

Replication of mPLUG, a SOTA for Image Captioning

Anurag Pendyala

app5997@psu.edu

Electrical Engineering and Computer Science



"man in black shirt is playing guitar."



"construction worker in orange safety vest is working on road."



"two young girls are playing with lego toy."

Figure 1: An example of Image Captioning where each image has a caption that best describes the contents [1].

1 Task

Image captioning generates some text describing the image. The basic task is when given an image, the model should output a natural language text as output. The text should be able to describe best what is in the image [2]. This is well described in Figure 1. However, there are a ton of issues while trying to figure out the description of the image.

Visual Understanding Capturing all the details from the image is not trivial. When a person looks at an image he can identify all the elements and find the relationships between these elements.

Grounding of Elements Capturing all the elements and grounding them is also necessary in case the caption needs to be specific to a particular object.

Evaluation Since the generation of captions is very subjective to every person looking at the image, there is no one valid answer. A proper metric needs to be established that captures human judgments [3].

One of the state-of-the-art methodologies for Image Cap-

tioning for the Microsoft COCO dataset is mPLUG which uses a cross-modal skip-connected network for predicting the caption of an image [4].

2 Related Work

Since this is an evolving problem, there have been quite a few works that build up to image captioning. A few of the notable works are given below.

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention [5] introduces an attention mechanism that concentrates on specific parts of the image. These parts are determined based on a grid that is created. This attention towards these parts helps the model focus on relevant parts of the image while generating captions. The work claims to be improving accuracy and details noted compared to other image captioning methods. However, one of the main flaws of this methodology is that sometimes the attention model might focus too much on a particular object and leave the remaining details a lot.

Tell Me What I Look At: Attention-based Explanations for Image Captioning [6] provides a methodology

where for a given object, the model predicts a caption with that object in focus. This helps the users understand how the model is trying to relate multiple objects. Despite, sounding successful, the explanations provided by the model are very complex and hard to understand. Sometimes, the explanations cannot be accurate at all.

Cider: Consensus-based Image Description Evaluation [3] gives a metric that can objectively give a score for image captioning which is very subjective by default. CIDER evaluates captions based on the consensus of multiple human judgments, rather than just the numeric accuracy of individual captions. This idea of subjectivity can prove to be a good metric. However, since it depends on human perceptions, getting scores can be time-consuming.

These works set up a good base to try mPLUG. **mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections** [4]. It is a state-of-the-art methodology. Attention mechanisms are computationally expensive, and sometimes, many models might lose a lot of details. mPLUG uses cross-modal skip-connections. The details of the model shall be described in Section 3.

3 Approach

mPLUG proposes a novel approach with cross-modal skip-connections for image captioning that addresses many of the existing methodologies [4]. The whole pipeline is represented in Figure 2. Each of the following subsections explains the various aspects of the overall model. The whole model is a summary of what has been given in the paper [4].

3.1 Architecture

The basic architecture is an encoder-decoder architecture.

Encoder processes the input image and extracts visual features. It can be any general image feature extractor like ResNet50, etc. Also, while training, the encoder even encodes the image captions into textual features.

Decoder has the *Fusion layer* that combines the text and visual features extracted by the encoder. A novel cross-modal skip-connection mechanism(explained in Subsection 3.2) that connects corresponding layers of the visual and textual encoders. The decoder also has an *attention* mechanism to focus on particular locations in the image. Finally, the *output layer* generates the required output.

3.2 Cross-Modal Skip-Connections

Cross-modal Skip-Connection consists of N skip-connected fusion blocks. In each of these blocks, a connected attention layer is connected to S asymmetric attention blocks. Text and Visual features are passed through an attention-block and S such blocks pass through a skip-connected fusion block.

SA is self attention layer, l^{n-1} is an input text feature, v^{n-1} is an input visual feature. LN is layer normalization. FNN is the final fully connected layer. The equations that represent the whole idea are as follows:

$$l_{SA}^n = LN(SA(l^{n-1}) + l^{n-1})$$

$$l_{CA}^n = LN(CA(l_{SA}^n, v^{n-1}), l_{SA}^n)$$

$$l^n = LN(FNN(l_{CA}^n) + l_{CA}^n)$$

4 Dataset

Microsoft Common Objects in Context Dataset(COCO) has around 330k images with 5 captions per image [7]. There are around 80 different categories they fall into. I have used only a subset of 500 images for both training and

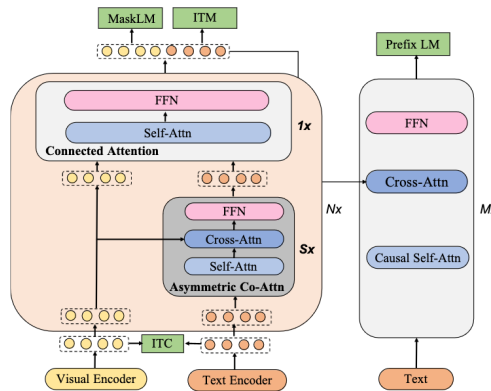


Figure 2: The pipeline has one text encoder and visual encoder passed through the cross-modal skip connections. Finally, fuse them.



The man at bat readies to swing at the pitch while the umpire looks on.



A large bus sitting next to a very tall building.

Figure 3: A couple of examples of the COCO dataset with the best available caption for the image. As can be seen, the caption tries to summarize what is there in the image comprehensively.

testing. The resource restrictions are the major reason behind shrinking the subset to this level. The dataset was specifically used for Image Captioning. For this application, the dataset is already well labeled with each image having 5 possible captions. Using natural language, these captions describe the image content, including objects, actions, and relationships between them. This format allows models to learn diverse and informative captions for each image. Figure 3 shows an example of such an image.

5 Results

5.1 Implementation Details

The project explores mPLUG’s state-of-the-art performance with Paperspace GPUs [8]. mPLUG uses ViT-B-16 [9] as the visual transformer. The GPUs provided were P4000 with 8GB of space. Since the resources are not exceptionally high, training and testing were performed in the mPLUG-base version. The final checkpoint was taken and trained for the 500-image subset of the Microsoft COCO dataset as discussed in Section 4. The model was trained for 5 epochs and the results achieved very pretty close to what was presented in the paper [4]. The losses and validation metrics during and after training are mentioned in the subsequent subsections 5.3.1 and 5.3.2. Due to the limitations of the machines available, the batch size was reduced to 16 from the traditional 64.

5.2 Metrics

mPLUG uses a variety of metrics to define its superiority. A few of them that are being focused on are BLEU and

CIDEr. BLEU-n is a Bilingual Evaluation Understudy that measures n-gram overlaps between the predicted and reference captions. The main focus is BLEU-4 which is a fraction 4-gram overlap. Overall, the higher the BLEU value, the better the model is performing[10]. The second metric is CIDEr, which is Consensus-based Image Description Evaluation. This metric focuses on the semantic similarity between the predicted and reference captions. It uses a consensus model trained on human judgments to evaluate the overall quality and informativeness of the predicted caption [3]. There are a couple more metrics to explain the accuracy [11]. They are:

- METEOR: Focuses on unigram precision and recall, considering word overlap between generated and reference captions.
- SPICE: Leverages semantic similarity between captions, evaluating how well they capture the image content in terms of meaning and relationships.

5.3 Replication

5.3.1 Training

The Loss vs Epochs graph can be seen in Figure 4. As expected, the loss drops from 0.555 to 0.331, suggesting that model training was successful. With the learning rate fixed at 10^{-5} , although the loss has been reduced, there is still a scope to reduce it further. Changing the learning rate on the fly could be more profitable. After training for each epoch, the model is validated/tested on a subset of the Microsoft COCO dataset to understand the results.

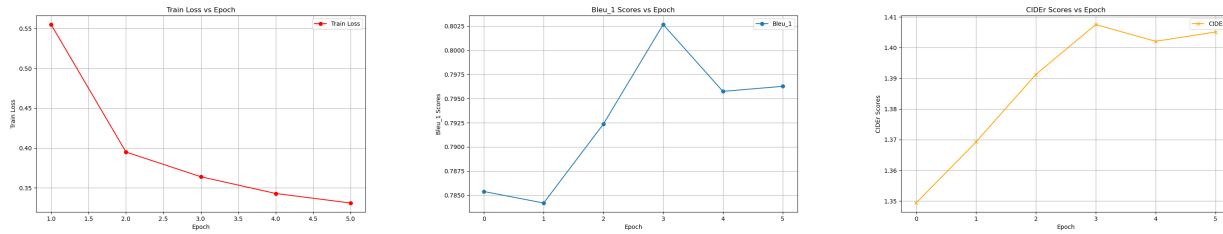


Figure 4: From the left, the plots exhibit the change in training loss, BLEU-4, and CIDEr values after every epoch. As expected, the loss keeps reducing after each epoch. The results are close to what mPLUG is doing for Image Captioning.

Model Name	BLEU4	METEOR	CIDEr	SPICE
<i>Replica</i>	<i>0.407</i>	<i>0.304</i>	<i>1.405</i>	<i>0.232</i>
mPLUG-large[4]	0.431	0.314	1.41	0.242
SimVLM_large[12]	0.43	0.334	1.426	0.247
LEMON_large[13]	0.406	0.304	1.357	0.235
VinVL[14]	0.385	0.304	1.308	0.234

Table 1: Comparison of Model Performance of the replicated one along with the SOTA mPLUG and some of the other contemporary models for image captioning.

5.3.2 Validation and Testing

The checkpoint is tested first and then after each and every training epoch, the model is tested on the subset of data. For every epoch, multiple metrics are used to evaluate the model. The following are the metrics calculated:

- BLEU-1
- BLEU-2
- BLEU-3
- BLEU-4
- METEOR
- ROGUE_L
- CIDEr
- SPICE

The main focus is on BLEU-4 and CIDEr which are explained in Section 5.2. The progression can be seen in Figure 4. Overall, the validation results indicate that the mPLUG model with ViT-B-16 backbone generates captions with high accuracy and quality. Despite the scores don't exactly overlap with the SOTA model, the results are very close. Probably further training and a bigger model can be

better.

5.4 Summary

To summarize the overall results, let us call the model that was trained as Replica. Table 5.3.1 summarize the overall table. Along with how it ranks with mPLUG, the SOTA, it also compares mPLUG with other current SOTA for image captioning. Again, mPLUG is SOTA for image captioning for the Microsoft COCO dataset alone [15].

A few images with predicted captions can be found in Figure 5. As you can subjectively see that, the caption for an image describes the image very effectively. With CIDEr score of close to 1.5, it is expected. The results match similar to the ground truth as well.



Figure 5: The prediction for the images are on the top. The results are very close to what is actually in the image.


6 Possible Improvements

Training was only possible for a few epochs from one of the checkpoints [16]. Besides this was only for the base model. The restrictions on the availability of resources have prevented the exact replication of mPLUG [4]. The authors have provided a large model as well which might yield better results. The learning rates can be tweaked a little to improve the overall model performance as well.

The method is SOTA specifically for the COCO dataset. Probably other datasets might present other results. But the hope is the model is formidable and should be one of the best for Image Captioning [15].

Finally, the mPLUG works for a lot more vision and language tasks than just image captioning. There has been updates on this code as well with the new iteration being mPLUG2 which is actually a state of the art methodology for visual grounding [17].

7 Code Repository

The forked codebase can be found in /mPLUG (forked and updated). The main repository is at [16]. There were quite a few changes to the original code to fit the limitations of the system provided. Along with this, there is another folder that is exclusively for the report which plots the training, evaluation plots. There is code also to give some predictions results with image captions.

...

References

- [1] P. Radhakrishnan, *Image captioning in deep learning*, en, Oct. 2017. [Online]. Available: <https://towardsdatascience.com/image-captioning-in-deep-learning-9cd23fb4d8d2>.
- [2] L. Ke, W. Pei, R. Li, X. Shen, and Y.-W. Tai, "Reflective decoding network for image captioning," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South): IEEE, Oct. 2019, pp. 8887–8896, ISBN: 9781728148038. DOI: 10.1109/ICCV.2019.00898. [Online]. Available: <https://ieeexplore.ieee.org/document/9009778/>.
- [3] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2015, pp. 4566–4575.
- [4] C. Li, H. Xu, J. Tian, et al., "Mplug: Effective and efficient vision-language learning by cross-modal skip-connections," *arXiv preprint arXiv:2205.12005*, 2022.
- [5] K. Xu, J. Ba, R. Kiros, et al., "Show, attend and tell: Neural image caption generation with visual attention," en, in *Proceedings of the 32nd International Conference on Machine Learning*, PMLR, Jun. 2015, pp. 2048–2057. [Online]. Available: <https://proceedings.mlr.press/v37/xuc15.html>.
- [6] M. K. Shaikh and M. V. Joshi, "Recursive network with explicit neighbor connection for image captioning," in *2018 International Conference on Signal Processing and Communications (SPCOM)*, Bangalore, India: IEEE, Jul. 2018, pp. 392–396, ISBN: 9781538638217. DOI: 10.1109/SPCOM.2018.8724400. [Online]. Available: <https://ieeexplore.ieee.org/document/8724400/>.
- [7] T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft coco: Common objects in context," in *European Conference on Computer Vision*, Springer, 2014, pp. 740–755.
- [8] en, Apr. 2022. [Online]. Available: <https://blog.paperspace.com/best-gpu-paperspace-2022/>.
- [9] A. Dosovitskiy, L. Beyer, A. Kolesnikov, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," no. arXiv:2010.11929, Jun. 2021, arXiv:2010.11929 [cs]. [Online]. Available: <http://arxiv.org/abs/2010.11929>.
- [10] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, Association for Computational Linguistics, 2002, pp. 311–318.
- [11] D. Raj, *Metrics for nlg evaluation*, en, Sep. 2017. [Online]. Available: <https://medium.com/explorations-in-language-and-learning/metrics-for-nlg-evaluation-c89b6a781054>.
- [12] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "Simvlm: Simple visual language model pretraining with weak supervision," no. arXiv:2108.10904, May 2022, arXiv:2108.10904 [cs]. DOI: 10.48550/arXiv.2108.10904. [Online]. Available: <http://arxiv.org/abs/2108.10904>.
- [13] X. Hu, Z. Gan, J. Wang, et al., "Scaling up vision-language pre-training for image captioning," no. arXiv:2111.12233, Mar. 2022, arXiv:2111.12233 [cs]. DOI: 10.48550/arXiv.2111.12233. [Online]. Available: <http://arxiv.org/abs/2111.12233>.
- [14] P. Zhang, X. Li, X. Hu, et al., "Vinvl: Revisiting visual representations in vision-language models," no. arXiv:2101.00529, Mar. 2021, arXiv:2101.00529 [cs]. DOI: 10.48550/arXiv.2101.00529. [Online]. Available: <http://arxiv.org/abs/2101.00529>.

- [15] *Papers with Code - COCO Captions Benchmark (Image Captioning)*, en. [Online]. Available: <https://paperswithcode.com/sota/image-captioning-on-coco-captions> (visited on 12/07/2023).
- [16] C. Li, H. Xu, J. Tian, *et al.*, *Mplug*, 2020. [Online]. Available: <https://github.com/x-plug/mplug>.
- [17] H. Xu, Q. Ye, M. Yan, *et al.*, "Mplug-2: A modularized multi-modal foundation model across text, image and video," no. arXiv:2302.00402, Feb. 2023, arXiv:2302.00402 [cs]. DOI: 10.48550/arXiv.2302.00402. [Online]. Available: <http://arxiv.org/abs/2302.00402>.