

OBSS AI Image Captioning Challenge

Participant Report

Name: Armağan Dağıstan

Email: armagandgstn@gmail.com

Phone Number: +905442245535

Kaggle Username: armaandastan

1. Tech Stack

- **Programming Language:** Python 3.10
- **Libraries:** PyTorch 2.1, Hugging Face Transformers 4.40
- **Models Used:** BLIP (Bootstrapped Language-Image Pretraining), BLIP Fine-Tuned
- **Tools & Platforms:** Google Colab, Matplotlib, Pandas
- **Evaluation Metric:** Frechet Generative Distance (FGD) with GTE-small embeddings

2. Summary

My solution used the pretrained **BLIP image captioning model** from Hugging Face. I experimented with two approaches: (1) zero-shot captioning using the pretrained model and (2) fine-tuning the same model on the provided dataset. Both models were evaluated using the FGD score. Interestingly, the zero-shot approach (score: 0.19) outperformed the fine-tuned version (score: 0.15), possibly due to overfitting on a small dataset or a loss in pretrained generalization.

3. Approach

Model 1: Zero-Shot BLIP (Score: 0.19)

The first approach directly used the pretrained BLIP model (Salesforce/blip-image-captioning-base) without any fine-tuning. Each test image was processed using BlipProcessor, and captions were generated with `.generate()` in a zero-shot setting.

Advantages:

- Leveraged rich generalization from pretraining on large corpora.

- Avoided overfitting risks.

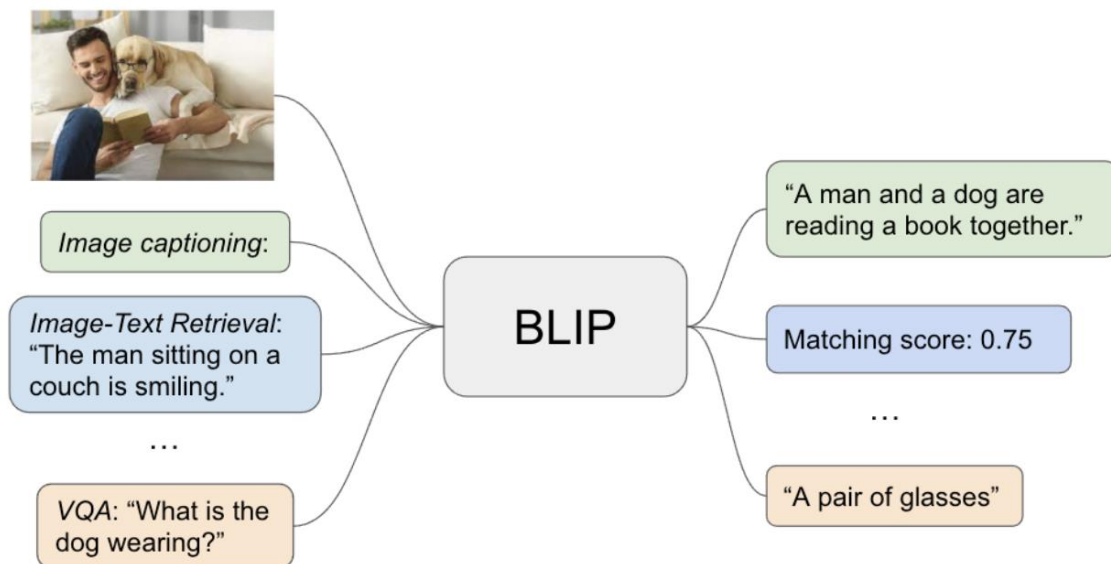
Model 2: Fine-Tuned BLIP (Score: 0.15)

In the second approach, I fine-tuned the same BLIP model using the image-caption pairs provided. The setup included:

- Optimizer: AdamW, Learning Rate: 5e-5
- Batch Size: 8, Epochs: 3
- Loss: Cross-Entropy with padding tokens ignored

Training Flow Diagram:

Image + Caption → BlipProcessor → BLIP → Fine-tuned BLIP → Generated Caption



Observations:




- Fine-tuning yielded fluent captions but with lower FGD alignment.
- Possible overfitting due to small dataset and lack of augmentation.
- Short training duration (3 epochs) might have prevented convergence .

Challenges:

- Fine-tuned captions possibly diverged semantically from original data in the embedding space.

- No validation set or scheduler was used, which could have improved performance. In addition, I had a limited time.

4. Sample Outputs

Image	Prediction	Comment
 <p>Image_100006</p>	<p>" the image shows a convenience store named " aguardit, " displaying various products for sale on the street."</p>	<p>Acceptable – general but scene is captured.</p>
 <p>Image_100011</p>	<p>" a black wine bottle labeled " pio cebare " with a circular design, indicating a red wine from brazil</p>	<p>Acceptable – text is not correct but general is good.</p>
 <p>Image_100156</p>	<p>the image features stacked boxes, showcasing various colors and designs, likely containing chocolates or cookies.</p>	<p>Poor, the model is bad at recognizing text.</p>



Image_100239

a baseball player in a
white and red
uniform prepares to
pitch on the field.

Good – accurate action and location.

Performance Discussion

- **Best Cases:** The zero-shot model performed well in identifying prominent subjects and context-rich scenes. It especially succeeded in images with common objects or clear backgrounds.
- **Failure Cases:** The fine-tuned model sometimes produced captions that looked fluent but were semantically misaligned (e.g., guessing objects not in the image). The lack of regularization and validation might have caused overfitting or semantic drift.

5. References

- [1] Li, J., Li, D., Xiong, C., & Hoi, S. C. (2022). *BLIP: Bootstrapped Language-Image Pretraining for Unified Vision-Language Understanding and Generation*. arXiv. <https://doi.org/10.48550/arXiv.2201.12086>
- [2] Hugging Face. (n.d.). *BLIP Image Captioning Base*. <https://huggingface.co/Salesforce/blip-image-captioning-base>
- [3] Salesforce AI Research. (2022, March 1). *BLIP: Bootstrapping language-image pretraining*. Salesforce AI Blog. <https://blog.salesforceairesearch.com/blip-bootstrapping-language-image-pretraining/>
- [4] OBSS Technology. (2025). *OBSS Intern Competition 2025 Dataset*. Kaggle. <https://www.kaggle.com/competitions/obss-intern-competition-2025>
- [5] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv. <https://doi.org/10.48550/arXiv.1908.10084>