Final Project
Project Progress Report
#2

Movie Genre Classification from Posters using Vision Transformers

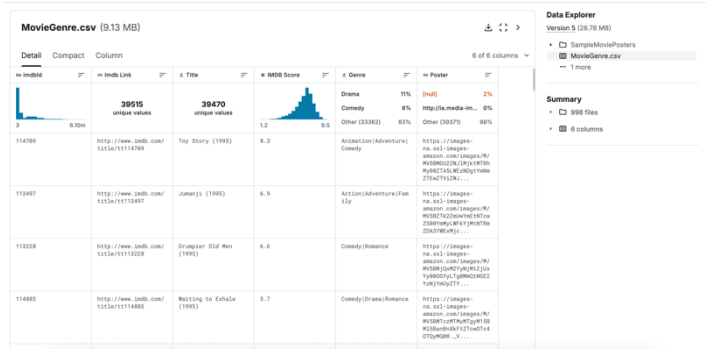
Mingu Hwang
mingu.hwang@stonybrook.edu

Sihoon Sung
sihoon.sung@stonybrook.edu

Revise Previous Report

In previous reports, the CIFAR-10/100 dataset was used for the initial experiments of the Spike-Driven Transformer.

The CIFAR dataset was suitable for validating model structure and learning pipelines because it is used by default in thesis and GitHub codes. It was also the most accessible dataset. However, our project goal is genre classification using movie posters, so we needed a movie poster dataset to see how it works in a real-world scenario, not just during model validation, and we found a movie poster dataset.



At first, we manually divided the task between two team members and tried to collect and label movie poster images and their genres ourselves. However, this process took too much time and effort due to the large number of movies, so we eventually had to give up that method. Instead, we decided to use a publicly available dataset from Kaggle: Movie Genre from its Poster. This dataset contains high-resolution poster images along with genre labels, and it fits well with our project's objective of genre classification based on visual content.

Accordingly, from this report, efforts have been made to confirm the performance of each model through the actual movie poster image. We focused on matching the project's core goal with the direction of the experiment, and as a result, we were able to come up with a more concrete plan.

In the future, we will also work on converting input data in a way suitable for the Spiking Neural Network and Vision Transformer structure.

Feature Extraction

In order to design a genre classification model based on movie posters using the Spike-driven Transformer, feature extraction using CNN is first performed. According to a paper on movie genre classification research by Stanford University, when a poster image is input using a pre-trained CNN model such as DenseNet-169, a feature vector that visually summarizes the image is output. This vector is a high-dimensional vector form of [1, 1664], and it abstracts visual features such as color, person arrangement, atmosphere, and background included in the poster. In this paper, genres are classified directly through classifier, and if this happens, there is

no point in creating an event-driven dataset. Therefore, we convert the feature vector into a spike form and process it along the time axis, rather than writing it directly for classification.

The method of threshold encoding will be used as the spiking encoding method, which converts the feature value to 1 if it exceeds the value of threshold, or 0 otherwise. The converted Spike sequence becomes a binary sequence consisting of 0 and 1 in the form of [time_steps, feature_dim], and the spike-driven transformer receives this sequence as input and learns how neurons react over time. The important thing at this time is that even if the data are simplified to 0 and 1, the information represented by the pattern is maintained. The model is provided with sufficient clues to distinguish genres through a combination of these patterns. In the Spike-driven Transformer, the model learns the relationship between spike patterns and genres by itself. We associate a spike sequence and a correct answer label, and the model learns about the genre classification by itself through iterative learning. As things stand now, feature extraction is being implemented and is expected to be completed within this week.

Vision Transformer (ViT)

ViT, unlike traditional convolutional models, works in an entirely different way. The ViT processes images in a manner that is more akin to natural language processing (NLP). It divides images into small fixed-size patches, treating each as a token in a sequence. Then, it flattens the patches, converting them into embedding vectors, which are passed through a Transformer encoder. Because of the potential loss of spatial information between patches, tokens are assigned positional encodings to denote their original locations in the image.

What makes this approach stand out is the use of self-attention. The model can analyze the entire image all at once, rather than focusing on regions, enhancing the understanding of complex relationships in visual data that CNNs could miss. This reasoning is especially important for genre classification—for instance, interpreting a movie poster's visual design.

In our project, it isn't simply a ViT that we are using. We are dealing with a Spike-Driven version of the Transformer, which means we had to reimplement some of the operations in SNN (Spiking Neural Networks) framework. For instance, instead of Softmax and multiplication with attention matrix float, the model does AND logic operations at the columns and addition at the rows-as well as Hadamard products. Attention in this form is called SDSA (Spike-Driven Self-Attention) and fundamentally removes the worst energy consuming sections of the attention transformers are built on.

While studying this model, one of the most intriguing aspects was how 'positional encodings' are dealt with in spikes, as well as how SDSA passes information without the most basic components such as Softmax. That was very challenging for us to grasp at the beginning, but it is what made us understand that attention does not need 'exact' operations to be performed; it just needs

structures that facilitate significant engagement between tokens. In general, this approach not only lowered the model's operational burden, but also enhanced its adaptability to event-driven systems which is critical for energy sensitive applications.

Expected Outcomes

Our studies to this point, together with the preliminary design of our pilot experiments, lead us to form a number of expectations regarding model performance, particularly on the balance of traditional CNNs and standard ViTs versus the spike-driven variant that we are currently developing.

As an initial guess, we expect reasonable performance from the CNN model on the movie poster classification task. The strength of CNNs derives from their capacity to extract local features such as edges, shapes, and textures which are definably constructive towards identifying visual patterns. Nevertheless, we believe that CNNs' reliance on local feature detection will hinder their ability to perform on more abstract images and visual patterns, such as in the case of movie posters. Take horror and thriller posters for instance. They may share a lot of color or visual tone, which could be difficult to capture through small-scale filters.

Self-attention mechanisms should help the Vision Transformer model capture long-range dependencies and complex relationships which span across the entire poster. Since every patch have access to every other patch in the image, ViT may better understand the relations of layout, composition and text placement given their importance in cue reasoning for genre classification. This is the reason why we expect the ViT to outperform the other models in classification accuracy.

To start with the expectations we have with the Spike-Driven Transformer, they are two-fold. One, with his model we expect to achieve at least similar classification accuracy as the standard ViT, if not slightly better, owing to the fact that it is based on a more efficient energy processing model. The other, and probably more interesting expectation, is that we anticipate significantly lower computational overhead. Because the SDSA (Spike-Driven Self-Attention) uses addition and logic gates in place of Softmax and other heavy multiplication operations, it is expected that far less energy and computation power will be required. This could be very beneficial in practical deployment scenarios, particularly for edge devices or resource constrained environments.

To summarize, the awaiting the most is whether the Spike-Driven Transformer can achieve the desired performance and efficiency ratio. A claim we would like to validate is that energy-aware models can maintain the needed power for vision tasks in real-time, such as genre classification from the poster design.

Challenges and Unsolved Issues

From the onset, there were a few critical difficulties we encountered while working on this project. One of the initial concerns pertained to the dataset. At first, we assumed it was feasible to construct a dataset from scratch. We began splitting tasks, and one of us tried to gather genre-labeled movie posters images from the internet. However, we reached a point where it became clear that the effort far exceeded the reward. The number of movie posters available as well as the accuracy of genre labels was far too overwhelming. Ultimately, we came up with a plan to fetch the dataset from Kaggle, a publicly available dataset repository. Although this provided us with a structured dataset, it also consumed a fair share of our time. In addition, the rate at which we could control the balance and quality of the data was greatly reduced due to the publicly available dataset.

One more major hurdle was preparing the data for the Spike-Driven Transformer model. Since this model did not accept images in their raw form, we had to devise a method of extracting features first, usually done using a CNN, and then transform those features into spike-based representations. Determining how to accomplish this took considerable effort. We had to figure not only how the spike encoding was done but also how to ensure the features we provide to the transformer were meaningful and well prepared.

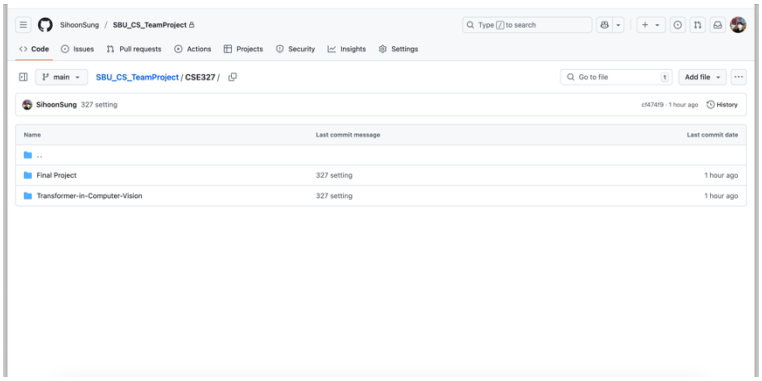
On top of this, the collaborative work on the open-source version available on GitHub was not as easy as we presumed. The complexity of the code base along with our requirements to tailor it towards our dataset and experimental purposes posed a great challenge. Attempting to execute the training scripts, we encountered numerous bugs and configuration problems that required batch renaming, learning adjustments, and even file-path alterations to achieve minimal operational ease. While such measures were a setback, they greatly improved our understanding of the training processes and setups for spike-based models.

Key learnings and insights

The major takeaway from this project for us was understanding how the Vision Transformer (ViT) works. Previously, we had a cursory understanding of Transformers from the context of NLP, but in the course of this project, we developed a much fuller understanding of how ViT processes image data. It was particularly interesting to learn how the model uses self-attention to capture relationships across different parts of the image, and how positional encoding aids in the preservation of spatial information despite the fact that the input is treated as a sequence of patches. This understanding greatly aided us when working on the spike-driven version’s structure during the later stages of the project.
A different significant realization arose from working with spike driven models. The way they ‘think’ about perception was different from regular vision models because they didn’t accept pixel data, but rather required information to be represented in biological event-driven formats. That’s when it hit us that feature extraction is not only useful, but essential. First, we needed to

transform raw image data into meaningful features, then encode them into spikes with the transformer. It made us rethink our entire data pipeline and appreciate the role of data preprocessing on model compatibility and performance.

GitHub



We've created a GitHub repertoire to create code and conduct experiments together.
We're starting right now, and we're going to set a time every week and always develop it together.

Team member's role

Mingu Hwang

- Feature extraction
- Find defects and re-schedule plan

Sihoon Sung

- Data pre processing
- Kaggle data set
- Find information about ViT

Plan

Weeks	Task Description
Weeks 1–2	Initial setup and testing using CIFAR-10 with the Spike-Driven Transformer (DONE)
Weeks 3–4	Attempted manual movie poster collection and labeling (discontinued)
Weeks 5–6	Switched to Kaggle movie poster dataset, preprocessing, and exploratory analysis
Weeks 7–8	Feature extraction using CNN and spike encoding for ViT-compatible input
Week 9	Spike-Driven Transformer fine-tuning using movie poster data
Week 10	Final experiments, comparison with CNN baseline, and result analysis + report