Extending the CapDec Model for Image Captioning

Lidor E.
Department of Industrial Engineering
Ben Gurion University of The Negev University
erezlid@post.bgu.ac.il

Dvir R.
Department of Industrial Engineering
Ben Gurion University of The Negev University
dvirre@post.bgu.ac.il

Abstract

Generating accurate and meaningful image captions remains a key challenge in artificial intelligence, as models must effectively bridge the gap between visual and textual representations. CapDec addresses this by injecting noise into CLIP embeddings to refine their alignment and improve caption generation. In this work, we extend CapDec by exploring alternative noise injection techniques, including T-Distribution noise, Gradient-based noise, and Cosine Similarity-Directed Gaussian noise, alongside hyperparameter tuning of the decoder architecture. We evaluate these modifications using semantic similarity analysis between generated captions and reference descriptions. Our results indicate that while different noise injection strategies influence performance, their improvements over Gaussian noise injection remain marginal. Similarly, increasing the number of transformer layers and attention heads has minimal impact on captioning quality.These findings suggest that CapDec's effectiveness is primarily shaped by how noise interacts with CLIP embeddings, and further improvements may require strategies beyond standard perturbation techniques.

1 Introduction

Image captioning is a fundamental task in artificial intelligence, enabling machines to describe visual content in natural language. It has widespread applications in accessibility, content creation, and human-computer interaction. However, traditional captioning models depend on large-scale paired image-text datasets, which are expensive to collect and often lack diversity. To address this, CapDec [1] introduces a text-only training approach, leveraging CLIP's pre-trained multimodal embeddings to generate captions without requiring paired supervision. While CapDec reduces dependency on labeled datasets, it still faces challenges due to the modality gap—the structural difference between image and text embeddings. To mitigate this, the original CapDec model applies Gaussian noise injection, yielding promising results. In this project, we explore alternative noise injection strategies and architectural modifications to further refine CapDec's performance. If successful, our modifications could enable a more effective text-only captioning model, reducing the need for large-scale paired datasets while improving alignment between text and image embeddings.

Final project report for the course *Introduction to Deep Learning*, IEM Department, Ben-Gurion University of the Negev, 2025.

2 Background

This section provides an overview of the key papers that form the foundation of our project. Our work primarily builds upon CapDec, which itself is based on CLIP, while also drawing insights from the Mind the Gap paper on modality gaps in multimodal models.

2.1 CLIP

CLIP [2] is a multimodal contrastive learning model trained on large-scale image-text pairs. It aligns visual and textual representations in a shared embedding space, enabling zero-shot classification and other vision-language tasks. The contrastive learning approach of CLIP inspired us to explore cosine-directed Gaussian noise as a means of refining the alignment between modalities.

2.2 CapDec

CapDec [1] introduces a decoder-based approach to reconstruct text from CLIP's textual embeddings. By training a decoder to reconstruct text from noisy embeddings, it compensates for the modality gap that exists between image and text representations. At inference, this decoder is applied to image embeddings, generating captions without requiring direct training on paired image-text data.

2.3 Mind the Gap: Understanding the Modality Gap in CLIP

While multimodal models like CLIP effectively align visual and textual representations, they still exhibit a modality gap, as demonstrated by [3]. This study shows that image and text embeddings occupy distinct regions within the shared embedding space due to model initialization and contrastive learning dynamics. Their findings provided a strong motivation for CapDec's noise injection strategy. Additionally, this paper inspired our own efforts to explore improved noise injection techniques to further mitigate the modality gap.

3 Contributions

We explored various strategies to enhance the CapDec model by addressing the modality gap through different noise injection techniques and hyperparameter tuning. In this section, we describe the enhancements and modifications we introduced in the training process and model architecture to improve performance. For convenience, we denote the text embeddings as e_{text} and the text embeddings after the noise injection as e^*_{text}.

3.1 Noise Injection Techniques

We experimented with different noise injection methods to better align the text and image embeddings.

T-Distribution Noise Injection Unlike Gaussian noise, which assumes normally distributed perturbations, T-distribution noise has heavier tails, making the model more robust to variations in the embedding space. We applied T-distribution noise to the text embeddings as follows:

$$e^*_{text} = e_{text} + t_{noise}, \quad t_{noise} \sim t(v) \tag{3.1}$$

where v is the degrees of freedom of the T-distribution.

Normalized Gradient Noise Injection Inspired by adversarial training, we perturbed the embeddings along the gradient direction of the loss function as follows:

$$e^*_{text} = e_{text} + \epsilon \cdot \frac{\nabla L}{\|\nabla L\|_2} \tag{3.2}$$

where ϵ is a hyperparameter controlling the magnitude of the gradient-based perturbation. The loss function remains the same as used in [1].

Cosine Similarity-Directed Gaussian Noise Instead of injecting random Gaussian noise, we introduced directional noise based on cosine similarity to better align with the CapDec model's foundation in CLIP embeddings. Since CLIP embeddings are inherently structured according to cosine similarity, we hypothesized that modifying text embeddings in a cosine-similarity-guided manner would be more effective than random perturbations. Our method is defined as follows:

1. Sample N noise vectors from a standard Gaussian distribution:
$$noise_i \sim \mathcal{N}(0, 1) \tag{3.3}$$
2. Create N noisy text embedding candidates:
$$e^{i*}_{text} = e_{text} + \beta \cdot noise_i \tag{3.4}$$
3. Filter candidates using a cosine similarity threshold:
$$\text{Keep } e^{i*}_{text} \text{ if } \frac{e^{i*}_{text} \cdot e_{text}}{\|e_{text}\|\|e^{i*}_{text}\|} > th \tag{3.5}$$
4. Compute weighted distance from the original text embeddings using the filtered noisy embeddings:
$$\bar{d} = \sum_i w_i \cdot (e^{i*}_{text} - e_{text}), \quad w_i = \frac{cosine(e^{i*}_{text}, e_{text})}{\sum_{j=1}^{n} cosine(e^{j'}_{text}, e_{text})} \tag{3.6}$$
5. Move e_{text} in the desired direction:
$$e^*_{text} = e_{text} + \bar{d} \tag{3.7}$$

Where β, th, N are hyperparameters that can be tuned.

3.2 Hyperparameter Tuning

In addition to noise injection techniques, we also fine-tuned CapDec's architectural hyperparameters to evaluate their impact on caption generation. Specifically, we experimented with:

Number of Transformer Layers (L): Changing the number of transformer layers allows the model to capture long-range dependencies in text embeddings.

Number of Attention Heads (H): The number of attention heads in the self-attention mechanism determines how well the model can focus on multiple semantic aspects of text embeddings.

Thus, we decided on trying different numbers of H and L.

3.3 Evaluation

To evaluate our model and compare it with CapDec, we adopted a different approach from the one used in the original paper. Instead of relying on the BLEU score, we employed the transformer-based model *all-MiniLM-L6-V2*[1] for semantic similarity analysis. This allowed us to measure the alignment between generated and reference captions more effectively, focusing on meaning rather than lexical overlap. We analyzed the top 15 images where our modifications achieved the highest scores and the top 15 images where CapDec performed best. This approach allows us to compare the performance of both models in their most favorable conditions, highlighting the strengths and weaknesses of each.

4 Experiments and results

The CapDec architecture consists of 8 attention heads and 8 transformer layers, along with Gaussian noise injection with a mean of 0 and a tunable variance ϵ. For consistency, all of our results were benchmarked against a baseline CapDec model configured with 8 transformer layers, 8 attention heads, and Gaussian noise injection with a fixed variance of 0.05. In every experiment we present one chosen value, which has the best results of all other values.

4.1 Experiment 1: Noise Injection methods

For each of the noise injection methods we tried different combination of hyper parameters to try and get the best results out of it.

T-Distribution: $t_3, t_5, t_7, t_{10}, t_{15}, t_{20}$

Figure 1: Comparison of the best results from the CapDec model (left) and our improved model (right) with 15 degrees of freedom. The blue bars represent CapDec's similarity scores, while the orange bars correspond to the model with the changes. The left plot highlights CapDec's highest-scoring outputs, whereas the right plot presents the best-performing results of our approach.

The left plot reveals a clear gap between the models, with CapDec consistently achieving high similarity scores, while the T-Distribution noise model underperforms on these images. In contrast, the right plot shows that CapDec's scores remain close to those of our model. While the T-Distribution Noise model frequently outperforms CapDec in its best cases, the difference between the two models is much smaller than in the first plot as shown in Figure 1.

Normalized Gradient Noise: $\epsilon \in \{0.1, 0.05, 0.01\}$

The left plot indicates that the Gradient Noise model generally underperforms on these images, with lower similarity scores in most cases. In contrast, the right plot shows that CapDec's scores remain close. Although Gradient noise outperforms CapDec in some instances, the overall gap between the models is small, suggesting that this method maintains stable performance without significantly impacting CapDec's effectiveness, as shown in Figure 2.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Figure 2: Comparison of the best results from the CapDec model (left) and the Gradient Noise model with a step size of 0.05 (right). The left plot highlights the 15 images where CapDec performed best, while the right plot showcases the cases where the Gradient Noise model achieved its highest scores.

Cosine Similarity-Directed Gaussian Noise:

- $N \in \{30, 50, 100, 200, 300, 400, 500\}$
- $\beta \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$
- $th \in \{0.7, 0.75, 0.8, 0.85, 0.9\}$

Figure 3: Comparison of CapDec and the Similarity-Driven Noise model with N=100, Beta=0.1, th=0.8. The left figure displays the top 15 images where CapDec achieves the highest scores, while the right figure highlights the best-performing cases for the noise model.

The left plot highlights a clear gap between the models, with CapDec consistently achieving higher similarity scores. This suggests that for these images, CapDec's embeddings, shaped by its own noise injection, already exhibit high cosine similarity with the reference captions, leaving little room for further refinement. In contrast, the right plot shows a much smaller gap, with the noise-injected model occasionally matching or surpassing CapDec. This indicates that leveraging cosine similarity to guide noise injection can refine embeddings in certain cases, though its impact varies across different images, as shown in Figure 3.

4.2 Experiment 2: Hyperparameter Tuning

Each of the hyperparameters of the architecture was tuned according to the following list:

- Transformer Layers: $L \in \{4, 6, 10, 16, 24\}$.
- Attention Heads: $H \in \{2, 4, 16\}$.

When we tuned one hyperparameter we fixed the other to the value they used in the real CapDec model. We chose to represent only the tuning we did on the number of transformer layers, whereas tuning the number of attention heads didn't improve the results the results.

Figure 4: Comparison of CapDec and the 10-layer model. The left plot presents the 15 images where CapDec obtained the highest similarity scores, whereas the right plot highlights the best-performing cases for the 10-layer model.

The left plot shows that the 10-layer model generally performs slightly worse, with lower similarity scores in most cases. However, the difference between the models is less pronounced compared to other methods. In contrast, the right plot, which highlights the best results for the 10-layer model, reveals that its scores are often close to or even exceed those of CapDec, as shown in Figure 4.

5 Code

Here you can find our work: https://github.com/erezlid/IEMDLProject.

6 Conclusions and future directions.

Our study aimed to extend the CapDec model by exploring alternative noise injection strategies and architectural modifications to further improve its ability to refine CLIP embeddings. While Gaussian noise injection, as originally proposed in CapDec, proved to be an effective baseline, our experiments investigated whether T-Distribution noise, Gradient-based noise, and Cosine Similarity-Directed noise could lead to more robust performance. Additionally, we examined the impact of increasing transformer depth and attention heads to determine whether a more complex decoder architecture would yield improvements.

Our results showed that while different noise injection techniques influenced CapDec's behavior, their overall impact remained marginal compared to the baseline Gaussian noise injection. Some methods, such as T-Distribution noise, performed well in certain cases but lacked consistency. Gradient-based noise maintained stable performance but did not significantly outperform CapDec. Cosine Similarity-Directed noise aligned embeddings more effectively but required careful tuning. Similarly, increasing the number of transformer layers or attention heads resulted in only minor variations in performance. These findings suggest that CapDec's effectiveness is primarily driven by the quality of noise injection, and further improvements may require more than just different perturbations or deeper architectures.

Future work could explore hybrid noise approaches, which combine techniques like T-Distribution and Cosine Similarity-Directed noise, could balance their strengths while mitigating weaknesses. Additionally, modifying CapDec's training process by incorporating contrastive objectives may improve robustness, potentially leading to more reliable performance gains. Finally, layer-wise noise injection, where different noise levels are applied at various transformer layers, could provide a more structured way to enhance CapDec's processing of CLIP embeddings.

Although our modifications did not significantly improve Gaussian noise injection, they highlight important directions for refining CapDec and enhancing text-only image captioning without large-scale paired datasets.

References

[1] David Nukrai, Ron Mokady, and Amir Globerson. Text-only training for image captioning using noise-injected clip. *arXiv preprint arXiv:2211.00575*, 2022. 1, 2, 3

[2] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[3] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022. 2