To analyze the effects of Knowledge Distillation on these image captioning models, we have taken a teacher and student architecture for each and trained separately. The distilled model takes untrained student architecture and learns along with the teacher's predictions in order to more efficiently converge. So, 9 different models have been trained and tested.

4.1 ResNet50 + Beam Search

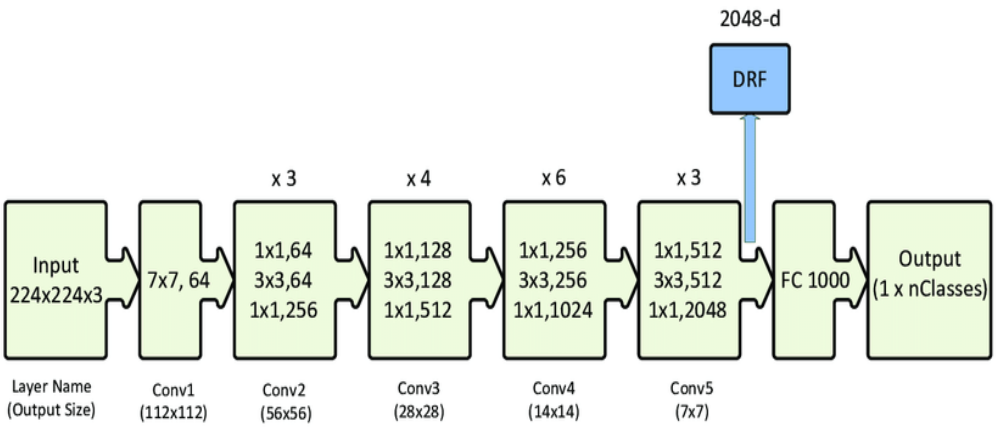

Figure 5. ResNet 50 Architecture

ResNet50 is the SOTA-object detection model used in this technique. It is a type of ResNet model having 50 layers, with 48 Convolutional layers, a MaxPool and an AveragePool layer. The model as in Figure 5 and Figure 6 are used for feature extraction and the embeddings are passed to the encoder-decoder network (RNN) which uses a beam search algorithm in order to give the probabilities of the next word for all the words in the pre-defined dictionary.

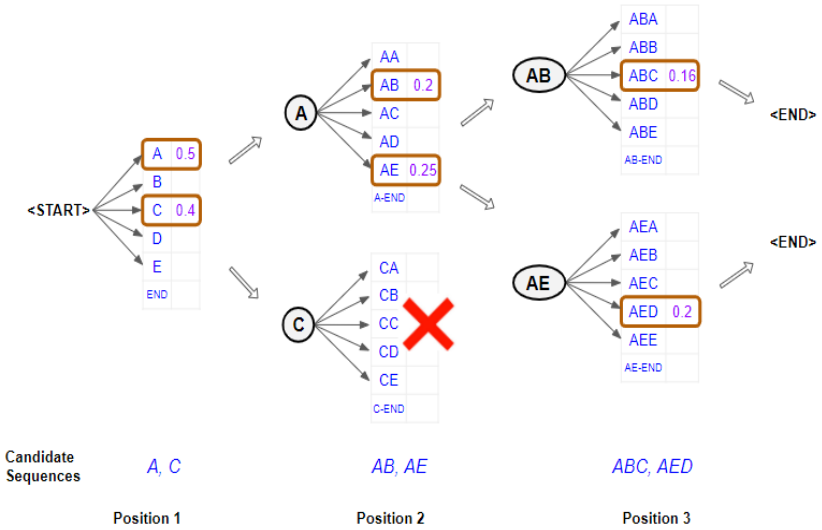

Figure 6. Beam Search

The model initially involves the ResNet50 model, whose final layer output is passed to the Encoder. The Encoder comprises of a fully connected linear/dense layer and a Normalization layer before being passed to the Decoder. The Decoder involves an embedding layer, an LSTM layer and the final fully connected layer of the vocab size for the probabilities. The below table 1 depicts the architecture details.

Table 1. Model 1 Teacher and Student Architectures

	Teacher Model	Student Model
Layer Type	Size	Size
ResNet	ResNet50	ResNet18
Fully Connected	512	512
BatchNorm	512	512
Embedding	5001,512	5001,512
LSTM	512,512,2	512,256,1
Fully Connected	512,5001	256,5001

4.2 Inception V3 + Encoder-Decoder with Attention

Because just the latest hidden state of the encoder RNN is used as the context vector for the decoder, the traditional seq2seq model is generally unable to effectively handle extended input sequences. In this model, we have used Inception V3 to preprocess all the images on the datasets. The captions are tokenized and the model is trained. The model consists of Encoder-Decoder architecture as in Figure 7 is similar to [17]. The output from the lower convolutional layer of Inception V3 is squashed and directly passed to the Encoder's Fully Connected Layer. It is then decoded by the RNN decoder (GRU with attention) and predicts the caption.

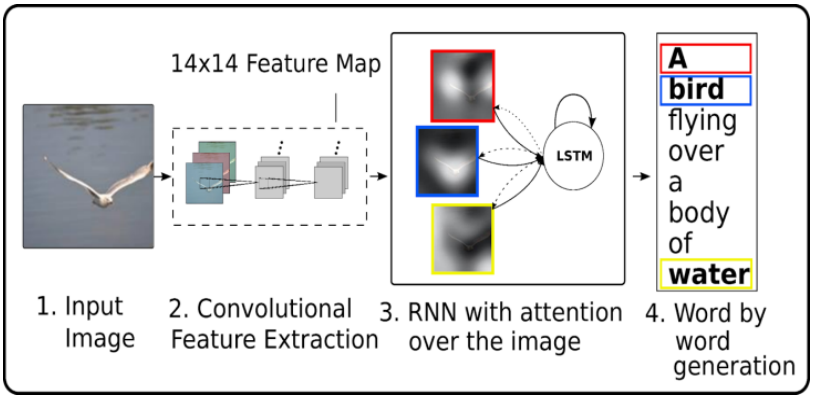

Figure 7. Flow of Model with Attention

The architecture details as in table 2 consists of the Inception V3 model for object recognition and image pre-processing. The final layer is directly passed through the dense/fully connected layer which then passes it to the decoder. The decoder consists of the Bahdanu attention layer, GRU layer for holding information for longer periods of time compared to RNN and is more efficient than LSTM, and dense/fully connected layers to predict the probabilities of the words for the image.

Table 2. Model 2 Teacher and Student Architectures

	Teacher Model	Student Model
Layer Type	Size	Size
Inception	InceptionV3	InceptionV3
Dense	256	256
Embedding Layer	5001, 256	5001, 256
Attention	512	256
GRU Layer	512	256
Dense	512	256
Dense	5001	5001

4.3 EffNet-B0 and Transformer Encoder-Decoder

The Transformer Neural Network is a unique design that tries to tackle sequence-to-sequence tasks while also being able to handle long-range dependencies. One major distinction in these networks is that the input sequence may be sent in parallel, allowing the GPU to be fully exploited while also increasing training speed. The vanishing gradient issue is also overcome by a substantial margin because it is based on the multi-headed attention layer. Transformers as in Figure 8 have been successfully adapted to many deep learning tasks, easily outperforming other network architectures.

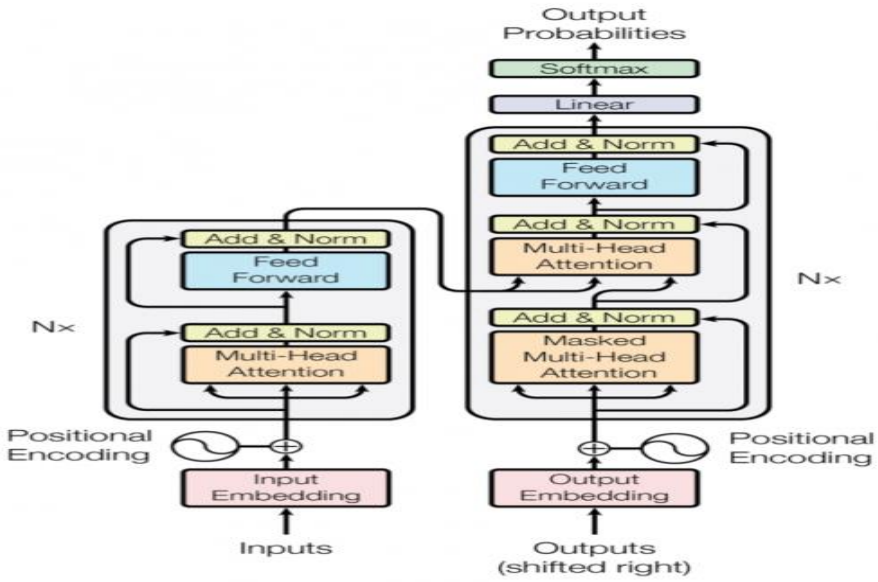

Figure 8. Transformer Architecture

In Model 3, the architecture as in table 3 involves the Efficient Net B0 model for object recognition and passes the final layer output to the Transformer-based encoder. The Encoder has a normalization layer and a dense layer for the inputs received from the EffNet model. It is passed through a Multi-Head Attention layer and is passed through another dense layer which connects to the Decoder. The Decoder involves an embedding layer for the inputs from the Encoder, followed by Multi-Head Attention layers which are normalized and squashed into dense layers with the final layer giving the probabilities.

Table 3. Model 3 Teacher and Student Architectures

	Teacher Model	Student Model
Layer	Size	Size
Efficient Net	EffNet-B0	EffNet-B0
Layer Normalization	-	-
Dense	512	256
MHA(1 Head)	512	512
Layer Normalization	-	-
Embedding Layer	512	512
MHA (2 Heads)	512	512
Layer Normalization	-	-
MHA(2 Heads)	512	512
Layer Normalization	-	-
Dense	512	256
Dropout	0.3	0.3
Dense	512	256
Layer Normalization	-	-
Dropout	0.5	0.5
Dense	5001	5001

5. Results

We have evaluated the results on BLEU and CIDER metrics. Bleu is the standard evaluation metric for measuring the amount of correspondence between the network output and the ground truth. The models were tested on a split of around 1000 images. Cider metric is a consensus-based metric to measure the similarity between generated output and the set of human-translated sentences. In Tables 4 and 5, we see how the performances of state-of-the-art architectures for image captioning and our architectures performances. Despite the smaller size of the knowledge distilled models, it shows comparable performance and is far more efficient than the larger models.

Table 4. Results on Flickr8k and MS-COCO datasets

	Results MS-COCO		Results on Flickr8k	
Model	Average BLEU-1 Score	CIDER	Average BLEU-1 Score	CIDER
Teacher(Model 1)	61.7	84.7	41.8	28.8
Student(Model 1)	48.0	76.8	32.1	22.1
DistilledStudent(Model 1)	54.3	80.4	34.3	26.3
Teacher (Model 2)	66.9	92.6	46.7	32.5
Student(Model 2)	53.4	81.0	37.3	24.6
DistilledStudent(Model 2)	58.1	84.5	39.6	27.9
Teacher (Model 3)	73.8	103.4	52.5	36.5
Student(Model 3)	64.6	91.1	42.3	28.2
DistilledStudent(Model 3)	67.3	96.2	46.6	29.1

Table 5. Karpathy Split Performances for MS-COCO

Model	BLEU-1	CIDEr
SCST[15]	78.1	114.7
GCN-LSTM[16]	80.2	117.9
Recurrent Fusion Network[17]	80.4	122.9
Meshed Memory Transformer [18]	81.6	129.3
Distilled Student (Model 3)	72.1	104.7

6. Conclusion and Future Discussion

In this work, we trained multiple teachers, students, and distilled models on the Image Captioning task. We used two standard datasets, namely Flickr8k and MS-COCO dataset to train the models. A comparison was made showing the results obtained for all the models, and we could see that the transformer models effectively outperformed their counterparts. We saw that in some cases, the student model outperformed teacher models of other architectures (Model 3 student vs Model 1 Teacher), whereas in other cases, the knowledge distilled model was given a boost was able to match/outperform other teachers (Model 3 distilled student vs Model 2 teacher). We can clearly see the use cases where a smaller model would be able to replace and to an extent, even outperform existing slower larger models. For future works, we can include more models in the study, as well as tuning the distillation parameters. We can also choose to study the effects of further distillation to establish a relationship function between performance and distillation.

References

[1] Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. "Distilling the knowledge in a neural network." arXiv preprint arXiv:1503.02531 (2015).

[2] J.-Y. Pan, H.-J. Yang, P. Duygulu, and C. Faloutsos, “Automatic image captioning,” in ICME, 2004.

[3] 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 ECCV, 2010

[4] Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[5] Anderson, Peter, et al. "Bottom-up and top-down attention for image captioning and visual question answering." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[6] You, Quanzeng, et al. "Image captioning with semantic attention." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[7] Xu, Kelvin & Ba, Jimmy & Kiros, Ryan & Cho, Kyunghyun & Courville, Aaron & Salakhutdinov, Ruslan & Zemel, Richard & Bengio, Y.. (2015). Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.

[8] Xuying Zhang, Xiaoshuai Sun, Yunpeng Luo, Jiayi Ji, Yiyi Zhou, Yongjian Wu, Feiyue Huang, Rongrong Ji; RSTNet: Captioning With Adaptive Attention on Visual and Non-Visual Words ; Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 15465-15474

[9] You, Quanzeng, et al. "Image captioning with semantic attention." Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, “Long-term recurrent convolutional networks for visual recognition and description,” in CVPR, 2015.

[11] Huang, Lun, et al. "Attention on attention for image captioning." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019.

[12] Vaswani, Ashish & Shazeer, Noam & Parmar, Niki & Uszkoreit, Jakob & Jones, Llion & Gomez, Aidan & Kaiser, Lukasz & Polosukhin, Illia. (2017). Attention Is All You Need.

[13] Herdade, Simao & Kappeler, Armin & Boakye, Kofi & Soares, Joao. (2019). Image Captioning: Transforming Objects into Words.

[14] Self-Distillation for Few-Shot Image Captioning Xianyu Chen, Ming Jiang, Qi Zhao University of Minnesota, Twin Cities

[15] Rennie, Steven J., et al. “Self-Critical Sequence Training for Image Captioning.” ArXiv Preprint ArXiv:1612.00563, 2016.

[16] Yao, Ting, et al. “Exploring Visual Relationship for Image Captioning.” Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 711–727.

[17] Jiang, Wenhao et al. “Recurrent Fusion Network for Image Captioning.” ECCV (2018).

[18] Cornia, Marcella, et al. "Meshed-memory transformer for image captioning." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.