EECS595 Final Project Report Group 10

Ke Liu
University of Michigan
kliubiyk@umich.edu

Zhongqian Duan
University of Michigan
duanzq@umich.edu

Lingjun Sun
University of Michigan
slingjun@umich.edu

Abstract

Image-to-Text is a common Vision-Language Task. Conversely, we have Text-to-Image tasks, which is helpful in data generation and can provide creative ideas. In our project, we tried to combine these two tasks together, to propose a feasible method of generating a suitable dataset for vision-language tasks such as Image Captioning, and to learn about different Image2Text models' performance in predicting the images created by Text2Image tools. We paid attention to one of the most popular image generators Midjourney AI recently, and some vision-language models(mPLUG, OFA). We will compare the models and utilize these tools to generate and analyze our own dataset.

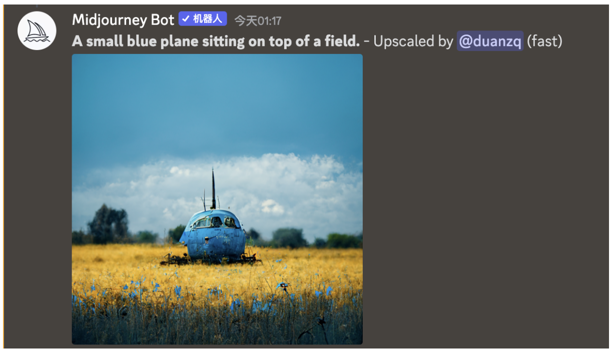

Figure 1: midjourney sample

1 Introduction

1.1 Image Captioning

Image-to-Text is a popular technology that converts images to text representations. It requires the vision-language model to extract important features from an image and generate natural language descriptions to accurately describe the content. One of its applications that we will focus on is Image Captioning.

Image Captioning(IC) is one of the most attractive topics in the research area. The objective of image captioning is to solve the semantic gap for computer vision, and allows computers to extract the features from graphics and transfer them to higher-level semantic information. Plenty of previous works showed remarkable developments in IC, and mainstreams for IC include the Transformer-based Encoder-Decoder approach, Attention Mechanism, and some other approaches(Conditional GAN, Reinforcement Learning to improve image captioning, etc.)

It is a challenging problem to achieve end-to-end training for Image Captioning since the visual encoder and language decoder doesn't share the same structure (Xu et al., 2022). From a most recent paper using mPLUG that achieves state-of-the-art performance on MS COCO Caption dataset, we were attracted by the unified Multi-modal Pre-training framework named mPLUG, which enables a cross-modal skip-connected network, and allows the fusion of visual and linguistic representations, thus provides an end-to-end model with achieved a high-efficiency performance. With such performance, it is useful on a wide range of vision-language tasks apart from images captioning, such as image-text retrieval and visual question answering(VQA). Similarly, the OFA model uses a Transformer as backbone architecture, and can also achieve high performance on a variety of vision-language tasks. We will mainly compare these two models and test them with the MSCOCO Caption, and the dataset generated by Text2Image generators.

1.2 Text-to-Image

Text-to-Image Tasks involve using text descriptions to generate corresponding images. These tasks are generally performed on Image generators, which are trained on large datasets of images with annotated captions. It covers a large range of applications, including the improvement of image recognition systems and the creation of personalized visual content.

As AI Image Generators becomes popular this year, many practitioners dedicate to grow the capabilities and ease-of-use of their image generator. The Midjourney AI is one of the examples, which is based on a deep generative model to generate images by descriptive text, and makes digital art more accessible to the public with a shareable discord channel. This popular trend arouse our interest in exploring the generated digital images. We wonder would computer recognize AI-generated images easier or harder than real-world images. We considered that Image Captioning is helpful in determining the understanding of the image. Therefore, we decided to use the digital images generated by Midjourney AI as our own dataset, and apply it to the Image Captioning Models such as mPLUG and OFA. After implementing training and fine-tuning, we will evaluate it by comparing the ground-truth image captions (same as the text descriptions we first used to generate our own dataset) with the newly-generated image captions (generated by Image Captioning Models), thus, we can evaluate the results to seek if the Text2Image tool can be helpful in generating fine datasets for visual-language tasks, and further reach a conclusion of computer image recognition system's performance.

2 Previous Work

For the Image Captioning Task, we explore the SOTA models in recent years. Generally, MSCOCO Caption is commonly being used to examine the performance of different models in Image Captioning Tasks. Looking into MSCOCO Captions's benchmark, the mPLUG model and OFA model by Alibaba Group were ranked the highest scores in BLEU-4, CIDEr, METEOR (Code).

2.1 mPLUG

As shown in Figure 2, mPLUG consists of two uni-modal encoders for image and text independently, a cross-modal skip-connected network, and a decoder for text generation.

First, it uses two unimodal encoders to encode text and images separately. The visual encoder directly applies the transformer on the image patches. The visual encoder encodes the input image patches into a sequence of embeddings $\{v_{cls}, v_1, v_2, ..., v_M\}$, and the text encoder encodes the input text messages into $\{l_{cls}, l_1, l_2, ..., l_N\}$. Next, these sequences of embeddings are fed into

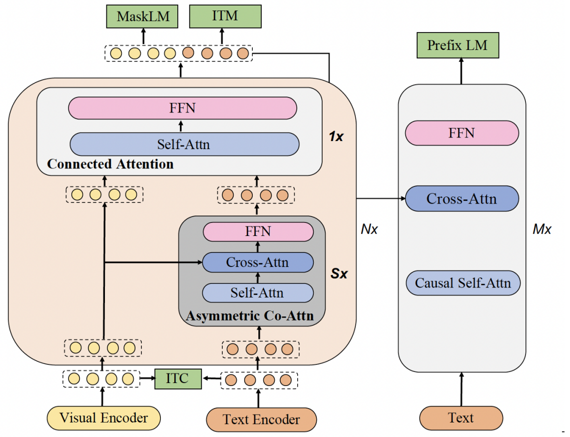

Figure 2: Mplug cross-modal skip-connected network (Li et al., 2022)

a cross-modal skip-connected network, which is used for cross-modal fusion of visual and linguistic representation.

The cross-modal skip-connected network includes multiple skip-connected fusion blocks. For each block, there are S asymmetric co-attention layers and a connected-attention layer. Explicitly, the asymmetric co-attention contains a self-attention (SA) layer, a cross-attention (CA) layer, and a feed-forward network (FFN), using the Linear layer for layer normalization. Once we fed the text feature l^{n-1} to the SA layer, its output will be calculated with the visual feature v^{n-1} in the CA layer, and we will get the visual-aware text representation l^n after passing the FFN. Equations (1) (2) (3) describe the process in the co-attention layer. For connected-attention layer, it is composed of a self-attention layer and a feed-forward network. It takes image feature v^{n-1} and text feature from the co-attention layer as input, and generates visual v^n and linguistic feature l^n as output for the next cross-modal skip-connected network(See equation (4) (5)).

Equations for each Co-Attention layer (Li et al., 2022)

$$l^n_{SA} = LN(SA(l^{n-1}) + l^{n-1}) \tag{1}$$

$$l^n_{CA} = LN(CA(l^n_{SA}, v^{n-1}) + l^n_{SA}) \tag{2}$$

$$l^n = LN(SA(l^{n-1}) + l^{n-1}) \tag{3}$$

Equations for each Connected-Attention layer (Li et al., 2022)

$$[v^n_{SA}; l^n_{SA}] = LN(SA([v^{n-1}; l^{n-1}]) + [v^{n-1}; l^{n-1}]]) \tag{4}$$

$$[v^n; l^n] = LN(FFN([v_{SA}^n; l_{SA}^n]) + [v_{SA}^n; l_{SA}^n]) \quad (5)$$

The output of the mPLUG cross-modal skip-connected network is a cross-modal representation, which will be fed into the transformer decoder and implemented with sequence-to-sequence learning to generate the result captions.

2.2 OFA

OFA is proposed with the purpose of achieving an omnipotent model, that is able to unify vision-language, vision-only, and language-only tasks. It is a Task-Agnostic and Modality-Agnostic sequence-to-sequence framework that once reached the state-of-arts in a various number of tasks such as Image Generation, Visual Grounding, Image Captioning, and Image Classification, to name a few. This model uses ResNet modules directly for visual feature extraction and follows the practice of GPT (Alec Radford and Sutskever, 2018), and BART (Mike Lewis and Zettlemoyer., 2020) to process the linguistic information and extract the features from text sequences.

By following the successful multimodal pretraining practices, OFA uses the Transformer encoder-decoder framework as unified architecture for all pretraining, fine-tuning, and zero-shot tasks (Wang et al., 2022). The encoder layer is composed of self-attention layer and a feed-forward network. The decoder layer consists a self-attention layer, a feed-forward network, and a cross attention for connecting the encoder's output and decoder together. Besides, OFA adds more implementations to improve its performance, such as stabilizing training and accelerating convergence. To reach the model's unification, it represents data of various modalities in a unified space and uses a unified vocabulary for all visual and linguistic representations.

2.3 Other Previous Works of Vision-Language Pre-training

Some other previous related works also achieved enormous success in Vision-Language Pre-training(VLP), such as CLIP(Alec Radford), OSCAR(Li et al., 2020), and VinVL (Zhang et al., 2021). According to the paper of mPLUG (Li et al., 2022), the typical approaches to VLP could be approximately divided into two types: dual encoder and fusion encoder. Dual encoders such as CLIP use two single-modal encoders for image and text separately and then apply straight-forward functions (dot product for example) to model the cross-modal interactions between them. This approach can achieve quite a computation efficiency as the image and text can be pre-computed and cached, however, they might fail for more complicated reasoning tasks such as visual question answering. Another approach, fusion encoder (OSCAR for example), is able to deal with complex reasoning tasks by utilizing deep fusion functions such as multi-layer self-attention or cross-attention networks.

From the Evaluation Results on the COCO caption (Figure 3), based on the same CIDEr Optimization approach, mPLUG has the highest score on BLEU-4, METEOR, and CIDEr than other models. mPLUG uses a visual transformer, which allows the model to be more computationally-friendly than using a pre-trained object detector to extract visual features of image patches. It also addressed the problem of information asymmetry that happens in the dual encoders model by introducing the cross-modal skip-connected network. Thus, we will mainly implement mPLUG model as well as OFA, which also ranked the second highest in COCO Image Captioning tasks, to fine-tune and test the dataset generated by the Text2Image tool.

Models	Data	COCO Caption							
		Cross-entropy Optimization				CIDEr Optimization			
		B@4	M	C	S	B@4	M	C	S
Encoder-Decoder	CC12M	-	-	110.9	-	-	-	-	-
E2E-VLP [19]	4M	36.2	-	117.3	-	-	-	-	-
VinVL [9]	5.65M	38.5	30.4	130.8	23.4	41.0	31.1	140.9	25.2
OSCAR [4]	6.5M	-	-	-	-	41.7	30.6	140.0	24.5
SimVLM$_{large}$ [7]	1.8B	40.3	**33.4**	**142.6**	**24.7**	-	-	-	-
LEMON$_{large}$ [33]	200M	40.6	30.4	135.7	23.5	42.3	31.2	144.3	25.3
BLIP [34]	129M	40.4	-	136.7	-	-	-	-	-
OFA [35]	18M	-	-	-	-	43.5	31.9	149.6	**26.1**
mPLUG	14M	**43.1**	31.4	141.0	24.2	**46.5**	**32.0**	**155.1**	26.0

Figure 3: Evaluation Results on COCO Caption "Karpathy" test split from paper (Li et al., 2022)

3 Methods Approaches

3.1 Dataset Generation

We used two tools: chatGPT and Midjourney to generate the dataset. This dataset is the first AI-generated dataset in the field of image caption.

There are several commands that were sent to chatGPT to help us to generate the textual side of the dataset. The commands are as follows:

1. Generate some random descriptive texts that are like image captions.

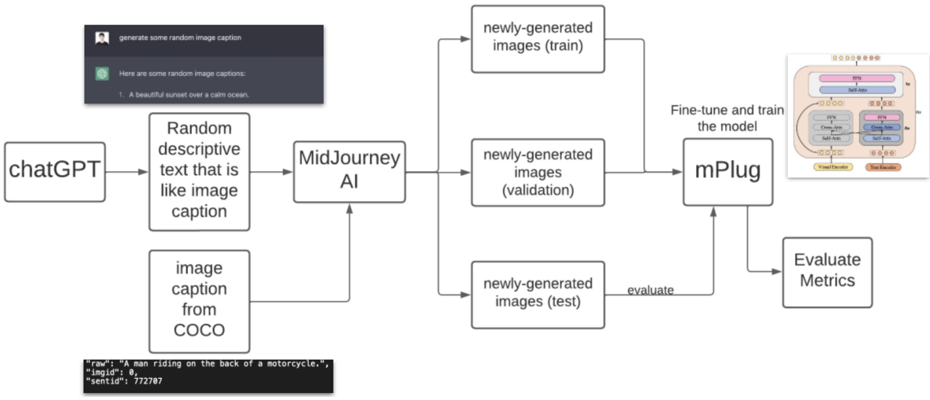

Figure 4: Our Method

2. For each sentence you input, give me five sentences with similar meanings, but with different structures.

3. For each sentence given below, extract the objects and combine the objects with "".

And the following three images are sample tests on each command on chatGPT.

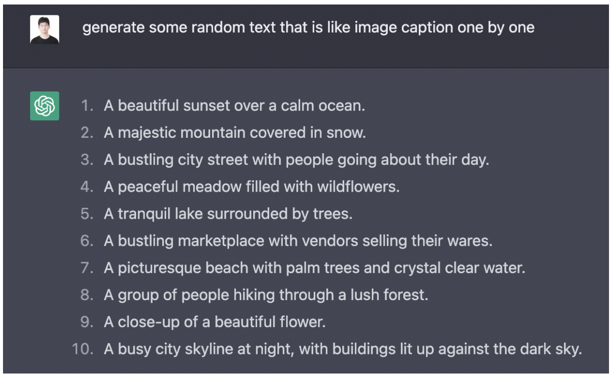

Figure 5: chatGPT sample result with command 1

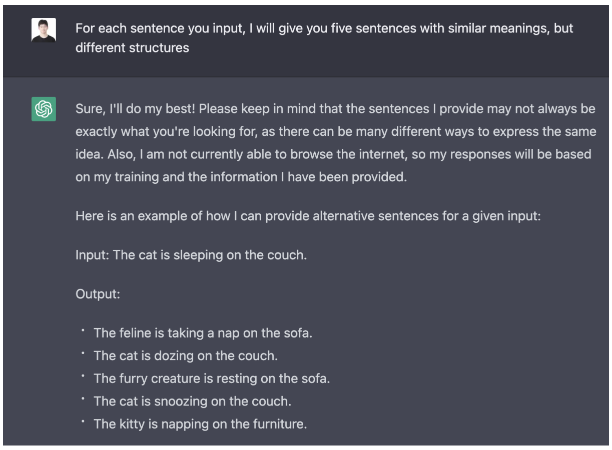

Figure 6: chatGPT sample result with command 2

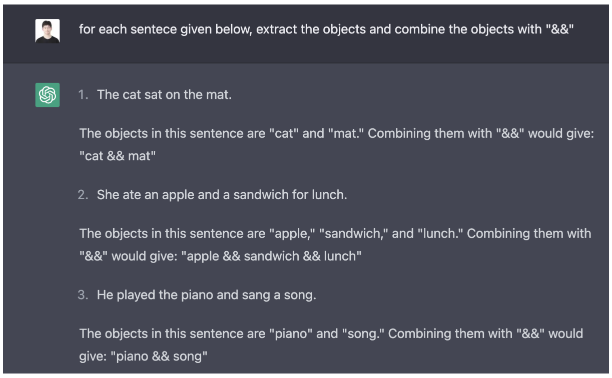

Figure 7: chatGPT sample result with command 3

We use Midjourney to generate the visual side of the dataset. The basic command is shown in Figure 8. We also used some advanced commands such as setting image size, upscaling, and making variations. An example of the image output is shown in Figure 1.

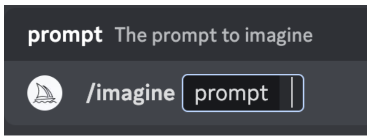

Figure 8: midjourney command

We generated two datasets based on the above method. In the first dataset, all image prompts were generated by chatGPT, and there are 45 images in total. And we split the dataset into training and testing partitions with 23 image-text pairs for the train and 22 image-text pairs for the test. In the second dataset, we mixed the image prompts generated by chatGPT with the captions given in the existing dataset COCO. The portion of them is 1 : 1. In this dataset, we generated 100 images using Midjourney in total. And we split the dataset into training and testing partitions with 80 image-text pairs for the training and 20 image-text pairs for the test. There are some sample train images with captions given in the appendix.

For each prompt and image pair, there are five image captions based on synonym conversion, and one object label generated by chatGPT. The final step is that we combined the image names, captions, and object labels together and transferred them into a JSON file in the same structure as the image caption JSON file for the COCO dataset using python. Note that the paring of the image and its corresponding text is based on image generated time since we generated each image based on the prompts line by line.

3.2 Reproduce the Paper Result

To implement mPLUG model and OFA model on our dataset, we first try to reproduce the paper's result with official open-source code.

For the mPLUG model, we used the pre-trained model mplug.en.based from code (Li et al., 2022). This model was pre-trained for 30 epochs with a total batch size of 1024. The text encoder and the skip-connected network are initialized with layers from the $BERT_{base}$ model, and the visual encoder is initialized by CLIP-ViT. The base architecture for the visual transformer is using the ViT-B/16 backbone. It uses an AdamW optimizer with 0.02 of weight decay as a preset, and with a learning rate warmed-up to 1e-5.

To run the test on MS COCO data, we use the Karpathy split the same as used in mPLUG paper, and set the learning rate unchanged as 1e-5, batch size equals to 64. After 5 epochs, the evaluation results in BLEU-4, METEOR, CIDEr, and SPICE all reach the baseline of mPLUG.

We also take a similar hyperparameter setting to test of OFA model. In the reproduction of baseline models on the MSCOCO image captioning dataset, we use the OFA_{Base} model, which uses ResNet101 as the backbone encoder, and has the same Hidden layer size as the mPLUG model we used. With a learning rate equal to 1e-5, and batch size equal to 64, we can get similar or even better results on top of the baseline results of the OFA_{Base} model(Table 2).

3.3 Fine-tune and Test on Midjourney Dataset

Our fine-tuning and testing is a two-stage process. In both stages, we initialize our model with the pre-trained weights of mPLUG.en.base. We first experiment with a small Midjourney dataset with only 45 images. In this stage, we finetune the images with the first 23 images and test the rest 22 images. We also apply random data augmentations such as flipping, shearing, or rotating. This stage is for verifying the basic functionalities of the model and helping us understand the gap between the pre-trained tasks and the specific downstream task on our Midjourney dataset.

For the second stage, we generate more images and divide the dataset with 100 images into 80 images for training and 20 images for testing. We finetune and test the images with the same process discussed above. Specifically, we finetune the dataset with different epochs (5, 30, 50), learning rates (1e-6, 1e-5, 1e-4), and batch sizes (1, 8, 16, 32). For every configuration, we record the highest result computed by the evaluation metrics discussed in section 4. This stage generates the final results used for analysis.

4 Evaluation and Analysis

For the purpose of image caption, a model must produce a relevant and fluid caption for each image. We analyze picture captioning on two datasets COCO Caption and our own Midjourney dataset. For the COCO Caption dataset, we are reproducing the results evaluated by the following metric techniques discussed below. As for the Midjourney dataset, we finetune the mPLUG by using the generated training dataset and then test the dataset using the same metrics. We split the dataset into a ratio of 4:1 for finetuning and testing the mPLUG on the Midjourney dataset. In accordance with mPLUG, we first adjust the model using cross-entropy loss and then for an additional 5 epochs using CIDEr optimization. (Li et al., 2022)

4.1 BLEU and BLEU-4

BLEU, or the Bilingual Evaluation Understudy, is a score for comparing a candidate's translation of the text to one or more reference translations. The BLEU metric ranges from 0 to 1. Few translations will attain a score of 1 unless they are identical to a reference translation. Due to this, even a human translator may not always receive a score of 1. For BLEU, it is significant to note that the score increases with the number of reference translations present in each sentence. When we analyze the results from the COCO dataset and our own dataset, we use the BLUE-4 metric that computes the cu-

Epoch	0	1	2	3	4	5
BLEU-4	39.6	43.11	45.21	46.83	46.99	**47.45**
METEOR	29.51	37.71	32.62	33.11	33.35	**33.59**
ROUGE_L	58.53	61.81	62.95	63.71	63.91	**64.24**
CIDEr	130.99	142.85	148.34	152.62	153.43	**155.53**
SPICE	22.99	24.39	25.02	25.57	25.61	**25.85**

Table 1: Reproduction Results of mPLUG.en.base Model (data shown in percentage form)

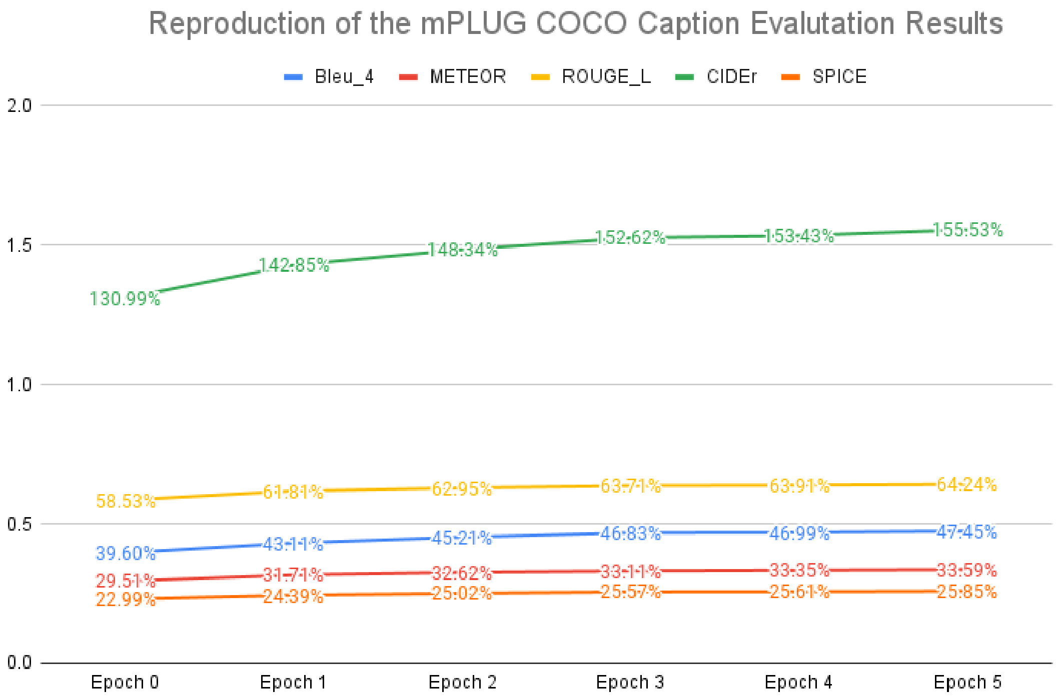

Figure 9: Reproduction Results in line chart form

mulative score which refers to the calculation of all individual 4-gram scores from 1 to 4, weighting them by computing the weighted geometric mean. (Papineni et al., 2002)

4.2 CIDEr

The CIDEr metric compares a generated sentence to a set of human-written ground truth sentences to determine how close they are. This metric has shown high agreement with consensus as assessed by humans. The concepts of grammaticality, saliency, relevance, and accuracy (precision and recall) are essentially captured by the CIDEr metric using sentence similarity. (Vedantam et al., 2014)

4.3 METEOR

The Meteor automatic evaluation metric scores machine translation hypotheses by aligning them to one or more reference translations. The criteria used to align words and phrases are exact, stem, synonym, and paraphrase matches. The alignments between hypothesis-reference pairings are used to determine the segment and system-level metric scores. (Banerjee and Lavie, 2005)

4.4 ROGUE-L

ROUGE, or Recall-Oriented Understudy for Gisting Evaluation, is used to assess automatic summarization and machine translation software in natural language processing. The L stands for the longest common subsequence (LCS). One advantage of using LCS is that it does not require consecutive matches but in-sequence matches that reflect sentence-level word order as n-grams. The other advantage is that it automatically includes the longest in-sequence common n-grams, therefore no

	BLEU-4	METEOR	CIDEr	SPICE
OFA_{Base}	**43.6**	28.2	139.8	**26.2**
OFA_{Base} Baseline	42.8	31.7	146.7	25.8

Table 2: Reproduction Results of OFA_{Base} Model

predefined n-gram length is necessary. (Lin, 2004)

4.5 SPICE

The SPICE metric is a relatively new metric that is used to analyze the ability that picture captions can identify objects, properties, and relationships between them. It has shown that SPICE reflects human judgment over model-generated captions on natural picture captioning datasets better than other n-gram metrics as Bleu, METEOR, ROUGE-L, and CIDEr. (Niu et al., 2022)

4.6 Choose between metrics

This downstream task mainly focuses on the ability to describe the generated image. Thus, we focus more on the precision of the generated sentence. So when finetuning the dataset, we choose the best configuration based on BLEU-4 and CIDEr score.

5 Discussion of Results

Let's donate the dataset that only has chatGPT generated prompts and 45 images as "Old Dataset" and the dataset which mixed prompts with captions from the COCO dataset and 100 images as "New Dataset".

For the image caption model, we use the pretrained model mPLUG base with Visual Backbone VIT-B-16, Text Enc Layers 6, Fusion Layers 6, and Text Dec Layers 6. The model was trained using the Midjourney dataset corresponding to the image-caption pairs, and it was applied to predict image captions for the unseen images in the test set. For each unseen image, we collected five image captions generated by chartGPT based on its image prompt.

We first tried different combinations of hyperparameters, such as learning rate, batch size, and epoch on "Old Dataset". We experimented with batch size equal to 64, learning rate(lr) equals to 1e-5 as default, and training 5 epochs, the best result was in epoch 3, which has reached 133.39 percent on CIDEr score. Then we lower the batch size to 8 and keep other hyperameters unchanged, we got lower scores on all evaluation metrics. We then altered the learning rate to 1e-4, the results seem improved a little bit, but still underperform than baseline tasks on Image Captioning.

For the Old Dataset, we found that no matter how we change the hyper-parameters, the training process tends to make the evaluation metrics worse. For example, in the case of batch size 8 and learning rate 1e-4, we trained for 50 epochs. But as the epoch increases, almost all the evaluation metrics decrease dramatically, where the CIDEr value decreases from 121.07 to 73.25.

Therefore, we tried to create the "New Dataset" from captions extracted from COCO and doubled the size of our dataset. We then use similar hyperparameters (batch size = 64 and learning rate = 1e-5) to test the "New Dataset". From Figure 13, we found that the best model finetuned to the New Dataset was generally better than the best model finetuned to the Old Dataset. For example, the value of CIDEr increased from 133.4 to 135.1.

6 Conclusion

In conclusion, we proposed a method that uses the Text2Image generator Midjourney AI to generate datasets for Image Captioning. We use Image Captioning tools such as mPLUG and OFA to predict the captions corresponding to those generated images. Our evaluation results with different evaluation metrics didn't provide as good results as the test on the MS COCO dataset, which shows that this method for dataset generation still needs to be modified and improved. Therefore, we provided some solutions and further improvements:

1. The descriptive captions generated by chatGPT is too abstract to create a graph. It's better to use a more simple and clear text as original captions.
2. Limitation of Image Captioning model. The pre-trained dataset resources are limited and less creative, and the word embeddings are also limited for predicting more complex words.
3. Improvement on dataset size. The quantity of images we are able to generate at these states is within a hundred, which is far less than the common dataset for Image Captioning. With wide-range and various datasets, we might generate a better result by using them for Image Captioning.

Prompt to Midjourney	Generated image	Predicted caption
"a young girl inhales with the intent of blowing out a candle."		"a little girl is looking at a cake with lit candles on it."
"a bathroom that has a broken wall in the shower."		"a bathroom is shown with a broken wall and a broken sink"
"an airport filled with planes sitting on tarmacs."		"a large group of airplanes parked on a snowy airfield."

Table 3: Some sample Image Caption test result on the New Dataset

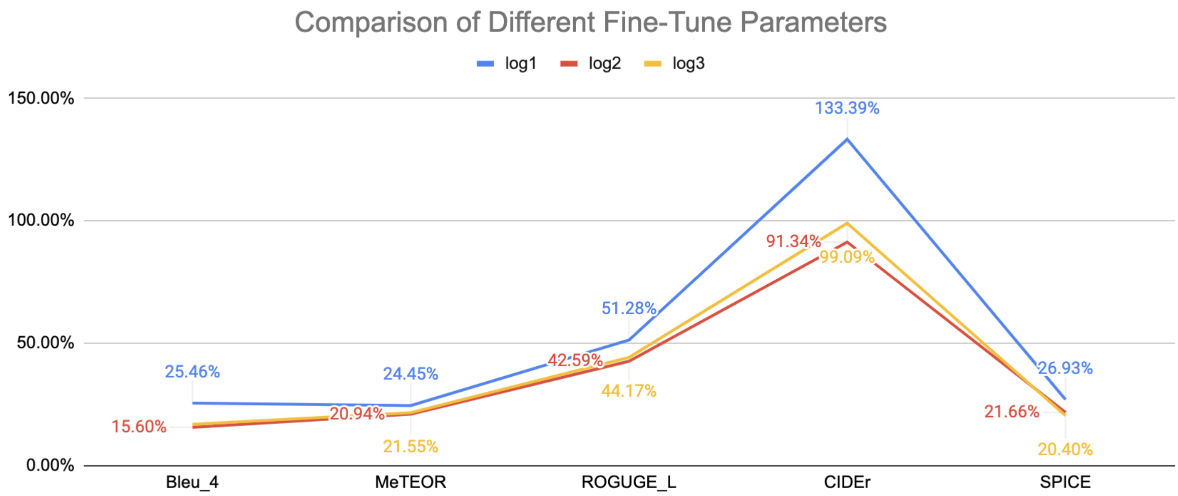

Figure 10: Fine-tune results picking from epoch with higher score. Log1: epoch 3, batch size = 64, lr=1e-5; Log2: epoch 5, batch size = 8, lr=1e-5; Log3: epoch 5, batch size = 8, lr=1e-4

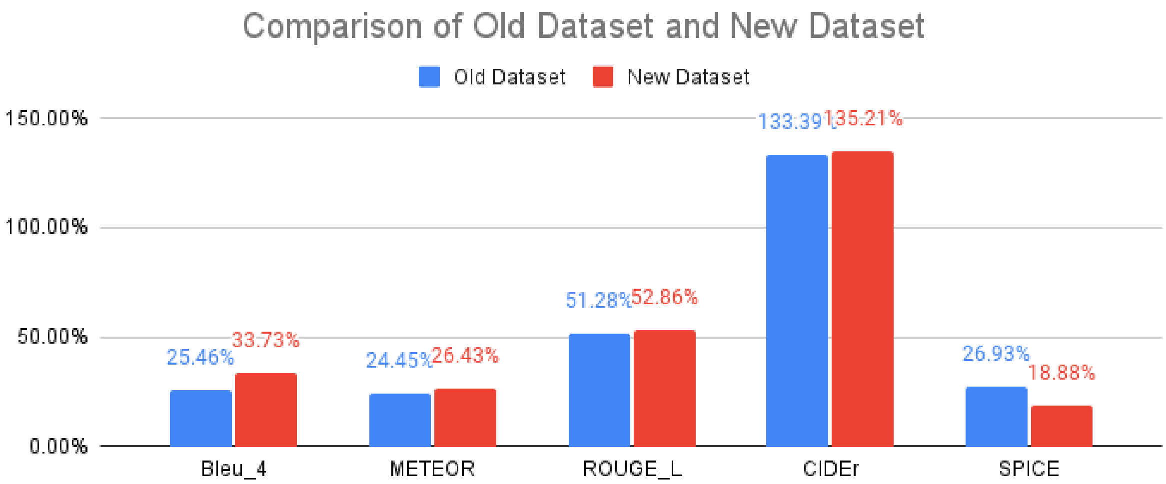

Figure 11: Comparison of Evaluation Results from Old Dataset and New Dataset

7 Division of Work

For the division of work, Ke Liu is responsible for finding previous work and reproducing the result using mPLUG and OFA model. Zhongqian Duan is responsible for generating the dataset. Lingjun Sun is responsible for the part of different evaluation metrics. We collaborated together to fine-tune the model and produce test results of our own dataset, and we also worked together to analyze our evaluation results as well as the problem of dataset generation.

References

Chris Hallacy Aditya Ramesh Gabriel Goh Sandhini Agarwal Girish Sastry Amanda Askell Pamela Mishkin Jack Clark et al. Alec Radford, Jong Wook Kim. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020.*

Tim Salimans Alec Radford, Karthik Narasimhan and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Paper With Code. Image captioning on coco captions.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mplug: Effective and efficient vision-language learning by cross-modal skip-connections.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Naman Goyal Marjan Ghazvininejad Abdelrahman Mohamed Omer Levy Veselin Stoyanov Mike Lewis, Yinhan Liu and Luke Zettlemoyer. 2020. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Chuang Niu, Hongming Shan, and Ge Wang. 2022. SPICE: Semantic pseudo-labeling for image clustering. *IEEE Transactions on Image Processing*, 31:7264–7278.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework.

Yang Xu, Li Li, Haiyang Xu, Songfang Huang, Fei Huang, and Jianfei Cai. 2022. Image captioning in the transformer age.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models.

A Appendix

Epoch	0	1	2	3	4
BLEU-4	20.61	24.55	24.74	25.46	24.77
METEOR	18.78	22.81	24.04	24.45	23.16
ROUGE_L	44.57	50.10	50.94	51.28	50.01
CIDEr	100.37	125.64	133.56	133.39	125.95
SPICE	19.73	25.84	26.07	26.93	26.36

Table 4: Evaluate result Log 1 with Dataset, batch size = 64, epoch = 5, default lr=1e-5

Epoch	0	5	10	15	20	30
BLEU-4	20.61	15.6	14.00	14.24	14.48	12.66
METEOR	18.78	20.94	20.30	20.54	20.18	19.53
ROUGE_L	44.57	42.59	40.29	41.04	42.13	40.81
CIDEr	100.37	91.34	84.25	89.13	85.59	82.52
SPICE	19.73	21.66	21.04	21.75	20.77	20.94

Table 5: Evaluate result Log 2 with Dataset, batch size = 8, epoch = 30

Epoch	0	5	10	20	30	50
BLEU-4	18.07	16.78	9.43	7.16	10.39	7.19
METEOR	22.16	21.55	19.64	19.07	20.46	19.72
ROUGE_L	46.59	44.17	41.83	38.37	38.47	38.37
CIDEr	121.07	99.09	83.62	72.41	77.16	73.25
SPICE	26.07	20.40	20.18	18.13	17.77	17.68

Table 6: Evaluate result Log 3 with Dataset, batch size = 8, epoch = 50, lr=1e-4

Prompt to Midjourney & Generated image	Image captions
A spaceship launching into the depths of space. 	• A spaceship is seen launching into space. • A spaceship is launching into the depths of outer space. • A spaceship is taking off and heading into space. • A spaceship is launching into the vast expanse of space. • A spaceship is being propelled into the depths of space.
People on bicycles ride down a busy street. 	• Cyclists are seen traveling on a busy street. • Bicycle riders are moving along a crowded street. • People on bikes can be spotted on a busy street. • Bicyclists are going down a busy street. • A busy street is filled with people on bicycles.
Animals hunting a man in the night in the large forest. 	• The animals were hunting the man in the night, stalking him through the large forest. • As the man ran through the dark forest, the animals pursued him, determined to catch their prey. • In the night, the animals hunted the man through the dense forest, their eyes shining in the darkness. • The man was the target of the animals' hunt as he ran through the forest, trying to escape their clutches. • The animals chased the man through the forest at night, their instinct to hunt driving them forward.

Table 7: Several example train images with their prompts and image captions