

Project Title: A Multimodal Deep Learning Framework for Image Captioning in Manuscript Collections


Project Description

This project investigates how **textual content in medieval manuscripts** can be computationally analyzed to generate **meaningful descriptions or captions** for their accompanying illustrations. The research goal is to explore the **thematic connection between manuscript text and imagery**, and to evaluate whether **automated summarization and keyword extraction techniques**, combined with **deep learning-based image captioning**, can support the creation of image metadata for manuscript studies.

Multimodal Deep Learning Approach

The project implements a **multimodal deep learning framework** that integrates **Computer Vision (CV)** and **Natural Language Processing (NLP)** to process manuscript images and their related texts simultaneously. The motivation behind this approach is that manuscript illustrations cannot be fully understood by relying solely on visual or textual information—**understanding both modalities provides a richer and more accurate interpretation**.

At its core, the framework builds an end-to-end pipeline that includes **data preprocessing, model design, training, and visualization**:

- **Image Understanding**
Manuscript images are processed using a **DenseNet201-based Convolutional Neural Network (CNN)**, which extracts high-level visual features. This CNN backbone acts as a feature extractor, capturing semantic and spatial details from raw images.
- **Text Understanding**
Manuscript transcriptions are pre-processed (cleaning, lowercasing, tokenization). Sequences are then embedded and passed through an **LSTM-based recurrent model**, which captures **contextual, grammatical, and thematic meaning** from the manuscript text.
- **Multimodal Fusion**
Extracted **image features** and **text features** are fused into a joint representation. This fusion enables the generation of captions for manuscript images that not only describe visual elements but also **retain the contextual meaning and historical significance reflected in the text**.
- **Training Optimization**
The training pipeline is optimized with modern deep learning techniques:
 -  **Adam optimizer** for adaptive learning rates.

- ☒ **EarlyStopping** to prevent overfitting.
 - ☒ **ModelCheckpoint** to save the best performing model.
 - ☒ **ReduceLROnPlateau** to dynamically lower the learning rate when training stagnates.
 - **Visualization & Analysis**

Model performance and learning dynamics are visualized using **Matplotlib** and **Seaborn**, providing insights into accuracy, loss trends, and training stability.
 - **Scalability**

The modular design allows easy extension: different CNN backbones (e.g., ResNet, VGG, DenseNet) or RNN variations (e.g., LSTM, Bidirectional LSTM) can be swapped in. This flexibility makes the system adaptable for experimentation and scaling to larger manuscript collections.
-