Image Captioning using Noise-Injected CLIP

	Member	Student ID
1	Mạch Vĩnh Phát	20125104
2	Đào Nhật Quang	20125108

I. Introduction

In recent years, advancements in computer vision and natural language processing have spurred innovations in the domain of image captioning, aiming to bridge the gap between visual content and textual descriptions. Our project was conceived with the primary objective of developing an effective image captioning system, known as CapDec, to automatically generate descriptive captions for images. Our methodology diverges from the conventional path of training a decoder model to reconstruct texts from CLIP embeddings, as we observed its limitations during inference. To address the known domain gap between image and text modalities, we utilize different injecting noise into the embedding to observe and find which one gives the best performance.

II. Model Architecture And Experiments

`CapDec`, or `Caption Decoder`, is a method for image captioning that leverages the `CLIP` model and additional text data. This methodology revolves around training a decoder to reconstruct text from the textual embeddings provided by CLIP. To bridge the inherent gap between the modalities of image and text within CLIP, CapDec employs a technique known as `noise injection`. This method introduces controlled randomness during training, enhancing the model's ability to generalize effectively to new and unseen data. Demonstrating remarkable performance, CapDec attains state-of-the-art results across various image captioning tasks, spanning standard, cross-domain, and style-oriented captioning. Notably, its versatility extends to enabling seamless style transfer, enabling the generation of captions in diverse styles namely, formal, informal, etc. by leveraging training data representing those distinct styles.

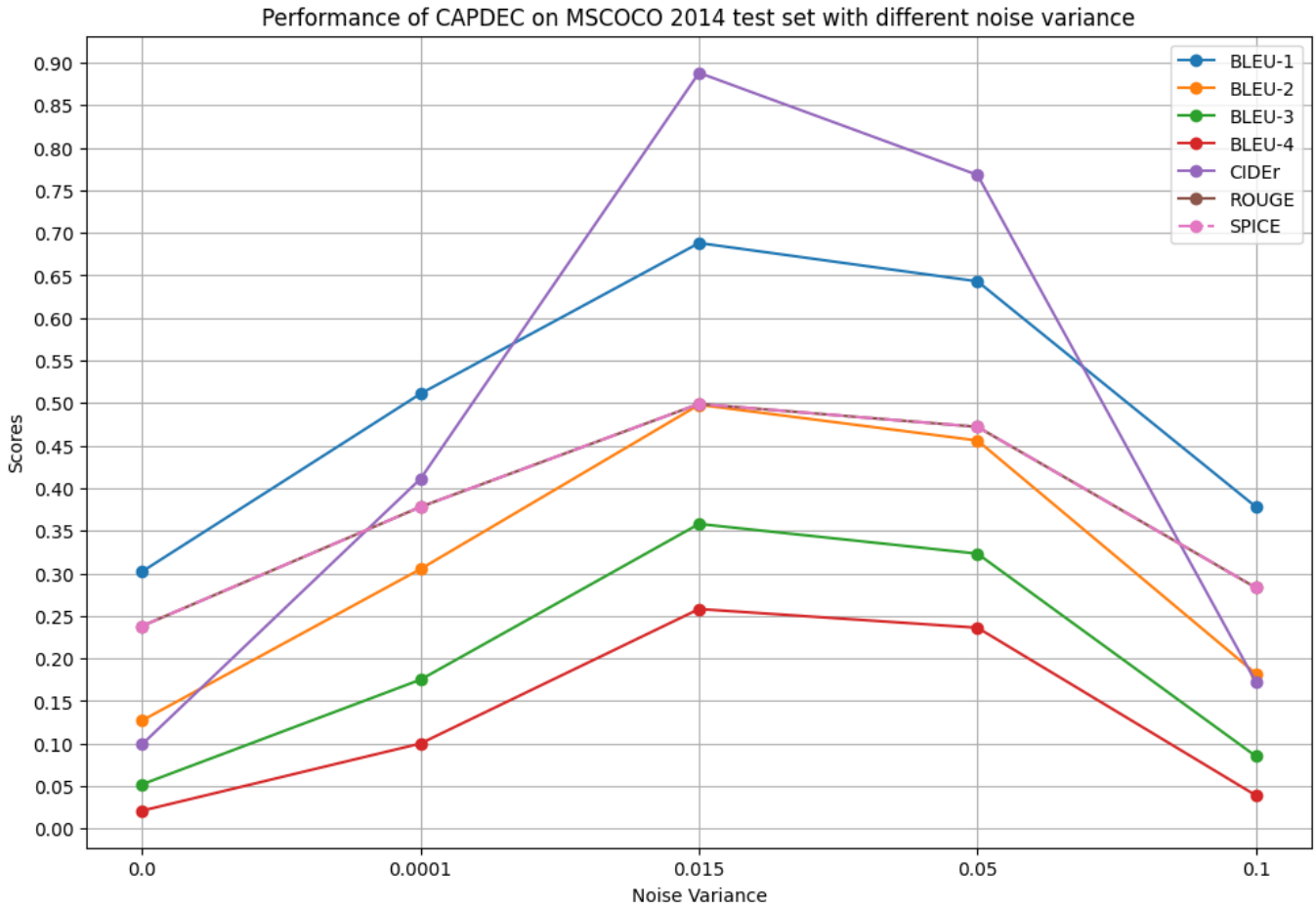
Our experimentation entails several key steps to evaluate and fine-tune the CapDec model:

- Noise variance: employing the pre-trained model, we conduct an investigation into different noise variance levels. This analysis aims to identify the optimal variance that establishes the best harmony between images and captions, refining the model's performance.
- Test datasets: utilizing `COCO 2014` as main dataset, followed by Nocaps, Flickr8k and VizWiz. This expanded evaluation allows for a more comprehensive assessment of the model's adaptability and performance across diverse datasets.
- Scoring metrics: to assess the model's performance, we utilize scoring metrics such as BLEU-1 to 4, CIDEr, ROUGE, and SPICE, where the implementation was the standard from pycocoevalcap.
- Manual evaluation: beyond automated metrics, a manual assessment is conducted by evaluating the model's output on a subset of 30 randomly selected images from both `COCO 2014` and `COCO 2017`. In addition, upload own images outside the datasets are available in our project. This hands-on evaluation provides additional qualitative insights into the model's performance.

III. Evaluation

1. Performance of different variance

- First, we use the pre-trained model with different nosie variance which is improving generalization and can be viewed as a data augmentation mechanism, namely `0.0`, `0.0001`, `0.015`, `0.05` and `0.1`.



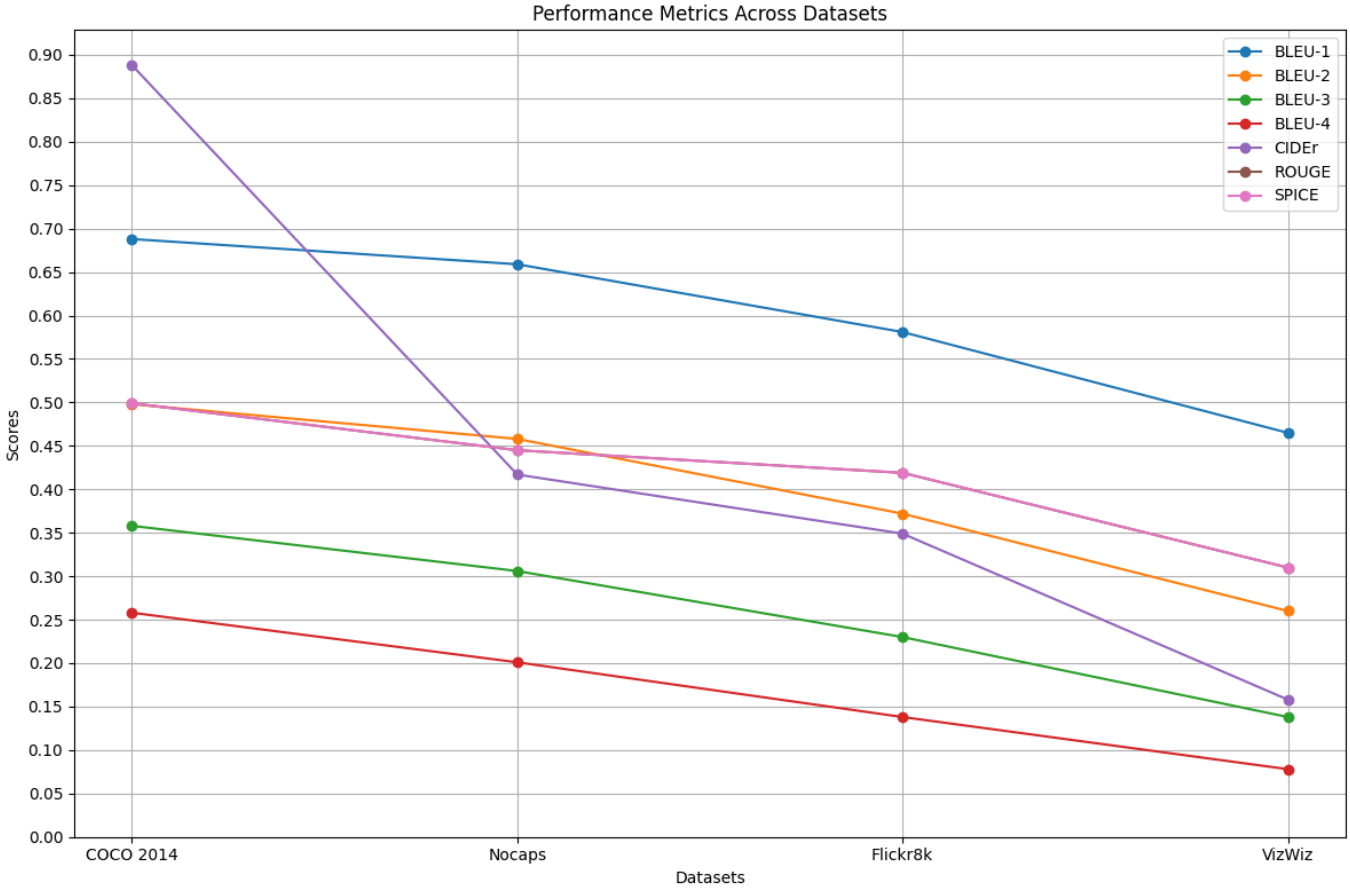
- Initially, the model exhibited with a variance of 0.0 where the scores are too low about 0.025 as BLEU-4 and 0.30 as BLEU-1. Nevertheless, a considerable increase in variance to 0.0001 led to a significant shift in performance, suggesting heightened variability in the model's predictions.

- The turning point was reached at a variance of 0.015, marking an archival peak in performance. This peak signifies a phase where the model excelled, demonstrating its highest level of accuracy as well as effectiveness. Following this peak, a gradual decline ensued, leading to a substantial plunge in performance when the variance spiked at 0.1. This drop illustrates a loss of reliability or accuracy in the model's outputs. Remarkably, ROUGE and SPICE metrics both provide the same scores.

- In conclusion, the model showcased a nuanced performance trajectory. The pinnacle of performance was achieved at 0.015 variance, indicating the model's optimal capabilities. Nonetheless, a subsequent sharp decline from 0.05 to the end highlighted a phase of decreased reliability or accuracy.

- Furthermore, we had evaluated the 0.01 of variance (but not included in this graph) and observed that the accuracy of them is comparable to 0.015 that is just a very little difference.

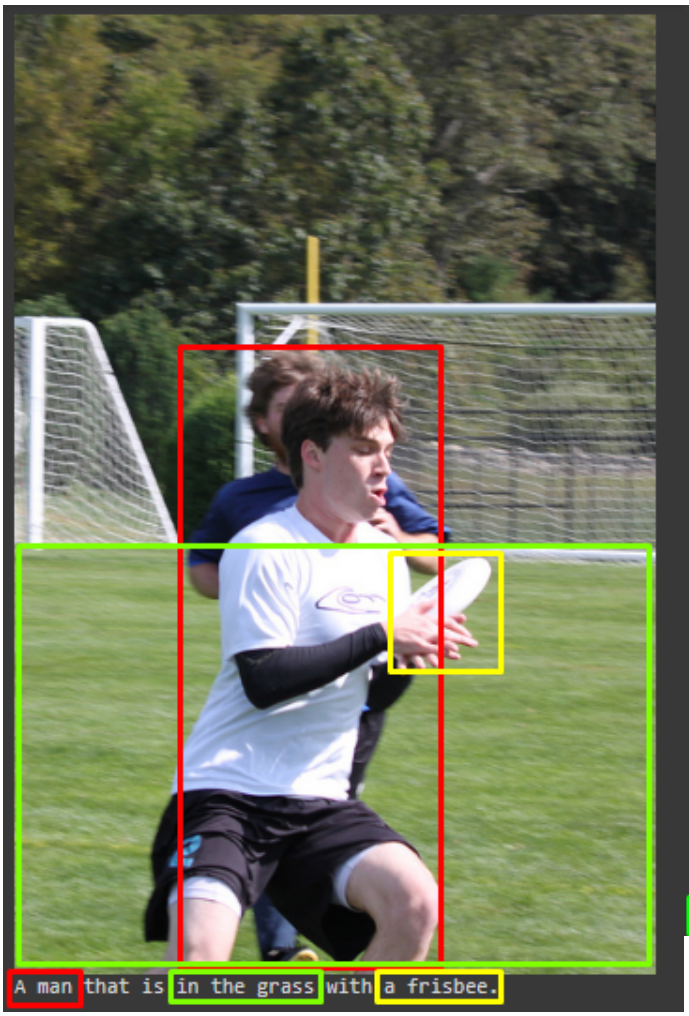
2. Performance Of Different Datasets



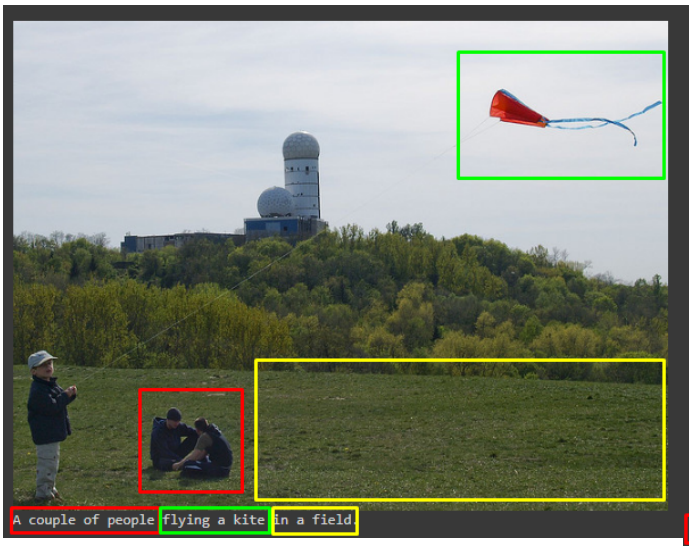
- After the evaluation, we observed that the model works well on 0.015 noise variance; therefore, we selected that variance to assess on different datasets.
- First, on `COCO 2014`, the model performs well on this dataset across most metrics, that means good generalization on similar data to its training set. While, although Nocaps has a slight lower than above test set, the model still maintains relatively good scores.
- Nevertheless, both `Flickr8k`, and VizWiz had a noticeable decrease in performance compared to `COCO 2014` and `Nocaps`, indicating some difficulty in generalizing to this dataset.

3. Results

Note: the rectangle frames are drawn by ourself to evaluate the accuracy, they're not a part of the model's output.

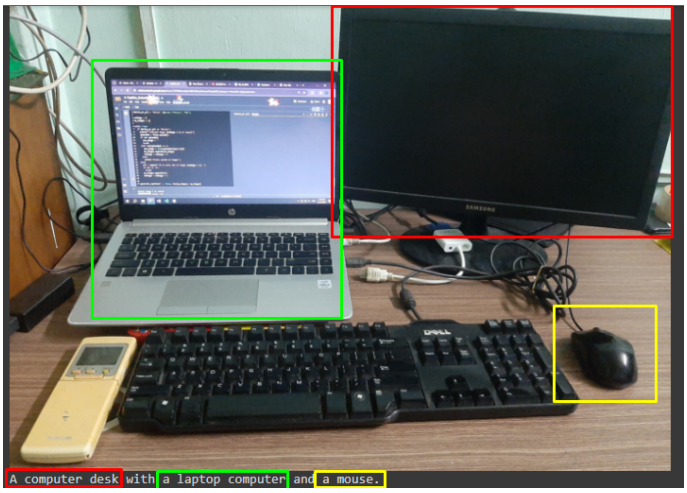

A man that is in the grass with a frisbee.


A bathroom with a white toilet sitting next to a sink.


A couple of people flying a kite in a field

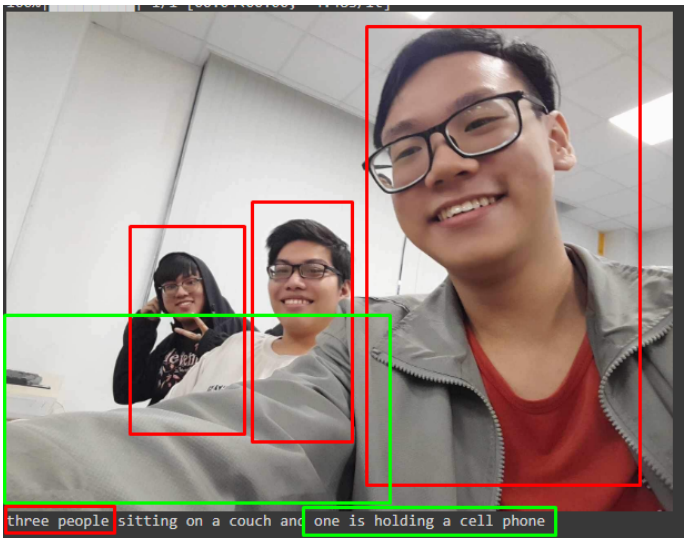

a couple of motorcycles that are parked next to each other

COCO 2014



Some our ground truth captions:

"A computer and a laptop are sitting next to each other."
"A computer and a laptop with a mouse and a keyboard."
"A remote is in front of the laptop and next to the keyboard."



Some our ground truth captions:

"Three people are sitting next to and one of them is holding the camera."
"A group of people are smiling and posing in front of the camera."
"Three men wearing the glasses are looking at the camera."

Our own test images

- According to test images from COCO 2014 and our own images, it can be clearly seen that most of the captions from the model capture the main context of the image as well as an acceptable level of conherence and fluency, but it sometimes still confuses some objects. The limitation for each caption is only generated by focusing on 2 or 3 objects to optimize the description.

IV. Fine-tuning And Futher Experiments

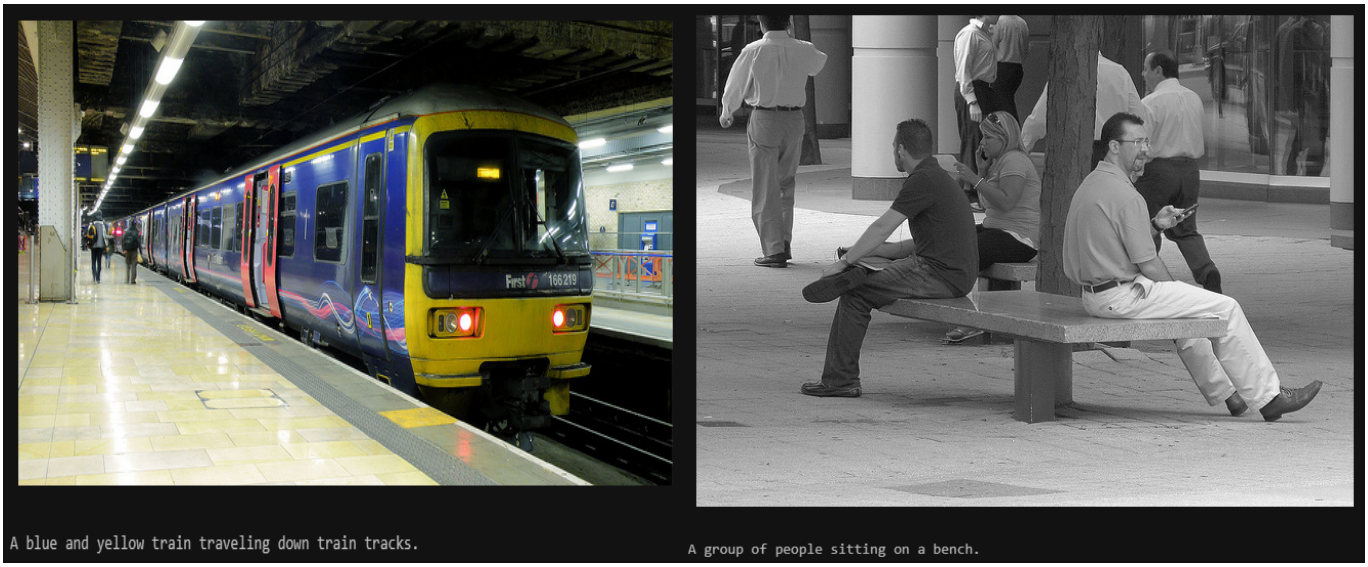
Fine-tuning Objective: building upon previous methods and proposals, the CapDec model undergoes fine-tuning. This process incorporates a combination of COCO captions and diverse open-text sources such as News and Shakespeare styles to generate new styles of captions. Based on prior evaluations, the model is fine-tuned using the 0.015 noise injection, identified as providing the most optimal caption-image match.

Testing Methodology: for post-fine-tuning comparison, cosine similarity calculations were performed using SentenceTransformer. This comparison analyzed embeddings of captions before and after fine-tuning.

Cosine Similarity Interpretation: a cosine similarity score above 0.5 suggests content retention in captions with modified sentence structures post-fine-tuning. Scores below 0.5 indicate significant changes or divergence in meaning/content.

Dataset / Open Text	News	Shakespeare
COCO 2014	0.64	0.66
Nocaps	0.51	0.54

Captions of CapDec-News



Captions of CapDec-Shakespeare

Before vs. After Fine-tuning (CapDec-News)

V. Dicussion

CapDec remarkable capability lies in achieving state-of-the-art image-captioning performance without any image input during training, a feat known as zero-shot learning. This unique attribute renders CapDec incredibly versatile and adaptable, showcasing its potential for a wide array of applications. Moreover, the model's prowess in style transfer stands out prominently. By harnessing unpaired textual examples, CapDec demonstrates the ability to generate captions in diverse styles, a feature not commonly found in other models, thereby offering a simple yet effective style transfer mechanism.

Nevertheless, certain weaknesses warrant consideration in the model's optimization and usage. CapDec heavily relies on text data for training, implying that the quality and diversity of the textual input significantly impact its performance. Any limitations or biases within the text data might consequently limit the model's capabilities. Additionally, the utilization of noise injection, while serving to bridge the gap between image and text modalities, poses a potential drawback. This technique, employed during training, could introduce instability or unpredictability in the model's outcomes, necessitating careful consideration and fine-tuning to mitigate these effects for optimal performance.

VI. Conclusion

The CapDec model is a revolutionary approach to image captioning that uses zero-shot learning to achieve high performance without any image inputs during training. Its versatility is evident in its ability to generate diverse caption styles from unpaired textual examples. However, the model's heavy reliance on textual data makes it vulnerable to biases and limitations in the input text. Noise injection, while bridging the gap between image and text modalities, introduces instability during training, causing unpredictable outcomes. Fine-tuning is crucial for optimal performance. The report highlights CapDec's strengths and suggests further exploration in image captioning, promising continued advancements and opportunities for improvement.