EEE 543 Preliminary Report

Bidirectional Vision–Language Networks for Cross-Modal Generation and Reconstruction

Group Members:

Yusuf Baran ATEŞ	21802263
İsmail Enes BÜLBÜL	22401367
Muhammet Melih ÇELİK	22003836

Project Description

This project investigates how neural networks can learn bidirectional mappings between visual data and natural-language descriptions. Following networks are:

i. **Image-to-Text Encoder** First generative model takes an image as input and outputs a detailed natural-language description.

ii. **Text-to-Image Decoder** Second generative model that takes the textual description generated by the first network as input and produces an image that matches the described content.

- The Historic Art dataset available on Kaggle is chosen .The dataset consists of 45k images.
- Since the original dataset does not include long-form textual descriptions, a subset of the images is augmented with generated captions to enable supervised learning. Each image is described with 100–120 words.
- All stages of the project are planned to be implemented; however, depending on time and computational constraints, some steps may not be finished

End-to-end architecture of the project can be seen in Figure 1.

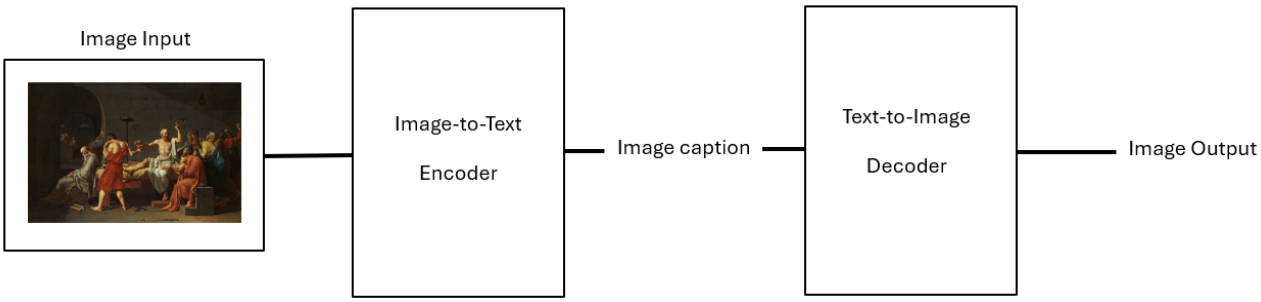

Figure 1: Project Architecture Diagram

Project Details

Step 1: Dataset Preparation

- 4,567 historical artwork images are randomly chosen. Each selected image is passed through GPT-4o-mini to generate a detailed natural-language description.
- Each caption is approximately 100–120 words, focusing on visual content such as composition, color usage, subjects, textures, and artistic style.
- These image–caption pairs form the supervised data used for training both models.The dataset is split into 80% training data and 20% testing data.

Step 2: Image-to-Prompt Network

- The first model learns to translate visual information into natural language.The input is an image from the Historic Art dataset.
- A vision encoder extracts high-level visual features.
- A language decoder generates a detailed textual description approximating the GPT-generated caption.
- This step evaluates how well visual semantics can be encoded into long-form language.

Step 3: Prompt-to-Image Network

- The second model learns to generate images from textual descriptions.The input is a vectorized representation of a 100–120 word prompt.
- A text encoder produces a latent embedding from the caption.
- An image decoder reconstructs an image conditioned on the text embedding.
- This step evaluates how effectively language can guide visual generation.

Step 4: Connecting the Two Networks

- After training both networks independently on the same dataset: An image from the test set is first passed through the Image-to-Prompt network to generate a caption.
- The generated caption is then passed into the Prompt-to-Image network.
- This forms an image $\rightarrow$ text $\rightarrow$ image translation pipeline.
- The reconstructed image is compared with the original image to assess semantic preservation.

Project Aim

For the evaluation, the architecture will be tested whether image-caption-image cycle preserves the semantic content and the visual structure of the original image.

It is expected that if proposed network architecture trained properly, **the reconstructed image should resemble the original image inserted to the architecture.** Results of the project would help to understand whether the generated text generated between independently trained vision-language networks is a good representation of the original image or not.