Convolution-augmented Vision Transformer for Improved Video Data Analysis

https://anonymous.4open.science/r/570Project-6E8E/

ABSTRACT

This paper presents a hybrid convolution-augmented vision transformer architecture designed to enhance video data analysis, addressing the challenges of efficiently capturing both fine-grained spatial details and broader temporal relationships within video content. The model leverages the complementary strengths of Convolutional Neural Networks (CNNs) for extracting local spatiotemporal details — such as object appearances and movements across frames — and Vision Transformers (ViTs) for capturing global and contextual information across video sequences. The architecture is further enhanced using reinforcement learning from human feedback, where a reward model guides the iterative refinement of generated captions to better align with human expectations. This integration of human-in-the-loop feedback ensures that generated video captions are not only syntactically correct but also contextually coherent, capturing both the intricate details of individual frames and the broader story conveyed by the video. The model is trained and evaluated on the MSR-VTT dataset, and preliminary results indicate a significant improvement in capturing nuanced video details, enhancing contextual coherence, and generating accurate, human-aligned captions. This hybrid solution holds promise for various applications in video understanding, including summarization, action recognition, and more.

INTRODUCTION

The field of video data analysis, particularly involving AI and machine learning, is rapidly evolving. Driven by the need to process large amounts of visual data for a variety of applications including autonomous driving, surveillance, healthcare, and entertainment, there arises a need for such a tool that analyzes both the small, unique details in every frame of the video while keeping the overall context of the scenes into perspective. Over the last 15 years, convolutional neural networks (CNNs) have become the backbone of computer vision research due to their ability to extract features efficiently from images by leveraging their spatial structure, allowing for identifying patterns and objects within an image regardless of its position or orientation. In 2012, the transformer architecture was introduced, giving a parallelizable approach to sequence processing primarily for applications in natural language processing. Such an architecture was then introduced for visual data, the vision transformer (ViT), which could capture global dependencies while extracting lower-level features from images. A fusion of the traditional CNN and transformer architectures, known as the Conformer, was introduced in 2020 for speech recognition to capture local acoustic patterns and the global linguistic context of speech. This paper discusses the implementation of a hybrid architecture involving the CNN and ViT models to efficiently analyze video data. To improve the model's subjective quality assessment, a reward model aimed to incorporate human feedback into generating responses is utilized.

PROBLEM STATEMENT

The rapid expansion of video content across numerous applications, such as surveillance, entertainment, healthcare, and autonomous driving, has created a pressing need for robust video data analysis tools that can interpret and summarize video content efficiently. Traditional models face significant challenges in efficiently capturing both the fine-grained, local spatiotemporal details within individual frames and the broader, global contextual relationships that span entire video sequences. CNNs excel that extracting local features from individual frames, while Vision Transformers (ViTs) are adept at capturing global dependencies across sequences. However, each architecture alone struggles to deliver a complete understanding of complex video content. This work seeks to bridge this gap by proposing a hybrid convolution-augmented Vision Transformer architecture that leverages the strengths of both CNNs and ViTs to improve video data analysis. Moreover, it integrates reinforcement learning with simulated human feedback to iteratively refine generated captions, aligning them more closely with human expectations. This novel approach aims to generate video summaries that are not only accurate and coherent but also contextually meaningful, addressing a critical need in video understanding tasks. The proposed architecture is designed to overcome the limitations of existing models in capturing the full scope of visual information in videos, thereby enhancing applications in video summarization, action recognition, and anomaly detection.

RELATED WORK

This project focuses on the following papers to gain a better understanding of current solutions for computer vision-oriented tasks and highlight the importance of human feedback in optimizing machine learning models.

- *Conformer: Convolution-augmented Transformer for Speech Recognition*, to further understand the default Conformer architecture and how it can be modified for video data
- *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*, to explore the vision transformer architecture currently built for computer vision tasks
- *Learning to Summarize with Human Feedback*, to leverage the potential of human feedback for aligning machine learning systems with preferences established by humans

The Conformer Architecture

The Conformer architecture aims to address the challenge of efficiently modeling local and global dependencies in sequential data used for automatic speech recognition (ASR) tasks.

Recurrent neural networks, an architecture commonly used for such tasks, excel at modeling time-series and sequential data problems due to their ability to remember previous inputs. However, they face limitations in modeling long-range interactions effectively. Transformers, a major breakthrough in the natural language processing (NLP) field, may not fully exploit local feature correlations despite being able to capture global dependencies by utilizing self-attention. Hence, the combination of transformers with CNNs, an architecture specifically designed to efficiently extract features from data by leveraging their spatial structure, allows for a hybrid model that captures local acoustic features and global linguistic contexts.

The proposed Conformer block includes four main components:

- a *feed-forward module*, responsible for the non-linear transformation of the input data to generalize to different types of speech by learning higher-level, abstract representations of the acoustic signal
- a *multi-head self-attention module*, designed to allow the model to attend to any part of the input sequence regardless of its position to understand the relationships between distant parts of the signal
- a *convolution module* to model local dependencies in the temporal domain, such as phonetic transitions or fine-grained acoustic features (pitch, intensity, or phoneme boundaries)
- an additional feed-forward module to further refine the combined global and local features to create rich, high-quality representations of the ASR model

The Conformer achieved state-of-the-art results on the LibriSpeech benchmark (2.1%-2.3% word error rate), outperforming the Transformer, LSTM-based models, and similar-sized convolution models. The authors of the paper also tested the architecture with a language model added, which achieved lower word error rates among all existing models. While there are benefits to the Conformer such as the use of depthwise separable convolutions to reduce the computational cost and its unique combination with convolution, further research is necessary to identify the applicability and effectiveness on other kinds of data while testing its computational complexity with larger model sizes and larger datasets.

The Vision Transformer Architecture

As mentioned in the previous section, the Transformer architecture was aimed at solving NLP problems and had fewer implementations for computer vision tasks. Vision Transformers (ViTs) are models that directly apply transformers to sequences of fixed-size patches. The ViT architecture consists of:

- a *patch embedding layer* that generates a sequence of non-overlapping patches (each treated as a token) and linearly embeds a vector-form of the tokens to match the input dimension of the transformer
- learnable *positional embeddings* that are added to the patch embeddings to preserve the spatial structure of the image
- several *transformer encoder* layers that include multi-head self-attention, capturing global relationships of the image patches across the entire image, and a feed-forward network applied to each token
- *layer normalization* and *residual connections* to stabilize the training process
- a *classification head* (or *classification token*) that collects information from all patches through attention layers and is passed through an MLP head to output class predictions

A crucial difference between the ViT and original transformer architectures is the adaptation to 2D image data instead of the typical 1D sequential data found in speech recognition tasks. Such an architecture is scalable and requires less inductive bias compared to CNNs. ViT achieved an accuracy of 88.55% on ImageNet and 99.50% on CIFAR-100, two popular image classification benchmarks.

Despite its superior performance when trained on large datasets, ViT struggles to compete with CNNs on smaller datasets as the convolutional inductive bias on locality is more beneficial with lesser data. Additionally, the self-attention mechanism in ViT scales quadratically with the input sequence length, meaning that high-resolution images or smaller patch sizes significantly increase computational demands. As mentioned in the project, training ViT models with human feedback through a process called *Reinforcement Learning from Human Feedback (RLHF)* can iteratively improve their performance when aligned with human expectations.

Human Feedback for Optimization

A common issue found in machine learning research is the misalignment between the objectives optimized by typical language models, which often rely on automatic metrics, and the true goal of generating high-quality outputs as humans expect. Focusing on language model training applications, the authors of the paper aim to demonstrate that directly incorporating human preferences into the training process through reinforcement learning can significantly improve the quality of the generated results.

To analyze the impact of human feedback on higher-quality results, the following experiments were used to draw conclusions:

- *Data Collection: Gathering Human Preferences on Reddit TL;DR*: Using the TL;DR dataset, the researchers used various models to generate summaries of Reddit posts. After using supervised baselines, the ”zero-shot” GPT-3 Model, and the current policy (using RL) for the summaries, the researchers presented human labelers with pairs of summaries for the same post and asked them to choose the better summary. By using their feedback and recognizing the potential for noise in the TL;DR dataset, several filtering steps such as focusing on posts within human -written summaries within a specific token length range were utilized to ensure data quality and control.
- *Reward Model (RM): Predicting Human Preferences*: The human comparisons were used as training data for the RM to predict which summary a human would prefer. This architecture involved a randomly initialized linear head on top to output a scalar value representing the ”reward” or preference score for a summary given a post (x) and two summaries (y_0, y_1). After training, the RM's outputs were normalized such that the average score assigned to the human written reference summaries in the TL;DR dataset was 0.
- *Reinforcement Learning (RL): Optimizing the Policy with PPO*: The paper optimized the summarization policy using Proximal Policy Optimization (PPO), a reinforcement learning algorithm that trains a computer agent to complete tasks. The RM's outputs, the log odds of human preference, are used as the reward signal for the RL agent. A KL divergence penalty is introduced to stabilize learning and a separate value function as used to estimate rewards more accurately during training, aiding the alignment of model outputs with human feedback.

The studies proved that the 1.3B parameter model trained with human feedback outperformed a 6.7B parameter model trained with supervised learning, proving the effectiveness of directly aligning the training objective with human judgment. The RMs mentioned in the study consistently outperformed metrics like ROUGE, commonly used to evaluate automatic summarization and machine translation software in NLP. As the paper focuses on binary comparisons for feedback, exploring other forms of feedback such as explanations or edits can further enhance model learning, while avoiding potential biases introduced by human labelers. Regardless, the applications of such models are endless especially in the field of video data analysis where human feedback can be used to ensure globally consistent outputs, mitigating issues of temporal incoherence. Human feedback can be used to develop more robust evaluation metrics for video-based models and fine-tune outputs on specific aspects of video quality that are difficult to capture with automatic metrics, such as video summarization. Tailoring the type and granularity of such feedback to the specific video-related task will be crucial for effective model training.

IMPLEMENTATION

Preliminary Implementation Details

Dataset: MSR-VTT, the dataset selected for this project, is a large-scale video benchmark consisting of 10,000 video clips and 200,000 captions that describe a wide range of human activities, events, and scenes. This dataset is widely used for evaluating video understanding and captioning models, as it contains diverse content covering multiple categories, such as sports, cooking, and travel. In this project, the dataset is used to improve video captioning, ensuring that the model is able to learn both specific details and general contextual features, which are critical for generating accurate and coherent captions that capture the essence of the visual content. The captions provided with the dataset also serve as ground truth references for the reward model, allowing the model to iteratively improve generated captions through reinforcement learning.

Preprocessing: To prepare video data for the hybrid architecture, the raw video file first needs to be decoded into individual frames. The frames are then sampled at fixed intervals or using a more sophisticated sampling strategy to capture important temporal events. The goal is to ensure that significant motion cues and scene changes are preserved, providing a representative summary of the entire video. All frames are resized to a consistent size of 224x224 pixels to standardize the input for the next stages, necessary for efficient processing by the CNN and the ViT.

The input for the CNN comprises a 4D tensor of resized video frames (ex. [batch_size, sequence_length, height, width, channels]) preserving the spatial and temporal information of the video. This structure ensures that both spatial (height, width) and temporal (sequence length) information of the video is preserved, allowing the CNN to capture detailed object appearances and movement across the timeline. By processing the video frames as a complete tensor, the CNN can learn spatiotemporal patterns and understand the changing dynamics of objects within the video, such as how objects interact, move, and transform over time.

On the other hand, the input for the ViT is a sequence of embedded 2D spatial patches with positional encodings.

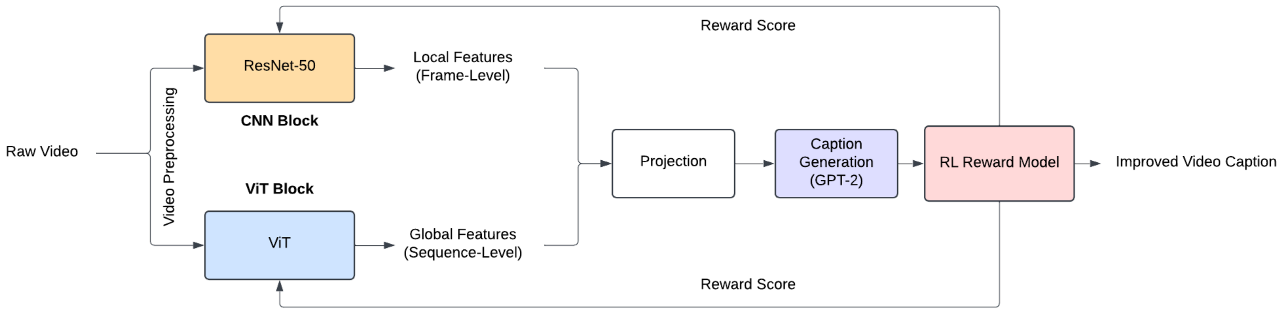

Figure 1: Model Architecture

For each resized frame, 2D spatial patches are extracted, flattened to a lower-dimensional embedding, and positional embeddings are added to each patch. This allows the ViT to model relationships between different regions of a frame and capture global contextual information across the entire scene, rather than focusing solely on specific local features. By processing spatial patches, the ViT effectively understands the connections between various segments of each frame, providing a comprehensive, high-level understanding of the visual content.

This complementary design — CNN for capturing detailed spatial features of individual frames and ViT for understanding global spatial relationships and context — allows for a richer understanding of the video content through a unified hybrid architecture.

Model Architecture: The parallel hybrid architecture is then implemented by incorporating 3D CNN layers for spatiotemporal feature extraction and ViT for global contextualization within frames.

It is necessary to extend the convolutional layers in the Conformer blocks from 1D to 3D due to the dimensions of the video. Hence, convolutional and self-attention modules are necessary to be adapted to process 3D feature maps, enabling the model to better capture the sequential nature of video data. Similarly, the ViT model is adapted to operate on 3D inputs, transforming its spatial self-attention mechanisms to include temporal information as well. This modification is crucial, as it enables the ViT to provide a global, sequence-level understanding of the entire video as it considers context both within and across frames. An attention-based fusion mechanism dynamically integrates the CNN and ViT models, balancing detailed frame-level features with the broader temporal context.

For the CNN, ResNet-50 was leveraged due to its balanced performance, efficiency, and effectiveness in extracting robust spatial features. ResNet-50's residual connections mitigate the vanishing gradient problem, facilitating deeper feature extraction without degradation in performance. Utilizing the pre-trained model for video tasks in combination with the 3D ViT ensures the information from each frame is represented in the summary generated of the video alongside global contextualization from the ViT. Figure 1 provides a visual summary of the architecture.

Reward Model and Reinforcement Learning: The reward model (RM) is crucial in aligning the video captioning model's outputs with human preferences. In this project, we design a reward model that utilizes a large language model (GPT-2) to simulate human feedback. Instead of relying on manually collected human feedback data, which can be challenging and resource-intensive to obtain, we use GPT-2 to evaluate the generated captions against reference captions from the MSR-VTT dataset. This approach measures the similarity between generated and reference captions, providing a reward score that reflects how well the generated output captures the key aspects of the video content.

Using GPT-2 in this context provides several benefits. First, it allows for the rapid evaluation of generated captions, providing immediate feedback without the need for costly human annotation. Additionally, leveraging a large language model ensures that the evaluation process is scalable, as GPT-2 can be used across a large volume of data without requiring manual intervention. This approach also takes advantage of GPT-2's extensive training on diverse text data, which allows it to approximate human-like judgments when scoring the quality and coherence of generated captions.

The reward model iteratively guides the captioning system through reinforcement learning. By generating multiple candidate captions and evaluating them against reference captions using GPT-2, the model identifies and selects the best-performing captions, improving both coherence and contextual alignment. This approach helps ensure that the generated captions generalize well to new video content while maintaining consistency with human expectations, making it an effective alternative to using human feedback directly.

Potential Challenges: Though the hybrid architecture presents a unique solution to improving video data analysis, the following are potential challenges and solutions related to the overall architecture and implementation of the model:

- *Computational Complexity and Memory Constraints*: Though the Conformer and ViT models are known for their high accuracies for their respective applications, combining them in a 3D architecture for video processing can lead to a very computationally expensive model that will require a lot of time to train. Processing the spatial and temporal dimensions of video data, especially those at high resolutions, will require significant computational resources.

  However, scaling the model size to learn more efficiently while reducing the input resolution to process video more efficiently are two solutions to tackle the computational expense of the model.

- *Training Data Requirements*: Training a model with the capacity of the hybrid architecture will require a large and diverse video dataset to learn the complex relationships between the spatial, temporal, and global features in the data. Unless the actions are carefully specified while testing, the model will need to be exposed to a large variety of actions, scenes, and other camera-related variations for accurate detections and classifications.

  As a solution to such an issue, ViT and language models have benefited from self-supervised pre-training. Pre-training on a large, unlabeled video dataset using objectives like contrastive learning, transfer learning, or masked prediction might be useful to enable a model to learn useful representations without relying on labeled data.

- *Reward Model Design and Generalization of Results*: Ensuring that the reward model (RM) accurately aligns generated outputs with human expectations is critical for high-quality video summarization. One major issue with using large language models (LLMs) in the reward model is the risk of hallucination, in which the model may generate information not directly supported by the provided data. Hallucinations are particularly problematic in video processing, as they lead to results that contain irrelevant or fabricated details.

  To combat this, the project emphasizes grounding the LLM-based reward model strictly in the input video data and reference captions. Techniques such as factual consistency checks and constrained decoding ensure that generated outputs remain faithful to the content of the video. By enforcing constraints and monitoring for consistency, the project aims to prevent hallucinations and ensure that the model's summaries accurately reflect the input data. This approach helps maintain the reliability and accuracy of the hybrid architecture, especially in sensitive applications where factual accuracy is paramount.

EXPERIMENTAL RESULTS

This section provides a summary and explanation of the current results received after training the hybrid convolution-augmented vision transformer architecture on the MSR-VTT dataset. As presented in Figure 2, the generated captions displayed a significant mismatch with the actual video content. The generated summaries often contained irrelevant or incoherent information, failing to capture the intended context of each video. Notably, many captions were nonsensical when compared to the associated video. Despite achieving a low training loss average of 0.075, the outputs highlight that the model overfitted to the training data rather than learning to generalize effectively. The observed issues after viewing the results include:

Video Content: A man is driving a car
Provided Original Caption: ('a demonstration of someone playing minecraft', 'a man in a black suit discusses election with a man appearing on a tv screen', '3 kids singing on the voice', 'a cartoon of a spider and the sun')
Generated Caption: This is a game about a group of people who are trying to save the world from a group that is trying to destroy it.

Figure 2: Results

- *Inaccurate and Irrelevant Captions:* The captions generated during testing were generic and did not reflect the specific content of the videos. For instance, summaries that should have been focused on specific gameplay were instead about unrelated topics like storylines or events that did not occur in the videos.
- *Possible Caption-Video Misalignment:* The nonsensical output combined with the poorly matched original captions points to potential misalignment in the dataset. If videos and captions are mismatched, this could significantly disrupt training as the model would not learn to associate video frames with correct narrative content.
- *Overfitting with Low Loss:* The low training loss typically suggest effective learning, but in this case the model's performance on unseen data is poor.

To improve model performance and prevent overfitting, it's essential to validate video-caption alignment by checking the video-caption pairs and correcting any misalignments. Regularization techniques, such as L2 regularization and early stopping, can further reduce overfitting by promoting simpler model representations and stopping training once validation loss stabilizes. Refining the reward model with frame-level checks and using enhanced sampling methods could yield captions more relevant to the video content. Finally, simplifying the model architecture and adjusting learning rates for pre-trained and newly added layers separately can help achieve better generalization and stability in training.

CONCLUSIONS AND FUTURE DIRECTIONS

A convolution-augmented vision transformer crafted for video data aims to combine the strengths of CNNs and ViTs to model spatial and temporal patterns in the data effectively. By incorporating parallel CNN and ViT branches, the architecture effectively splits the responsibility for local feature extraction and global context modeling. This division enables the model to capture both intricate frame-level details and long-range dependencies, thereby providing a richer understanding of the video content. The use of reinforcement learning, via a reward model, further aligns the model's outputs with human preferences, allowing for iterative improvement based on evaluative feedback.

The modular design of this architecture, with clearly defined CNN and ViT branches, provides not only better interpretability but also the flexibility to adapt to various video-based tasks. The proposed solution is well-suited for tasks such as video summarization, action recognition, and video quality assessment. Beyond these, the model's ability to flexibly handle both local and global video features suggests its potential for specified applications like those mentioned below, where human feedback can be critical in assessing nuanced outcomes:

- *Action Recognition with UCF101 Dataset*: The UCF101 Dataset is a commonly-used dataset containing 13,320 video clips from 101 different human action categories including sports, exercise, and other such physical activities. To recognize actions within video data, the hybrid architecture can be trained on this dataset by utilizing the CNN to capture local movements and the ViT to capture the temporal context of the action. The reward model can then be applied by incorporating human feedback on the clarity and accuracy of the predicted actions, leading to improved human-aligned recognition.
- *Anomaly Detection in Surveillance Videos*: The hybrid architecture can also be utilized for anomaly detection in surveillance footage, where identifying rare events such as unusual activities or security breaches is crucial. In this application, the CNN captures detailed spatial features to understand local activities while the ViT helps maintain an overall understanding of typical behavior across time. The reward model can be used to learn from human feedback, where labeled anomalies are used to guide the model to correctly identify and distinguish between normal and unusual events, improving precision and reducing false alarms.

In future work, several areas can be further explored to enhance the proposed architecture. One direction could involve leveraging more sophisticated fusion mechanisms to integrate the CNN and ViT features more effectively, potentially through attention-based weighting schemes. Additionally, scaling the model through self-supervised pre-training on larger, diverse video datasets could help improve its generalizability and reduce reliance on labeled data. Finally, incorporating an expanded reward model that utilizes large language models, such as GPT-2, for assessing the quality of generated captions could provide an automated alternative to human feedback, especially when access to such labeled datasets is limited. These future improvements can further solidify the architecture's capacity to perform well in diverse, real-world video analysis scenarios.

References

Gulati, Anmol and Qin, James and Chiu, Chung-Cheng and Parmar, Niki and Zhang, Yu and Yu, Jiahui and Han, Wei and Wang, Shibo and Zhang, Zhengdong and Wu, Yonghui and Pang, Ruoming Conformer: Convolution-augmented Transformer for speech recognition *Interspeech* 2020 `https://www.isca-archive.org/interspeech_2020/gulati20_interspeech.pdf`

Stiennon, Nisan and Ouyang, Long and Wu, Jeffrey and Ziegler, Daniel M. and Lowe, Ryan J. and Voss, Chelsea and Radford, Alec and Amodei, Dario and Christiano, Paul F. Learning to Summarize with Human Feedback *Conference on Neural Information Processing Systems (NeurIPS)* 2020 `https://proceedings.neurips.cc/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf`

Dosovitskiy, Alexey and Beyer, Lucas and Kolesnikov, Alexander and Weissenborn, Dirk and Zhai, Xiaohua and Unterthiner, Thomas and Dehghani, Mostafa and Minderer, Matthias and Heigold, Georg and Gelly, Sylvain and Uszkoreit, Jakob and Houlsby, Neil An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)* 2021 `https://openreview.net/pdf?id=YicbFdNTTy`

He, Kaiming and Zhang, Xiangyu and Ren, Shaoqing and Sun, Jian Deep Residual Learning for Image Recognition *Computer Vision and Pattern Recognition* 2015 `https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf`