COMP411/511: Project Proposal

Deadline: November, 8th

Tolga Ermiş
Cem Ozan Doğan
Azad Aslanlı

1 Problem and Motivation

- **Problem Definition**

  Our project will focus on action recognition in videos, where the goal is to accurately identify and classify actions. We will develop a model capable of capturing motion details and can compete with the state-of-the-art models while being optimized for efficiency and shorter training time.

  To achieve this, we will develop a neural network architecture that handles the features in video sequences. This architecture will make use of recent advancements in deep learning for video analysis, with modifications to increase accuracy and efficiency.

- **Motivation**

  Action recognition is a very important task in computer vision and has wide-ranging applications in many fields. With the increasing availability of video content and advancements in technologies, the demand for high-performing action recognition models is growing.

  However, current state-of-the-art models often require too much computational resources and training time, which can limit their practical use. Developing a model that is both accurate and quick to train would make significant contribution to the field. This project could contribute to the development of intelligent systems capable of understanding and responding to human actions in variety of domains.

2 Plan

- **Starting Point**

  We will base our work on the Multi-View Transformer architecture[5]. The paper expands on the earlier work on vision transformers and addresses the need to capture both fine-grained motions and long-duration events by processing multiple "views" or perspectives of video inputs simultaneously. Each view represents different temporal resolutions, which are processed in separate transformer encoders with lateral connections to integrate information across views. The model improves accuracy while balancing computational efficiency, as demonstrated through extensive ablation studies and state-of-the-art performance on multiple video classification datasets. Additionally, MTV benefits from large-scale pretraining, further enhancing its effectiveness in complex video recognition tasks.

- **Related Work**

  There has been many papers and models for the action recognition task. One of the models that was used earlier in the field is the ViVit: A video vision transformer[1]. The ViViT paper introduces a pure-transformer model for video classification, leveraging the self-attention mechanism to handle spatio-temporal data. The authors propose four model variations that factorize spatial and temporal dimensions differently to improve efficiency and scalability, addressing challenges posed by video data's long token sequences. They apply regularization techniques and use pretrained image models

to train ViViT effectively on smaller datasets, despite vanilla transformers traditionally requiring large datasets. Extensive experiments on the paper showed that ViViT achieves high performance on video classification benchmarks like Kinetics and Epic Kitchens[1].

Another important related work is Multiscale Vision Transformer(MViT2) model. It is a model designed for various visual tasks, including image classification, object detection, and video recognition. It introduces key improvements, such as decomposed relative positional embeddings and residual pooling connections, which help reduce computational complexity and memory usage while improving accuracy. The authors evaluate MViTv2 across multiple benchmarks, achieving high scores in ImageNet classification, COCO object detection, and Kinetics video classification. The architecture employs pooling attention mechanisms and hybrid window attention to balance performance and efficiency, outperforming earlier windowed approaches like Swin Transformers[3].

- **Plan and Timeline:**
  Week 1-2:
  - Dive deeper into key papers in literature
  - Select and preprocess appropriate dataset.
  - Set up data augmentation.

  Week 3-4:
  - Outline the architecture.
  - Decide on hyperparameters.
  - Conduct initial tests on model.

  Week 5-6
  - Set up training pipeline.
  - Train a simple baseline model.
  - Experiment with batch sizes and learning rates.

  Week 7-8
  - Fine-tune hyperparameters.
  - Test impact of different parameters.
  - Evaluate the model on a validation set.

  Week 9-10
  - Document the achitecture,training and optimizations.
  - Prepare graphs and tables to present our results.
  - Final evaluation on test data.
  - Project presentation.

3 Datasets and Metrics

- **Datasets**
  We decided to use the Something-Something-V2 dataset [4] as our main dataset to train our action recognizer model. Something-Something-V2 has 220,847 labeled video clips of humans performing pre-defined, basic actions with everyday objects. Additionally, we might use Kinetics-400/600/700 [2], HMDB-51, UCF-101, Charades, AVA (Atomic Visual Actions) and Epic-Kitchens datasets to further train or fine-tune our model.

- **Metrics**

  We will use these metrics to test and compare our model: Accuracy, Precision, Recall, F1 Score, Mean Average Precision(mAP), Top-k Accuracy, Confusion Matrix, latency and inference time.

- **Evaluation Goals**

  The success based on these metrics will be:

  1. Getting at least 75% accuracy.
  2. Getting at least 60% top-1 accuracy.
  3. Getting at least 75% top-5 accuracy.
  4. Having approximately the same latency and inference time as the baseline paper model[5].

4 Team Members

Cem Ozan Doğan: A computer engineering student with a strong interest in AI research since the release of GPT-3 in 2020. He is passionate about developing AI models capable of performing increasingly complex tasks and aims to continue his work in the AI field after graduation.

Tolga Ermiş: A CHBI & COMP double major student with interest in use of AI and on biological and medical problems. My aim is to dive deeper into the field of computer vision and apply my knowledge on such problems in the future.

Azad Aslanlı: 4th year Computer Engineering Student who has previously done 2 different internships. One of them is about AI based recruitment systems. The other one is an ML/Data Science project at Setur. Mainly interested in how AI and ML algorithms can be used in different areas.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[2] Joao Carreira and Andrew Zisserman. Kinetics-400: A large-scale video dataset for human actions. `https://deepmind.com/research/open-source/kinetics`, 2017. Accessed: 2024-11-07.

[3] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[4] Qualcomm AI Research. Somethingsomething v2 dataset. `https://www.qualcomm.com/developer/software/something-something-v-2-dataset`, 2018. Accessed: 2024-11-07.

[5] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.