Paper Review: "ViViT: A Video Vision Transformer"

Munish Monga
Centre for Machine Intelligence & Data Science
Indian Institute of Technology Bombay
Mumbai 400 076, INDIA
22m2153@iitb.ac.in

Aniket Thomas
Centre for Machine Intelligence & Data Science
Indian Institute of Technology Bombay
Mumbai 400 076, INDIA
22m2162@iitb.ac.in

Motivation

Taking motivation from the success of the Vision Transformers (ViT) in image classification, the authors in this paper explore the potential of extending transformer architectures to video classification. The paper delves deep into the challenges of video understanding, highlighting the criticality of capturing both spatial and temporal dynamics within videos. Drawing inspiration from the Vision Transformers (ViTs) in image classification, the authors saw an opportunity to bridge the gap between static images and dynamic videos. They were motivated to adapt and extend the principles of ViTs, which have proven effective for images, to address the increased complexity of video data and to develop models that can efficiently process and classify videos.

Novelties

The authors introduce two novel methods for mapping a video to a sequence of tokens, and they also propose four transformer models for video in the paper:

- The novelty lies in how embedding for video frames is done. In ViT, images are segmented into uniform patches, each treated as a "token". These tokens are flattened, linearly transformed, combined with position embeddings, and then fed into the transformer model. ViViT extends this concept to videos. The **Uniform Frame Sampling** approach is straightforward and involves selecting frames from a video at uniform intervals. Each frame is then treated similarly to an image in ViT, where it's segmented into patches, and each patch is transformed into a token.
- Their second novel approach uses **Tubelet Embeddings** that extracts non-overlapping spatiotemporal "tubes" from the video, extending ViT's embedding to 3D. Instead of processing individual frames, this approach extracts consecutive, non-overlapping 3D "tubelets" from videos.
- The introduction of the **Factorised Encoder** and **Factorised Self-Attention** models showcase their innovative approach to video understanding using transformers. The **Factorised Encoder** distinctly separates spatial and temporal dimensions, allowing for an optimized fusion of these two critical aspects of videos. This ensures efficient processing without compromising on the depth of understanding. On the other hand, the **Factorised Self-Attention model** re-imagines the traditional self-attention mechanism. Instead of a holistic attention computation, it factorizes the process, attending first to spatial features and then to temporal dynamics.

Major Contributions

The major contributions of the paper include:

- Presentation of two distinct methods for video tokenization: **Uniform Frame Sampling** and **Tubelet Embedding**.
- Introduction of the **Factorised Encoder** model, which consists of two separate transformer encoders and corresponds to a "late fusion" of temporal information.
- Proposal of the **Factorised Self-Attention** model, which factorizes the multi-headed self-attention operation to compute attention spatially and then temporally.
- Development of the **Factorised Dot-Product Attention** model, which factorizes the multi-head dot-product attention operation.

Critical Analysis

While the ViViT paper presents potentially great innovations in video understanding using transformers, certain aspects require critical examination. The introduction of the **Factorised Encoder** and **Factorised Self-Attention** models, though novel, raises questions about their scalability and adaptability to diverse video content. The distinct separation of spatial and temporal dimensions in the Factorised Encoder might lead to potential information loss, especially in videos where spatial-temporal interdependencies are crucial. Additionally, the factorized approach in self-attention, though computationally efficient, might not always capture the intricate relationships within video frames, especially in complex dynamic scenes. The paper's emphasis on efficiency, while commendable, might come at the cost of depth in video understanding.

A standout in the paper is Model 2 (Factorized Encoder). Its superior performance across datasets like Kinetics-400, 600, Epic Kitchens, and SSv2 is noteworthy. However, the necessity to initialize its larger variants using weights from the private JFT-300 dataset for the Kinetics dataset highlights the challenges inherent to the video classification domain. However, the authors' innovative approach of using image classification weights as an initializer is intriguing, suggesting potential cross-modal benefits. Despite the above critical analysis, it is clear that the paper sets the stage for further research in this domain, emphasizing the need for efficient and scalable transformer architectures for video understanding.