COSE474-2024F: Final Project: Skin Cancer Classification with CLIP

Andi Analta Dwiyanto Palalangan Tunru [1]

1. Introduction

1.1. Motivation

Automated diagnosis tools can assist dermatologists in analyzing lesions efficiently and accurately. Multimodal models like CLIP (Radford et al., 2021) offer an opportunity to combine textual and visual modalities, enabling flexible and robust classification pipelines. One of CLIP's core strengths is it's generalization ability leverages extensive pretraining on diverse image-text pairs, allowing it to adapt to unseen data with minimal task-specific fine-tuning. This capability makes CLIP particularly well-suited for medical applications, where high variability in image characteristics and textual labels is common.

1.2. Problem Definition

The task involves classifying skin lesions into predefined diagnostic categories using dermoscopic images and their corresponding diagnostic labels. The primary challenge lies in handling the high variability in lesion appearances across categories while ensuring that the model generalizes effectively to unseen cases based solely on visual data and descriptive textual prompts derived from the diagnosis.

1.3. Concise description of contribution

1. Integrated CLIP for multimodal learning with custom preprocessing for textual descriptions and image augmentations.
2. Applied design experimentations, such as hyperparameter tuning , regularizations and prompt testing to achieve competitive performance.
3. Enhanced training data variability through augmentations like random cropping, horizontal flips to make the model more robust to augmentated data

2. Methods

2.1. Significance/Novelty

Leveraging CLIP, a pretrained multimodal model, reduces the need for task-specific pretraining and adapts it for medical image analysis.

2.2. Main Figure

The diagram below outlines the workflow of the classification pipeline designed for the skin lesion diagnosis.

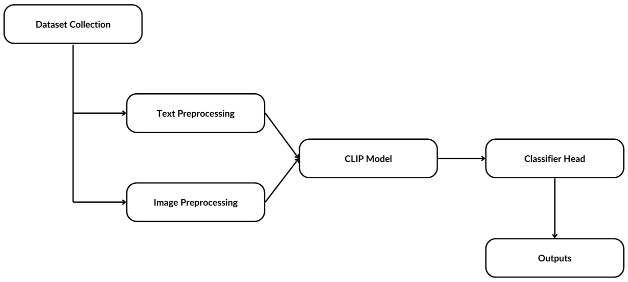

Figure 1. Main Figure

The figure represents the comprehensive workflow of the classification pipeline implemented for diagnosing skin lesions. The process begins with Dataset Collection, where image and text data related to skin lesions are gathered. This data is then bifurcated into two preprocessing pipelines—**Text Preprocessing** and Image Preprocessing.

1. Text Preprocessing: The textual descriptions of the lesions are tokenized and transformed into a numerical format suitable for the CLIP model. This ensures compatibility with the multimodal nature of CLIP, which processes both text and images simultaneously. 2. Image Preprocessing: The image data undergoes resizing, cropping, normalization, and other augmentation steps to ensure uniformity and enhance the model's generalization capability.

Both the preprocessed text and image data are passed into the CLIP Model, which generates embeddings that align the textual and visual representations in a shared latent space. These embeddings are then fed into a Classifier Head, a fully connected neural network designed to map the embeddings to specific skin lesion classes.

Finally, the Outputs of the classifier head represent the predicted skin lesion class, completing the diagnostic pipeline.

This end-to-end architecture leverages the multimodal capabilities of the CLIP model, combining textual and visual cues for enhanced diagnostic accuracy.

2.3. Pseudocode

```
Initialize dataset with images and diagnostic labels (dx)
Define train and validation augmentations
For each epoch:
    For each batch in train_dataloader:
        Preprocess images with augmentations
        Convert dx to textual prompts
        Tokenize text inputs and process images
        Forward pass through CLIP model
        Compute logits using the classifier head
        Backpropagate loss and update model weights
    Validate on val_dataloader
```

3. Experiments

3.1. Dataset

The dataset utilized in this study is the HAM10000 (Human Against Machine with 10000 Dermatoscopic Images) dataset (Tschandl, 2018), a comprehensive and widely recognized dataset for skin lesion classification tasks. It is publicly available through platforms like Hugging Face and has been extensively used in dermatological research to benchmark machine learning models.

HAM10000 consists of 13,354 high-quality dermatoscopic images, collected from multiple international clinics and spanning a diverse range of patient demographics. These images represent seven diagnostic categories of skin lesions, including both benign and malignant conditions:

1. Melanocytic nevi: The most common category, accounting for a majority of the dataset.
2. Melanoma: A malignant category critical for early detection due to its potentially life-threatening nature.
3. Benign keratosis-like lesions: Including seborrheic keratoses and solar lentigines.
4. Basal cell carcinoma: A form of skin cancer that is generally less aggressive but still significant for diagnosis.
5. Actinic keratoses and intraepithelial carcinoma: Precancerous lesions that can develop into squamous cell carcinoma.
6. Vascular lesions: Including angiomas, angiokeratomas, and pyogenic granulomas.
7. Dermatofibroma: Benign lesions that may be mistaken for more serious conditions.

Each image is annotated with its corresponding diagnostic condition (dx) and additional metadata, such as patient age, sex, and lesion localization, which can provide valuable context for model training.

The dataset was divided into three subsets:

1. Training Set: 9,577 images
2. Validation Set: 2,492 images
3. Test Set: 1,285 images

3.2. Computing Resource

The research was conducted in a Google Colab environment, utilizing widely used frameworks and tools for model development and evaluation. The stack included the following:

- GPU: NVIDIA Tesla T4 (16GB).
- CPU: Intel Xeon.
- Frameworks: PyTorch 2.0, Hugging Face Transformers.
- OS: Ubuntu 20.04.

3.3. Experimental Design/setup

In order to find an analyze the models performance on the dataset, a few different setups were used to see the impact of how the hyperparameters and setups would affect the performance of the model

- Base Setup
  For the base setup, the original dataset was used without augmenting any of the data, using the following hyperparameters
  1. Learning Rate: 5e-5 (with scheduler).
  2. Batch Size: 16.
  3. Epochs: 20.
- Regularizations
  For this setup, regularization techniques were used to see how it affects the models performance. All other hyperparameters are the same with previous experiment setup
  1. Weight Decay : 0.01

- Augmented Dataset
  In order to increase the robustness of the model, I have augmented the dataset and included it as part of the training dataset, the augmentation steps that are done to the images are below

  1. Resize((256, 256)): Resizes the input images to size of 256x256 pixels.
  2. RandomCrop((224, 224)): Crops a random region of size 224x224 pixels from the resized image.
  3. RandomHorizontalFlip(p=0.5): Flips the image horizontally with a 50% probability

3.4. Qualitative Results

- The class with the highest accuracy is melatocytic nevi, which is the class with the most data available

- The model could not classify any data correctly from dermatofibroma class

- Predicted labels match well with ground truth for most classes.

3.5. Quantitative Results

Below are the results of the accuracy from the experiments

	Training	Validation
Base	83.5%	81.8%
Regularization	81.2%	80.4%
Augmented Data	78.8%	79.1%

Table 1. Results Table for Experiments

The base model achieved the highest accuracy, with 83.5% on the training set and 81.8% on the validation set, indicating that it was able to learn the patterns in the data effectively without significant overfitting. When I applied regularization, I observed a slight decrease in performance, with the training accuracy dropping to 81.2% and the validation accuracy to 80.4%. This suggests that regularization helped in reducing overfitting, but it came at the cost of a small reduction in overall performance. In the case of the augmented data model, I noticed a further drop in training accuracy (78.8%), but an improvement in validation accuracy (79.1%). This result indicates that while the augmented data approach helped the model generalize better to unseen data, it reduced the model's ability to fit the training data as efficiently. Overall, these results highlight the trade-offs between enhancing model generalization and maintaining training accuracy, offering insights into how different techniques affect model performance.

3.6. Figures and Analysis

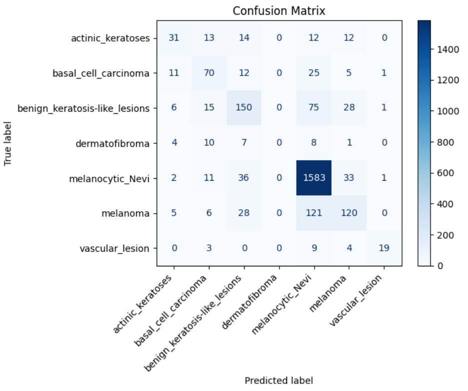

Figure 2. Confusion Matrix

Melanocytic nevi show the highest accuracy, with 1583 samples correctly classified and very few misclassifications. However, for other classes like actinic keratoses and basal cell carcinoma, the misclassification rates are higher. For example, actinic keratoses were often misclassified as benign keratosis-like lesions or melanoma. Similarly, benign keratosis-like lesions show significant confusion with melanoma and actinic keratoses. Notably, dermatofibroma and vascular lesion have low correct classification rates, likely due to their limited sample size and visual similarities with other classes.

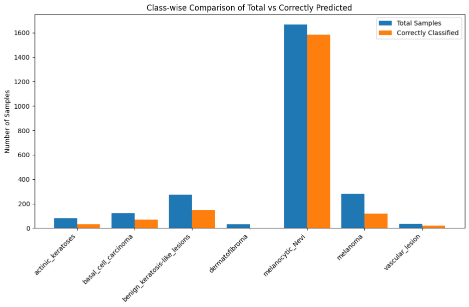

Figure 3. Class wise comparison for classes

the class-wise comparison between the total samples and correctly classified samples highlights key insights into the performance of our classification model. The largest category, melanocytic nevi, contains over 1600 samples, with the model achieving high accuracy by correctly classifying the vast majority of these cases. For benign keratosis-like lesions and melanoma, which each have around 300 samples, the model shows moderate success, indicating potential areas for further improvement. Smaller categories such as actinic keratoses, basal cell carcinoma, dermatofibroma, and vascular lesion demonstrate lower sample sizes and varying degrees of correct classification, with dermatofibroma and vascular lesion showing particularly low accuracy. This suggests that more training data or enhanced model techniques may be required to improve performance for these underrepresented classes. Overall, the findings emphasize that while the model performs well for more common categories, there are challenges in accurately identifying rarer skin lesion types.

4. Future Direction

- Dataset Expansion: Include additional dermoscopic datasets to improve class diversity and generalization.
- Advanced Prompt Engineering: Explore context-aware prompts like ”The lesion shows signs of label”.
- : Model Enhancements: Incorporate domain-specific pretrained encoders for images (e.g., models fine-tuned on medical datasets).
- Explainability: Incorporate attention maps to provide interpretability for predictions, aiding clinical adoption.

References

Radford, A., Kim, J., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., and et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Tschandl, P. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions, 2018. URL https://doi.org/10.7910/DVN/DBW86T.