Final Project
Proposal Document

Movie Genre Classification from Posters using Vision Transformers

Mingu Hwang
mingu.hwang@stonybrook.edu

Sihoon Sung
sihoon.sung@stonybrook.edu

Title

Movie Genre Classification from Posters using Vision Transformers

Project Vision and Objectives

A movie poster is not a random promotional photo, but an important photo that visually conveys the genre and feeling of a movie. However, it is difficult for most viewers to understand a specific genre of a movie by only looking at the poster. In fact, I once went to the theater and chose a movie only after seeing the movie poster, and then chose a horror movie that I was afraid of. When I looked at the poster again after watching the movie, the overall atmosphere was scary, but when I chose the movie, I did not feel this part properly because I was a person. I was curious about what the results would be if machines, not humans, ruled out emotions and judged them, so I decided to create a model for predicting the genre of the movie by looking at the pictures on the movie poster.

In this project, we will first develop a movie poster genre classification model based on CNN and Vision Transformer (ViT), compare the experimental results, and then find out which model is more accurate and efficient. According to our plan, uploading movie posters through the classification model will allow us to know the genre of the movie. This model will make it easier for audiences to choose a genre that suits their taste when choosing a movie, and reduce the difficulties caused by the unexpected genre I experienced.

Through this CSE327 course, the ultimate goal of the project is to develop a movie recommendation program. For this project, we propose to build a model for predicting genres by analyzing only the visual elements of movie posters, and develop a recommendation system based on them. At the end of the spring semester in 2025, we will develop this model and create a movie recommendation program based on the atmosphere of movie posters. This is the biggest difference from the existing movie recommendation model based on text data such as plot and lead actors.

In this project, which we start with great interest, rather than developing new models, we aim to implement and compare existing CNNs and Vision Transformer (ViT) models in movie poster genre classification. Our approach focuses on classifying genres by analyzing visual elements of movie posters, providing a unique perspective when compared to existing recommendation systems that rely on textual information. Furthermore, we will create a model that can organize the photographer's characteristics by categorizing not only movie posters but also any photo's atmosphere into an objective analysis.

This is our vision and the end of the project we started in this CSE327 class.

Background and Related Work

Movie posters are an important marketing tool that visually represents the mood and story of a movie. We are going to use these posters to create a model that categorizes the genres of movies. We tried to find data that had previously been conducted to infer genres through visual elements of movies. There were few studies involved, but we were able to find studies that were still relevant(Samuel Sung, Rahul Chokshi, https://cs230.stanford.edu/projects_winter_2020/reports/32643471.pdf). The Convolutional Neural Network (CNN)-based research had a lot of references in our project. Using CNN, it extracts visual features and uses them to classify genres of movies. Based on this study, we confirmed that it is possible to classify genres by analyzing movie posters.

The Effect of Movie Poster Characteristics on the Box Office Performance: Analysis Using an Image Deep Learning Model paper, which is listed on kci.go.kr , analyzed 19,551 movie poster images using deep learning techniques to study the effect of extracted genre scores on a movie's box office performance. Through this study, we can learn how to analyze visual elements.

Vision Transformer (ViT), which has recently applied Transformer architecture to computer vision, has attracted attention. ViT has shown superior performance in image classification tasks by segmenting images into patch units and inputting them into Transformer models. Specifically, ViT is a way to improve generalization performance through prior learning on large datasets. We intend to develop a model that utilizes ViT to classify the genre of movies only with movie poster images(Utsav Kumar Nareti .et al, 2023, https://arxiv.org/abs/2309.12022). It is differentiated from existing text-based or CNN-based research and aims to achieve high-accuracy genre classification with only visual elements by utilizing the strong expressive power of Transformer architecture.

Dataset Selection and Justification

CNN model

1. Dataset of 36,898 movie posters collected by Kaggle(contain movie titles, genres, IMDb scores, and movie poster URLs)
2. Resize the poster image of 256 x 256 x 3, genre converted to multi-hot vector format
3. 6 Major genres with more than 4,500 samples
4. Data divided into training(80%) - validation(10%) - test(10%)

ViT model

1. Dataset source from IMDb's non-commercial datasets(13,882 movie posters labeled with 13 genres, multi-label)
2. Each poster has up to three genre labels, encoded using multi-hot encoding.
3. Select movies released after 2000, movies with over 10,000 user votes, and runtime of at least 60 minutes.
4. Split the dataset into 3 parts. 80% Training(10,942 posters), 10% Validation(1,470 posters), 10% Test(1,470 posters).
5. Preprocess the dataset. Uniformly resize the image, use multi-hot encoding, and extract features by using ResNet50V2-based convolutional layer to extract deep feature embeddings instead of raw images.

Methodology

CNN model

1. Model Selection & Modification
   - Implemented ResNet-50, VGG-16, and DenseNet-169 with modified fully connected layers for multi-label classification
2. Loss Function & Optimization
   - Used Binary Cross-Entropy Loss with class weighting to address dataset imbalance. Adam optimizer applied for training
3. Training Process
   - Trained models with mini-batch size of 32 in ResNet-50(20 epochs), and VGG-16 & DenseNet-169(10 epochs).
4. Evaluation Metrics
   - Assessed performance using F1-score and AUC-ROC instead of accuracy due to multi-label nature.
5. DenseNet-169 outperformed other models, achieving the highest F1-score(0.77) and AUC(0.67), showing the best classification performance.

ViT model

1. Problem Formulation
   - Multi-label classification where each movie poster can belong to multiple genres, represented using multi-hot encoding
2. Model Architecture
   - Residual Dense Transformer (RDT) -> Uses transformer encoders with multi-head self-attention (MSA) and deep feature embeddings for genre classification
   - Ensemble Model (ERDT) -> Combines Residual Network (R), Residual Transformer (RT), and RDT for enhanced performance
   - Probabilistic Module (PrERDT) -> Filters out irrelevant genres by estimating conditional probabilities
3. Loss Function & Optimization
   - Uses Asymmetric Loss Function (ASL) to handle class imbalance
   - Trained with Adam optimizer and early stopping (10 epochs).
4. Training Details
   - Batch size: 32, 4 transformer layers, 6 attention heads, 256 embedding dimensions.
   - Decision thresholds $\tau = 0.3$, $\tau' = 0.03$ in the probabilistic module
5. Evaluation Metrics
   - Assessed using Precision, Recall, Specificity, Balanced Accuracy (BA), F1-score (FM), and Hamming Loss (HL).

Expected Outcomes

The project's purpose is to combine performances of CNN and ViT models on multi-label movie genre classification solely from poster images. The baseline would be the DenseNet-169 CNN-based model with an F1 score of 0.77 and an AUC-ROC of 0.67. To better this score, the ViT model must, therefore, meet or exceed an F1 score of 0.78 and an AUC-ROC above 0.68. The detailed analysis of both models will include comparisons of Precision, Recall, Specificity, Balanced Accuracy, F1 score, and Hamming Loss.

ViTs instead forward the use of transformer encoders and self-attention to capture global relationships within images rather than using CNNs on local spatial structures. Such structural disparity would include, in essence, that the ViT model, particularly ensemble-based Eurdra with Pruning or Probabilistic Filtering (PrERDT), will yield results more reliable for predictions in multi-label classification tasks.

Further analysis of genre-wise performances shall also establish which model has more ability to differentiate between visually close genres. In the end, if ViT is better than CNN, it certifies transformers as the superior model in the vision-based classification arena, especially with respect to movie poster-driven genre tagging. On the other hand, though, If CNN performs equally or even better, then it is easy for CNN to extract features relevant to the genre from images as against the complexity posed by the transformers. These results may help in the building of more sophisticated movie recommender systems based on visual cues rather than traditional metadata systems.

Timeline and Milestones

Week	Task Description
Week 1	Data Research & Data Preprocessing and Experimental Setup for CNN model
Weeks 2-4	CNN Model Implementation and Initial Experiments
Week 5	Data Research & Data Preprocessing and Experimental Setup for ViT model
Weeks 6-8	ViT Model Implementation and Initial Experiments
Week 9	Comparison and Performance Evaluation of CNN and ViT models
Week 10	Result Analysis and Report Writing