COSE474-2024F: Final Project Proposal
Enhancing CLIP Performance on Art Style Classification through Detailed Class Descriptions

Geonyeong Koh

1. Introduction

Motivation In the field of computer vision, image classification has achieved remarkable progress, yet the task of classifying artistic styles remains a significant challenge. This difficulty arises primarily from the ambiguous boundaries between styles. Many art movements have evolved through mutual influence, making it hard to define clear distinctions. Furthermore, the classification of a single artwork can vary depending on the viewer's perspective. Artistic style classification criteria extend beyond mere visual features to encompass historical and cultural contexts, which traditional computer vision models relying solely on visual features struggle to capture.

In this context, state-of-the-art multimodal models such as CLIP have demonstrated the potential to provide richer contextual understanding through textual descriptions. However, the zero-shot performance of CLIP in artistic style classification remains suboptimal. Thus, this study aims to explore methods to enhance the performance potential of CLIP in the classification of artistic styles.

Problem Definition This study addresses the following research questions to improve CLIP's classification performance for artistic styles:

1. Can replacing ambiguous style names with detailed descriptions improve classification accuracy?

2. Which types of artistic styles benefit the most from detailed descriptions?

Contribution This study systematically compares and analyzes the effectiveness of simple class names versus detailed class names in artistic style classification. It also examines the impact of detailed class names in both zero-shot learning and fine-tuning scenarios. Furthermore, the study investigates the varying effects of detailed class names depending on the characteristics of the artistic styles. Finally, it presents effective prompt design guidelines for artistic style classification.

2. Methods

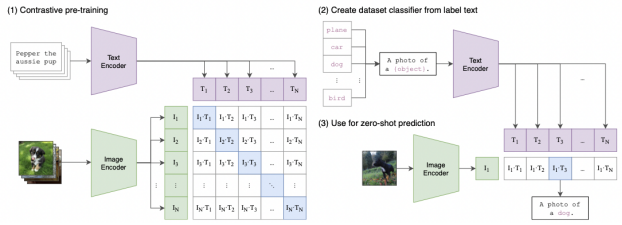


Figure 1. The architecture of CLIP

This section describes the specific methods employed in this project. This study is conducted using CLIP, and Figure 1 illustrates how CLIP operates.(1) The significance of this study lies in unlocking the potential of the CLIP model for artistic style classification. While previous studies did not modify the class names used in prompts input to CLIP's text encoder, this research attempts a novel approach by focusing specifically on the class names themselves.

2.1. Approach

In this study, we compare two different prompting approaches for the same artistic style classification task:

Basic prompt The prompt follows the conventional format used for writing prompts in CLIP: "An artwork in the style of {style_name} (e.g., "An artwork in the style of Baroque").

Detailed prompt The prompt uses detailed descriptions that include characteristic elements of each artistic style as the class name itself (e.g., "a Baroque painting with strong light-dark contrasts, rich colors, and dramatic compositions with diagonal lines"). The detailed descriptions for each class were generated considering the visual features and historical background of the styles. The detailed class names used in the experiments are available in the implementation code.

2.2. Zero-shot Classification

We compare the performance difference between Basic and Detailed prompts through CLIP's zero-shot capabilities. The confusion matrices are used to analyze class-specific differ-

ences between the two approaches.

Algorithm 1 Zero-shot CLIP Style Classification

Input: Image I, Set of basic/detailed prompts P
Output: Art style classification result
$f_{img} \leftarrow \text{CLIP}_{image}(I)$
$f_{txt} \leftarrow \text{CLIP}_{text}(P)$
$f_{img} \leftarrow f_{img}/\|f_{img}\|$
$f_{txt} \leftarrow f_{txt}/\|f_{txt}\|$
$s \leftarrow f_{img} \cdot f_{txt}^T$
Return: $argmax(s)$

2.3. Fine-tuning Approach

To enhance CLIP's performance in art style classification, we implement a fine-tuning strategy and compare the performance differences between basic and detailed prompts throughout the training process. Our fine-tuning approach modifies the pre-trained CLIP model while maintaining its core architecture.

Algorithm 2 CLIP Fine-tuning for Style Classification

Input: Dataset, labels, pre-trained CLIP model, optimizer, loss function
Output: Fine-tuned model
Load pre-trained CLIP model and freeze initial layers
for epoch in training epochs **do**
  for batch of (images, labels) **do**
    Compute image and text features: $f_{img}, f_{txt} \leftarrow \text{CLIP}(I, P)$
    Compute similarity scores: $s \leftarrow f_{img} \cdot f_{txt}^T$
    Calculate loss: $\mathcal{L} \leftarrow \text{CrossEntropyLoss}(s, \text{labels})$
    Update weights using optimizer
  end for
end for
Return: Fine-tuned model

2.4. Confusion Matrix Analysis

To analyze the effectiveness of our approach in detail, we utilize normalized confusion matrices for both basic prompts and detailed prompts. The confusion matrix provides insights into not only the overall accuracy but also the specific patterns of correct classifications and misclassifications across different art styles.

2.4.1. Normalization Method

We normalize each element of the confusion matrix by dividing the number of predictions for each class by the total number of samples in the corresponding true class. This normalization accounts for class imbalance in the test dataset and allows for fair comparison between classes of different sizes. Formally, for each element (i,j) in the confusion matrix:

$$\hat{CM}(i,j) = \frac{CM(i,j)}{\sum_k CM(i,k)} \tag{1}$$

where $CM(i,j)$ represents the number of images from class i that were predicted as class j, and $\sum_k CM(i,k)$ is the total number of images in class i.

3. Experiments

3.1. Dataset and Implementation

3.1.1. Dataset

We utilized the WikiArt dataset, which contains 80,020 unique images representing 27 different art styles.(2) We sampled 300 images per class for the training set, and this sampling strategy ensures balanced training. We test the model using the standard test set of 15,860 images provided by WikiArt to evaluate its performance.

3.1.2. Implementation Details

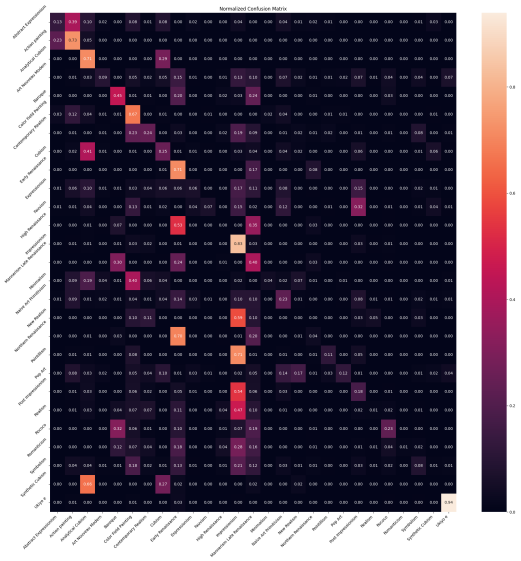
All experiments were conducted using Google Colab Pro with NVIDIA V100 GPU. We employed the CLIP ViT-B/32 model for both zero-shot testing and fine-tuning experiments. During training, we used a batch size of 32 and increased it to 1024 for testing to improve evaluation efficiency. The model was optimized using AdamW optimizer with a learning rate of 2e-6 and weight decay of 0.01.(3) Training continued for 25 epochs with early stopping monitored on validation loss. We implemented label smoothing (0.1) and gradient clipping (max norm 1.0) to stabilize training.

3.2. Results and Analysis

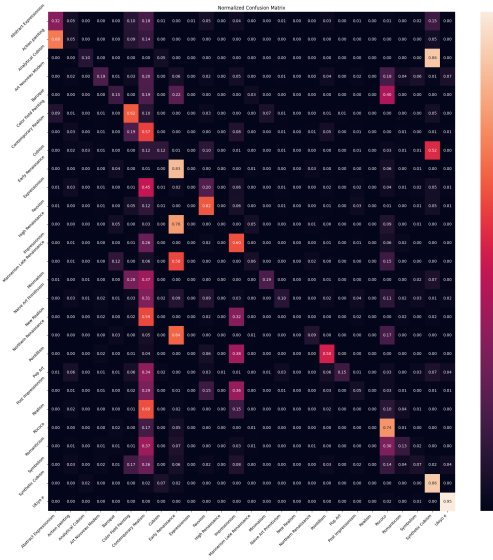
We compare our approach with two baseline models: a ResNet50 model achieving 50.1% accuracy, which serves as a traditional computer vision baseline, and EnAET, the current SOTA model reaching 82.61% accuracy through advanced self-supervised learning techniques.(4)

Table 1. 성능 비교 (% 정확도)

Method	Zero-shot	Fine-tuned
Basic Prompts	23.61	37.73
Detailed Prompts	26.67	41.03
ResNet50	-	50.10
EnAET (SOTA)	-	82.61



(a) Detailed Prompts



(b) Basic Prompts

Figure 2. Confusion matrices for art style classification using different prompt types

3.2.1. Quantitative Analysis

Zero-shot evaluation demonstrates the effectiveness of detailed descriptions, achieving 26.67% accuracy compared to 23.61% with basic prompts, showing a 3.06% improvement. This performance advantage is maintained in fine-tuning, where detailed descriptions reach 41.03% accuracy versus 37.73% with basic prompts, representing a 3.3% improvement.

Analysis of the confusion matrices reveals significant improvements in class-specific accuracy. In particular, Action painting shows a substantial increase in diagonal elements from 0.73 to 0.86, while Synthetic Cubism improves from 0.66 to 0.86. Traditional styles such as Ukiyo-e maintain high accuracy across both approaches (0.94 and 0.95 respectively).

3.2.2. Qualitative Analysis

The confusion matrices reveal distinct patterns of success and failure across different art styles. Better performance is observed in styles with clear temporal characteristics (e.g., Baroque, Early Renaissance), distinct technical approaches (e.g., Action Painting), and strong regional identity (e.g., Ukiyo-e). In contrast, styles with more complex or abstract characteristics show limited improvement, suggesting inherent challenges in their classification.

3.3. Discussion

This study demonstrates the effectiveness of detailed class descriptions in art style classification. By providing clearer visual characteristics and historical-cultural context, detailed descriptions enhance the distinction between similar styles, leading to improved CLIP classification performance. This improvement is consistently observed in both zero-shot testing and fine-tuning scenarios.

The success can be attributed to several factors. First, detailed descriptions effectively capture the unique technical aspects of each style, particularly beneficial for styles with distinct methodological characteristics. Second, explicit descriptions of visual elements help reduce confusion between visually similar styles.

However, our approach shows limitations when compared to baseline models. The classification performance of CLIP remains insufficient for artistic style datasets, falling significantly below the SOTA performance of 82.61%. Additionally, we observed performance degradation in certain styles, particularly those with complex or abstract characteristics. This limitation might be attributed to the inherent challenge of capturing nuanced artistic elements through text descriptions, and the potential semantic gap between visual features and textual descriptions in more abstract art forms.

4. Future Directions

Future research should focus on optimizing prompt engineering strategies through investigating optimal description lengths and developing efficient methods for generating effective detailed class descriptions. Furthermore, extending this approach to other domains with subjective classification challenges could validate its broader applicability. The potential integration with other techniques might also help improve the current performance.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, et al., "Learning transferable visual models from natural language supervision," in Proc. Int. Conf. Mach. Learn. (ICML), 2021, pp. 8748–8763.

[2] B. Saleh and A. Elgammal, "Large-scale classification of fine-art paintings: Learning the right metric on the right feature," arXiv preprint arXiv:1505.00855, 2015.

[3] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in Proc. Int. Conf. Learn. Representations (ICLR), 2019.

[4] A. Joshi, A. Agrawal, and S. Nair, "Art Style Classification with Self-Trained Ensemble of AutoEncoding Transformations," arXiv preprint arXiv:2012.03377, Dec. 2020.