1. Tech Stack

- **Programming Language:** Python 3.11
- **Libraries and Frameworks:**
    - PyTorch 2.1
    - HuggingFace Transformers 4.39
    - torchvision 0.16
    - scikit-learn 1.3
    - pandas 2.2, matplotlib 3.8
    - pycocoevalcap (for CIDEr evaluation)
- **Models:**
    - BLIP (Bootstrapping Language-Image Pretraining)
- **Tools & Environment:**
    - Google Colab (with GPU acceleration)
    - Google Drive (for model checkpoints)
    - tqdm (training progress visualization)

---

2. Summary

This solution uses a fine-tuned version of the BLIP model for generating image captions. The pipeline involves data preprocessing with basic text cleaning and caption augmentation, followed by training on a custom dataset using mixed precision. Validation is performed using the CIDEr metric. The model is incrementally improved by re-including difficult samples identified during validation, and data augmentation is applied both to images and captions for better generalization.

---

3. Approach

The base model used is BLIP, a vision-language transformer pretrained on large-scale image-text pairs. I fine-tuned this model on the **N/A** dataset with several enhancements:

- **Pretrained Model:** BLIP was fine-tuned from HuggingFace checkpoints.
- **Training Strategy:**
    - Training was resumed from epoch 12 after earlier fine-tuning.
    - Automatic Mixed Precision (AMP) was utilized for faster and memory-efficient training.
    - Used label smoothing with CrossEntropyLoss to mitigate overfitting.
    - Cosine annealing scheduler was applied with a warm-up phase.
- **Data Augmentation:**
    - Image transforms: random rotation, cropping, horizontal flips, color jitter.
    - Caption transforms: synonym replacement for diversity.
- **Curriculum Learning:** Incorrect predictions on the validation set were re-added to the training set to focus on challenging samples.
- **Beam Search:** Used with beam size of 5 during generation for higher-quality captions.

---

4. Sample Outputs

Prediction: a vibrant sunset with hot air balloons floating above a rocky landscape, surrounded by rocks and distant hills.

Prediction: a young lady holds a camera close to her face, focused on the lens, ready to take a photo.

Prediction: a person stands on a ledge at the edge of a grand canyon, silhouetted against the landscape under a cloudy sky.

---

5. References

- **BLIP:** https://arxiv.org/abs/2201.12086
- **HuggingFace Transformers:** https://huggingface.co/transformers/
- **pycocoevalcap:** https://github.com/tylin/coco-caption
- **Google Colab:** https://colab.research.google.com/
- **ChatGPT (OpenAI, GPT-4.5):** Used for technical guidance, debugging, optimization suggestions, and report drafting assistance. https://chat.openai.com/