AI6126 Project 1

Fashion Attributes Classification Challenge

Kozhimadam Shahin Shah, G2303660L

1 Introduction

Tackling the nuanced problem of multi-label multi-class classification presents a unique challenge, especially within the domain of digital fashion imagery. This endeavor seeks to develop an algorithm capable of accurately identifying and classifying multiple attribute labels from a single image, navigating through a dataset of 6000 images split into 5000 for training and 1000 for validation, with further evaluation on an online-hosted test set. Each image is annotated with six attributes across 26 possible labels, detailing common garment descriptions and grouping them into six major categories. This setup not only demands the identification of various attributes within an image but also requires the precise classification of each attribute into its respective category. Such a task is compounded by the complex interaction of attributes within the images, challenging the model to maintain accuracy and specificity across a diverse array of labels.

2 Methodology

2.1 Model Architecture

This work introduces an adaptation of the Query2Label model [2], extending its utility from multi-label to multi-label multi-class classification tasks. The model integrates a Transformer-based model [5] with a Swin Transformer Large (Swin-L) [3] for feature extraction. The Swin-L, a pre-trained model on the ImageNet-1K dataset [4], is selected for its robust ability to capture detailed spatial hierarchies and features. This capability is crucial for processing the complex and varied images typically encountered in this classification task. The spatial features extracted are then adapted through a Conv2D projection layer to meet the Transformer's input specifications. Moreover, 2D positional encoding is employed to maintain spatial context, crucial for accurate image content interpretation.

The architecture is further enhanced by incorporating learnable query embeddings for each category, enabling the model to focus on relevant features for precise attribute recognition within categories, thus significantly improving its classification performance. Alongside, the model features a custom Transformer block, composed of an encoder and two decoder layers, that refines the query embeddings through the process of cross-attention with the encoded spatial features. This cross-attention mechanism allows the model to selectively concentrate on specific parts of the image that

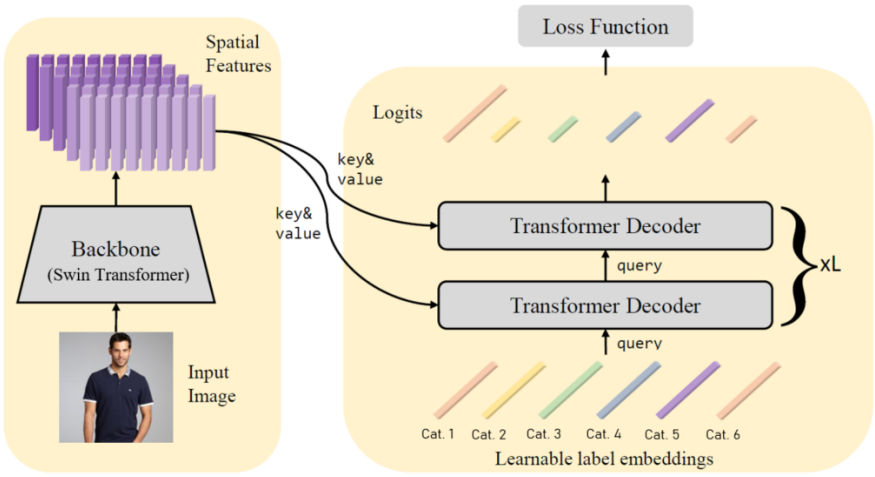

Figure 1: Architecture of the adapted multi-label multi-class classification model, incorporating a Swin Transformer Large (Swin-L) as the feature extraction backbone, and a custom Transformer block for precise attribute recognition across multiple categories. This figure is based on and modified from [2].

are most relevant to each category. The Transformer block is specifically designed with a hidden dimension of 768, a feed-forward dimension of 3072, and four attention heads, making it well-suited to handle the intricate details of the images. To address overfitting, a common issue given the dataset's size, a dropout rate of 0.3 has been introduced, providing a necessary regularization effect to enhance the model's ability to generalize.

Classification is achieved through a sequence of linear projection layers, each mapped to a specific category. These layers translate the Transformer block's output into class numbers for each category, effectively facilitating multi-label multi-class classification. The model outputs a set of logits for each category, indicative of the presence of particular class attributes. Illustrated in Figure 1, the architecture demonstrates the seamless integration of the Swin Transformer backbone with the Transformer block and classification layers, collectively geared towards navigating the intricacies of multi-label multi-class classification.

2.2 Loss Function

To address class imbalance within the dataset for the classification task, an evaluation of loss functions was conducted, focusing on Categorical Cross-Entropy with Class Weights and Focal Loss [1]. Applying class weights to Categorical Cross-Entropy adjusts misclassification penalties based on class rarity, aiming to enhance the model's sensitivity to under-represented classes for a more equitable learning process.

Despite the potential of Focal Loss to focus on hard-to-classify examples by modifying the loss contribution from well-classified instances, Categorical Cross-Entropy with Class Weights was determined to be more suitable for this model and dataset after extensive testing. It demonstrated superior performance on the validation set, indicating the importance of adjusting class weights to improve model accuracy and handle class imbalances effectively in multi-label multi-class classification tasks.

2.3 Data Processing

The challenge of working with a limited training dataset necessitates strategic measures to enhance model generalization and mitigate the risk of overfitting. To this end, a comprehensive approach to data augmentation was employed as a critical component of the data processing pipeline. The experiments encompassed a variety of augmentation methods, including random horizontal flips, random crops, random rotations, color jitter, MixUp [7], and CutMix [6].

After rigorous experimentation to ascertain the optimal combination of augmentations for improving model performance while avoiding overfitting, it was determined that the best results were achieved utilizing only random horizontal flips and random crops. This pared-down approach effectively increased the dataset's variance without introducing excessive noise, leading to better generalization on unseen data. Additionally, to ensure a consistent input distribution and facilitate model training, all images were normalized using the mean and standard deviation values of the ImageNet-1k dataset.

3 Training

3.1 Training Procedure

The model underwent training for a total of 20 epochs, with a uniform batch size of 64 applied across both the training and testing phases. This batch size was selected to achieve an optimal equilibrium between computational efficiency and the effectiveness of gradient updates during the training process.

Optimization was facilitated through the Adam optimizer, configured with a learning rate of 1×10^{-4}. This learning rate was carefully chosen to encourage steady progress towards convergence, avoiding the risks associated with overly aggressive learning rates. The optimizer's parameters were set with beta values of 0.9 and 0.999 for the first and second moment estimates, respectively. Additionally, a weight decay of 1×10^{-3} was employed as a regularization measure to prevent overfitting, and an epsilon value of 1×10^{-8} was used to ensure numerical stability during optimization. A cosine learning rate scheduler was implemented, devoid of warmup steps, to allow for an immediate commencement at the specified learning rate. This scheduler is designed to reduce the learning rate following a cosine curve across the epochs, aiding in the refinement of the model's weights through a gradual and controlled optimization trajectory. The model's performance was assessed using mean class accuracy to gauge its consistent classification ability across all classes.

3.2 Training Machine Specifications

Training was conducted on a high-performance setup featuring two NVIDIA RTX 6000 Ada Generation GPUs, each with 48GB of memory. This configuration was selected for its advanced computational capabilities and ample memory, essential for efficiently managing the demands of the training process. The dual GPU setup significantly expedited training times and supported the utilization of larger batch sizes, enabling more extensive experimentation with model configurations and hyperparameters, thereby enhancing the overall model training and development process.

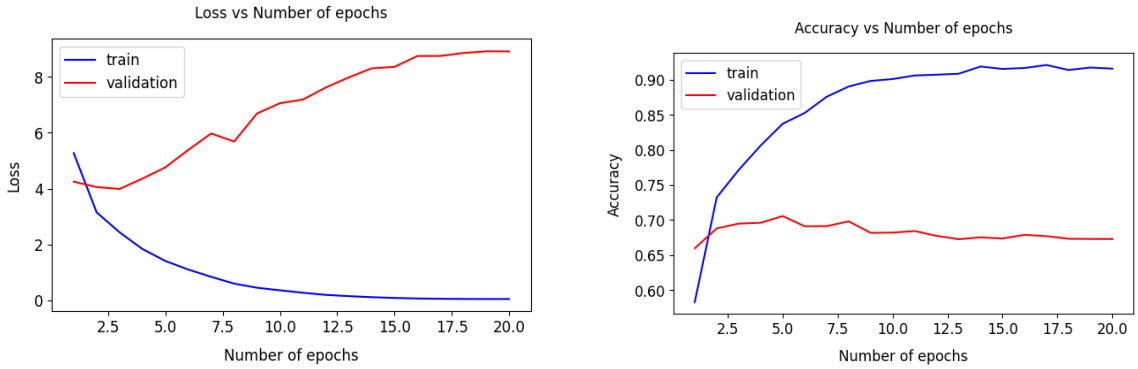

Figure 2: Training and validation dynamics over 20 epochs. (a) shows the loss on the training and validation sets, indicating convergence of the model. (b) presents the accuracy, illustrating the model's learning progression. The divergence between training and validation curves suggests overfitting.

4 Results

4.1 Training Curves

The training curves, depicted in Figure 2, present a nuanced picture of the model's learning behavior over 20 epochs. The loss and accuracy trends during training highlight a clear case of overfitting. The training loss decreases significantly and the accuracy increases, while the validation loss plateaus and the validation accuracy remains relatively stagnant after initial improvements. This divergence between training and validation performance metrics is a classic indication of the model's over-specialization to the training data.

In response to the overfitting observed, various regularization techniques were employed. Despite these efforts, there was a noticeable decrease in validation accuracy, suggesting that the measures to prevent overfitting might have constrained the model's ability to generalize. Ultimately, the model displaying overfitting was chosen for its superior validation accuracy. The rationale behind this selection lies in the pragmatic approach towards model performance, accepting a certain level of overfitting to achieve better generalization on unseen data, which is a critical factor in practical deployments of machine learning models.

4.2 Model Parameters

The architecture employs a total of 217.5 million parameters, a testament to its capacity for learning and adaptation. A substantial share of the parameters, amounting to 195 million, originates from the Swin Transformer Large (Swin-L) backbone, which forms the crux of the feature extraction process.

4.3 Test Dataset Results

Upon evaluation on the test dataset, the model demonstrated a mean class accuracy of 72.66%, showcasing a performance level that is comparable to the validation

results. This consistency reinforces the model's generalization capabilities across unseen data.

5 Discussion

In this study, the strategic employment of Swin Transformer Large (Swin-L) as the model's backbone, pretrained on the ImageNet-1k dataset, emerged as a foundational strength. The immediate benefit of pretraining was visible in the model's remarkable validation accuracy right from the first epoch, setting a solid groundwork for subsequent learning. This advantageous starting point was crucial for the task at hand, enabling the model to extract more representative features conducive to multi-label classification.

A key innovation in the model's design was the implementation of cross-attention between label embeddings and encoded features, which significantly enhanced its ability to concentrate on relevant portions of the image. This approach starkly contrasts with global average pooling, which could potentially dilute the distinct features pertinent to each category. The choice of Categorical Cross-Entropy with Class Weights as the loss function played a pivotal role in navigating the challenges of class imbalance. This strategic selection underscored the model's capacity to maintain a balanced approach to learning across diverse classes, contributing significantly to the nuanced understanding required for the task at hand.

The exploration of data augmentation techniques underscored a preference for simplicity, with basic methods like random horizontal flips and crops outperforming more complex strategies such as Mixup and CutMix. The latter, while theoretically beneficial for introducing variation and robustness, appeared to inject noise, thereby complicating the model's training process. Nonetheless, the model encountered overfitting, a challenge attributed to the limitations of simpler augmentation methods and the dataset's size. Despite efforts to mitigate this through various regularization strategies, the decision to accept a model exhibiting some overfitting was based on its superior validation performance. This pragmatic choice highlights the delicate balance between achieving theoretical model ideals and practical efficacy, with an emphasis on maintaining high accuracy in real-world applications.

6 Conclusion

This report outlines the journey and outcomes of tackling a sophisticated multi-label multi-class classification problem within the fashion domain. By harnessing the Query2-Label model enhanced with a Swin Transformer Large (Swin-L) backbone, the project achieved notable success in attribute identification and classification. Despite facing challenges like class imbalance and overfitting, strategic data augmentation, and a nuanced approach to loss function application allowed for considerable model performance, demonstrating strong generalization capabilities on unseen test data. The success of this approach demonstrates the utility and potential of sophisticated machine learning methods for tackling multi-label multi-class classification problems, setting a precedent for future investigations aimed at unraveling similar complex classification tasks across diverse domains.

References

[1] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[2] Shilong Liu, Lei Zhang, Xiao Yang, Hang Su, and Jun Zhu. Query2label: A simple transformer way to multi-label classification. *arXiv preprint arXiv:2107.10834*, 2021.

[3] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[4] Microsoft. Swin transformer large (swin-l) model pre-trained on imagenet-1k. `https://huggingface.co/microsoft/swin-large-patch4-window7-224`, 2021. Accessed: 2024-03-20.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[6] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[7] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.