Final Project
Project Progress Report
#1

Movie Genre Classification from Posters using Vision Transformers

Mingu Hwang
mingu.hwang@stonybrook.edu

Sihoon Sung
sihoon.sung@stonybrook.edu

Review of relevant research papers

We found research papers about Spiking Neural Network (SNN) and Spike-driven Transformer (SDT). First is about (arXiv 2022.10) Spikformer: When Spiking Neural Network Meets Transformer.

Spikformer is a model that combines the powerful self-attention mechanism of Transformer with the low-power event-based computational characteristics of the Spiking Neural Network and has been proposed to solve the high computational volume and energy consumption around the Spiking Self-Attention module, which is SSA, and this module removes Softmax and multiplication operations and redesigns self-attention to suit the spike-form computational structure of SNN.

Spikformer performs convolution-based Spiking Patch (SPS) on the input image to divide it into patch units and generate features in the form of spike. Subsequently, Spikformer Encoder Block uses SSA to model the relationship between patches and delivers information through Multi-Layer Perceptron (MLP) and Residual Connection SSA converts Query, Key, and Value into spike values represented by 0 and 1 and calculates the attention using only logical AND operations and addition. This contributes to simultaneously reducing the amount of computation and energy consumption by maximizing the sparsity of SNN. Finally, the classification results are output through Global Average Pooling (GAP) and the Fully Connected Layer.

Spikformer proved its potential as an SNN-based model by recording excellent performance on various static and neural visual datasets such as ImageNet, CIFAR-10/100, and DVS128 Gesture. It significantly improved the computational efficiency while maintaining higher accuracy than the existing SNN structure. However, there was a limit to seeing it as a completely non-multiplication structure because some multiplication or floor operation-like structures remained inside SSA.

We also looked up for Spike-driven Transformer (arXiv 2023.07 Spike-driven Transformer), since it has more energy efficiency and less computation by omitting multiplications. So this is what Spike-driven Transformer paper says.

The Spike-driven Transformer model builds upon Spikformer's methodology by removing all multiplication and accumulating (MAC) from transformer operations and enabling complete event-based processing characteristic of SNNs. This model proposes Spike-driven Self-Attention (SDSA), a self-attention mechanism that completely avoids the use of multiplication, and implements Membrane Shortcut without floating point residue accumulation, thus constructing the entire computational framework purely using additions and logic.

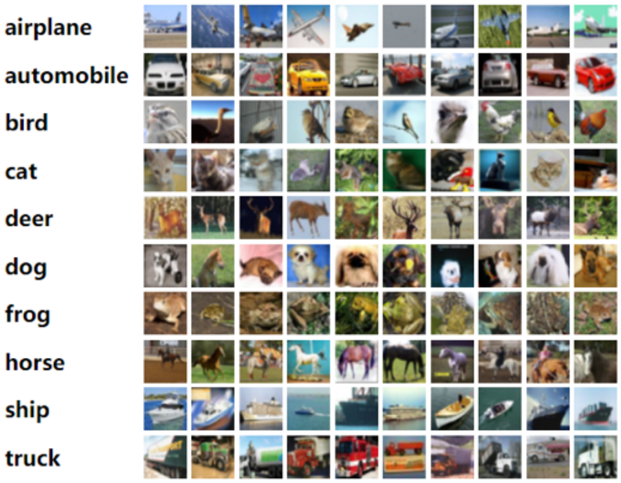
In the part of SDSA, Query, Key, and Value transforms into spikes, skipping the calculation of SoftMax, scaling, and multiplications, while the attraction value is determined by

Hadamard Product, column addition, and the spike neuron layer. This change decreased the complexity from O(N^2d) to O(Nd) with the computation, which performs up to 87.2 times better in energy efficiency.

Through these enhancements, Spike-driven transformer outperforms Spikformer in both energy efficiency and accuracy, with the decreased energy consumption while improving accuracy based on Datasets with a fixed parameter count. As a result, Spike-driven Transformer is a model that has developed into a complete addition-based transformer structure that can be applied directly to real SNN hardware, based on Spikformer's ideas.

Dataset

The dataset selected in this project is CIFAR-10/100. CIFAR-10/100 is one of the most widely used benchmark datasets in the field of computer vision. CIFAR-10/100 consists of 60,000 color images classified into a total of 10 classes. Each image is an RGB image with a size of 32x32 pixels and is divided into 50,000 for learning and 10,000 for testing.



There are many reasons why we used the CIFAR-10/100 dataset in this project. We will introduce three of the most important reasons.

1. **This is because the dataset we use in the research paper is CIFAR-10(/100).**

   This project is implemented based on the Spike-Driven Transformer paper and the GitHub repository published in NeurIPS 2023. In this paper, CIFAR-10/100 was used as a dataset to verify the model performance. Therefore, the code used to conduct this experiment was also made based on CIFAR-10/100. We are planning to use the code used in this paper as it is to proceed with the experiment, so we decided that it would be most helpful to use the dataset used in this paper as it is.

   Therefore, the most accurate result value can be obtained by using the same dataset in this project and can be directly compared with the data in the paper.

2. **There are various studies that have already been conducted and data that can be helped.**
   CIFAR-10/100 is a dataset used in a variety of studies. Even it is a dataset basically supported by major deep learning frameworks such as PyTorch and TensorFlow and can be easily used by anyone. Because it is easy to access, there are numerous public models, tutorials, and experimental references. Therefore, it is easy to solve problems in the learning process, and there are plenty of materials that can be referenced for model improvement and tuning.
3. **It is suitable for comparing the model architecture.**
   This project compares the performance of Convolutional Neural Network (CNN), Vision Transformer (ViT), and Spike-driven Transformer (SDT) and examines the advantages of a Spike-based vision model. A dataset of sizes that can be used in these models is generally not easy to create. However, CIFAR-10/100 is a dataset that meets all of these conditions and can be applied to anything we need to do.

In addition to this reason, CIFAR-10/100 is a dataset that can be used in the Spiking Neural Network through simple processing. Therefore, the comparison with CNN and Spike-driven Transformer, and the dataset we use in the papers and experiments we refer to, is the reason why we selected CIFAR-10/100 as our dataset and should be selected.

The training approach

The training process in this project is based on the official implementation provided by the Spike-Driven Transformer GitHub repository(https://github.com/BICLab/Spike-Driven-Transformer), which was released alongside the paper accepted at NeurIPS 2023. The training strategy follows a spike-based adaptation of the Vision Transformer, with special attention to computational efficiency and energy-aware operations.

First of all, the Spike-Driven Transformer is different from the traditional ViT structure. First, the input data is converted into a binary spoke form rather than a real value and processed. And during the attraction operation, the operation efficiency is increased by using only the masked addition operation without multiplication. This part can be said to be the biggest difference. These modifications allow the transformer to process input images in the form of time-dependent spikes, enabling efficient computation.

We will proceed using the CIFAR-10/100 dataset. GitHub supports this dataset by default, so you can use it more conveniently. During preprocessing, the dataset is normalized and optionally converted into spike trains depending on the encoder setting.

- **Framework**: PyTorch
- **Loss Function**: Cross-Entropy Loss
- **Optimizer**: Adam
- **Learning Rate**: Configurable via .yaml files, typically with cosine decay
- **Epochs**: 100~300 depending on final experiment setup
- **Batch Size**: 128 or 256
- **Evaluation Metrics**: Accuracy (Top-1), FLOPs, spike sparsity, and latency

In addition to the Spike-Driven Transformer, the CNN-based model and the existing ViT model will be used and compared together. We will do a comparative analysis focusing on accuracy and efficiency.

Result

Currently, we have completed the data preprocessing and experimental setup for the Spike-Driven Transformer model. We are currently preparing to conduct the experiment and are in the process of learning about the CIFAR-10/100 dataset. The full results are not yet available, but we expect the Spike-Driven Transformer to show higher accuracy and efficiency compared to the CNN and ViT models.

Challenges and Unsolved Issues

We still need to build and develop Spike-driven Transformer from the GitHub, and we need to test the datasets. After testing the Spike-driven Transformer model, we need to find the CNN model for classification, build and learn the model through the same dataset. Then we compare two models and find out which model is more sufficient to classification.

Until we submitted this assignment, it was most difficult for us to study new concepts on our own. We think the hardest part was trying to understand the contents of the thesis. But we were able to find a lot of helpful materials, and we think we're going well even though we haven't finished it yet.

Key learnings and insights

Like we said in the proposal, we need to find out that Spike-driven Transformer is better than CNN, which certifies transformers as the superior model in the vision-based classification arena, especially with respect to movie poster-driven genre tagging. This was not included because we have not yet been able to derive experimental results using models. However, we were able to understand CNN, and Spike-driven Transformer in more detail in the course of studying the thesis. We're sure that we can learn more important points in the future.

Team member's role

Mingu Hwang

- Build and train the Spike-driven Transformer model from GitHub
- Build and train the CNN model from GitHub

Sihoon Sung

- Data collection and preprocessing for both models
- Test both of models
- Write about the comparison analysis and final report

Plan

Weeks	Task Description
Weeks 1–2	Data research, preprocessing, and experimental setup for Spike-Driven Transformer (**DONE**)
Weeks 3–4	Spike-Driven Transformer model implementation and experiments
Weeks 5–6	Data research, preprocessing, and experimental setup for CNN model
Weeks 7–8	CNN model implementation and experiments
Week 9	Comparison and performance evaluation of CNN and Spike-Driven Transformer
Week 10	Final result analysis and report writing