CAST: Cross-Attention in Space and Time for Video Action Recognition

DATA MINING TERM REPORT
Rana Kashyap Chitrang, Abbir Sarkar, Siddharth Jain
(Group No:1)
25 April 2024

***Abstract*—This paper addresses the challenge of recognizing human actions in videos by proposing a novel two-stream architecture called Cross-Attention in Space and Time (CAST). Unlike existing models,CAST achieves a balanced spatio-temporal understanding of videos solely using RGB input.The proposed bottleneck cross-attention mechanism facilitates information exchange between spatial and temporal expert models, resulting in synergistic predictions and improved performance.Extensive experiments on public benchmarks, including EPIC-KITCHENS-100, Something-Something-V2, and Kinetics-400, validate the effectiveness of the proposed method. CAST consistently outperforms existing methods across these datasets, demonstrating robust performance regardless of dataset characteristics. The code is publicly available for further exploration.**

I. Introduction

To accurately recognize human actions in videos, understanding both spatial and temporal contexts is crucial. Models lacking fine-grained spatial understanding may misinterpret actions, leading to incorrect predictions. Similarly, a deficiency in temporal context understanding can also result in inaccurate predictions. Therefore, achieving a balanced spatio-temporal understanding is essential for accurate action recognition. Despite advancements in action recognition using Transformers, balancing spatio-temporal understanding remains challenging. Models performing well on static-biased datasets may not generalize to temporal-biased datasets, and vice versa. Multimodal learning, such as two-stream networks, addresses this challenge but can be computationally expensive. In this study, we introduce a two-stream architecture, Cross-Attention in Space and Time (CAST), to address the balanced spatio-temporal understanding challenge using only RGB input. CAST employs spatial and temporal expert models that exchange information through cross-attention, facilitating more effective learning. Extensive experiments across diverse datasets demonstrate CAST's superior performance in achieving balanced spatio-temporal understanding. Our contributions include introducing CAST, conducting extensive experiments demonstrating its effectiveness, and providing insights into the design choices behind achieving balanced spatio-temporal understanding.

II. Related works

In the domain of Video Action Recognition, the adoption of Convolutional Neural Network (CNN)-based methodologies has proliferated, leveraging both 2D and 3D CNN architectures alongside 2D and 1D separable CNNs, as well as two-stream CNN frameworks. These methodologies have demonstrated significant advancements, primarily owing to their robust inductive biases. In recent times, Transformer-based approaches have ascended in popularity due to their proficiency in long-term context modeling. Following this trajectory, we propose a novel two-stream transformer architecture comprising a spatial expert and a temporal expert, diverging from conventional two-stream CNNs by exclusively utilizing RGB input sans flow.

The integration of Cross-attention mechanisms has been pivotal in multi-modal learning paradigms, facilitating the exchange of information across diverse modalities such as audio, visual, and textual domains. Recent advancements have demonstrated the efficacy of cross-attention in fostering interactions between distinct views within the same video. Aligned with this progress, our proposal entails a cross-attention approach employing a solitary RGB input, featuring two discrete expert models: a spatial expert and a temporal expert. These experts engage in mutual attention through cross-attention mechanisms to cultivate a harmonized spatio-temporal comprehension.

Foundation models, honed through self-supervised learning on expansive web-scale datasets, epitomize adaptability and versatility. Exhibiting prowess across an array of computer vision, natural language processing, and audio recognition tasks, foundation models serve as a cornerstone in our methodology. Specifically, we leverage CLIP as our spatial expert, given its remarkable performance across more than 30 computer vision tasks.

The conventional "pre-training and fine-tuning" paradigm, while potent, imposes computational overheads, rendering it often impractical to fine-tune the entire model. Numerous studies advocate for parameter-efficient transfer learning strategies, advocating for the learning of a subset of parameters while maintaining the remainder frozen, proving effective in both NLP and computer vision domains. Extending image foundation models through adapter architectures has yielded promising outcomes in action recognition. Our proposed methodology adopts an adapter architecture with cross attention between two experts, empirically showcasing its superiority over existing adapter-based video models in fostering a nuanced spatio-temporal understanding.

III. CAST architecture

The model architecture of each expert is the same as the ViT except for adapters and the B-CAST module. All the other parameters are frozen, while the adapter and B-CAST parameters are learnable. For completeness, we first define the operations used and then describe the entire model architecture. Given an input X, we define Multi-Head Self Attention (MHSA) operation as follows:

$$MHSA(X) = Softmax((XW_Q)(XW_K)^T(XW_V)) \quad (1)$$

The model architecture for each expert in the CAST for VAR methodology closely resembles the Vision Transformer (ViT), with the addition of adapters and the B-CAST module. The parameters of all components except for the adapters and B-CAST are kept frozen while the adapter and B-CAST parameters are trainable.

The Multi-Head Self Attention (MHSA) operation is defined to capture spatial and temporal dependencies. An

adapter operation with linear down and up projection matrices is applied for feature refinement.

For each attention block, independent MHSA is applied for each expert with skip connections. The spatial path undergoes spatial attention, while the temporal path undergoes space-time attention.

To exchange information between the experts, a B-CAST operation is applied, allowing the experts to benefit from each other's strengths. This involves Multi-Head Cross-Attention (MHCA) operations and temporal-to-spatial and spatial-to-temporal cross-attention mechanisms.

The B-CAST architecture incorporates MHCA modules into bottleneck-shaped adapters, facilitating efficient learning. This involves plugging MHCA modules into adapters and adding new learnable positional embeddings for each MHCA. The output of B-CAST is fed into a feed-forward network for further processing. The B-CAST operation for T2S $\Phi_S(\cdot)$ is defined as follows:

$$\Phi_S(Y_s^{(l)}, Y_t^{(l)}) = \sigma(MHCA(E_s + LN(Y_s^{(l)} W_{U,s}), \\ E_t + LN(Y_t^{(l)} W_{D,t})))W_{U,s} \quad (2)$$

Overall, the CAST for VAR model architecture leverages MHSA, adapters, and the B-CAST module to capture spatio-temporal dependencies effectively, leading to improved performance in video action recognition tasks.

IV. The Key to CAST's Success

The CAST (Cross Attention in Space and Time) architecture represents a significant advancement in the field of video action recognition, offering a comprehensive approach to understanding the dynamics of actions in video sequences. At its core, CAST employs a sophisticated encoder-decoder structure, integrating both spatial and temporal cross-attention mechanisms to capture relevant visual cues and temporal dependencies within video frames. Spatial cross-attention modules embedded within the encoder layers enable the model to focus on important spatial regions across all frames, facilitating a nuanced understanding of the visual content. Concurrently, temporal cross-attention modules within the decoder layers allow the model to dynamically attend to relevant frames in the temporal sequence, aggregating contextual information over time. This fusion of spatial and temporal attention mechanisms ensures that CAST effectively captures both the spatial layout of objects and the temporal evolution of actions, leading to robust performance in fine-grained action recognition tasks. Through multi-level fusion of spatial and temporal features, CAST achieves a balanced representation of both fine-grained and coarse-grained temporal dynamics, enabling accurate prediction generation based on the combined spatial-temporal context. Moreover, CAST is designed with computational efficiency in mind, leveraging parallelization and optimization techniques to minimize computational overhead while maintaining high performance. Overall, CAST demonstrates the effectiveness of integrating cross-attention mechanisms in space and time for achieving state-of-the-art results in video action recognition, paving the way for further advancements in the field.

V. Implementation Details

In this section, the researchers offer a detailed account of their experimental setup and implementation methodology for each dataset. The experiments were conducted utilizing 16 NVIDIA GeForce RTX 3090 GPUs, with CAST implemented using PyTorch and built upon the existing codebase of VideoMAE.

Regarding data preprocessing, the researchers sampled videos to 16 frames and performed random cropping and resizing to 224×224 for each frame. They applied various data augmentation techniques, including mixup, label smoothing, horizontal flip, color jitter, and randaugment, along with repeated augmentation to diversify the training data. Notably, horizontal flip was not used on the SSV2 dataset. Additionally, to accommodate the time stride value of 2 in the patch embedding layer of the temporal expert (e.g., VideoMAE), only even frames were selected for the spatial pathway. After the patch embedding layers, both experts were fed input data of dimensions 3 channels × 8 frames × 224 width × 224 height. The same data preprocessing protocol was maintained across all experiments.

For model training, experiments were conducted using 2 nodes, each equipped with 8 GPUs. The researchers utilized the DeepSpeed library to ensure efficient multi-node training. Additionally, they augmented the effective batch size by implementing gradient accumulation to update the model weights. For the EK100 dataset, the update frequency was set to 4 iterations, resulting in a total batch size of 24. Since VideoMAE did not provide pre-trained weights specifically for the EK100 dataset, the researchers pre-trained VideoMAE on the EK100 without incorporating extra video datasets, following the recipe described in the VideoMAE paper. For all other experiments, pre-trained weights provided by the respective model repositories were utilized to ensure consistency and reliability in the results.

VI. Dataset

The B-CAST module was evaluated on two video datasets for action recognition: Kinetics400 (K400) and Something-Something-V2 (SSV2).

i) K400 is a large-scale third-person video dataset containing approximately 300K video clips and 400 human action classes. The dataset is divided into train/val/test sets, with 240K/20K/40K video clips, respectively. All videos are trimmed to around 10 seconds and sourced from various YouTube videos.

ii) SSV2 comprises over 220K short video clips depicting humans performing predefined basic actions with everyday objects. The dataset is split into train/val/test sets, with 168K/24K/27K video clips, respectively. It encompasses 174 human-object interaction categories.

For fine-grained action recognition, the B-CAST module was evaluated on the EPIC-KITCHENS-100 (EK100) dataset. EK100 is a large-scale egocentric video dataset spanning 100 hours of recording, capturing unscripted kitchen activities over several days. It comprises 90K action segments split into train/val/test sets of 67K/10K/13K. Unlike the previous datasets, EK100 defines actions as a combination of a verb and a noun, posing a more challenging recognition task compared to datasets where actions are represented by a single label, such as Kinetics-400 and Something-Something-V2.

VII. Qualitative Analysis

The provided data showcases the effectiveness of CAST (Cross Attention in Space and Time) through qualitative analysis conducted on additional sample frames extracted from the EK100 dataset, as illustrated in Figure 15. In this analysis, expert models specialized in either noun or verb prediction tasks were compared with CAST.

The results indicate that each expert model excels in its respective task of expertise, providing more accurate predictions. However, these specialized models demonstrate weaker performance when tasked with the other prediction task.

In contrast, CAST consistently delivers correct predictions for both the noun and verb prediction tasks across the sample frames. This suggests that CAST is capable of achieving a balanced spatio-temporal understanding, effectively capturing both spatial and temporal cues in the video data. This balanced understanding is crucial for fine-grained action recognition tasks, where the accurate identification of both the action itself (the verb) and its involved objects or entities (the noun) is necessary for comprehensive recognition and understanding of the action being performed.

Therefore, the qualitative examples presented in the analysis serve to highlight the efficacy of CAST in achieving robust performance across multiple aspects of video action recognition, emphasizing its capability to integrate spatial and temporal information effectively for improved accuracy and comprehensiveness in action recognition tasks.

VIII. B-CAST Architecture And Methodology

In the appendix section, the researchers delve into various aspects of the B-CAST architecture, offering comprehensive insights into its design and implementation.

Architecture Details: This subsection outlines the architecture of B-CAST, assuming the utilization of CLIP as the spatial expert and VideoMAE as the temporal expert. It elucidates the classification strategy for conventional and fine-grained action recognition datasets, emphasizing the different approaches adopted for each task. Additionally, the classification head used for predicting actions is described, highlighting the distinction between conventional action recognition datasets like Kinetics400 and Something-Something-V2, and fine-grained action recognition datasets such as EPIC-KITCHENS-100.

Implementation Details: Further elaboration on the experimental setup and implementation process is provided here. Details regarding data preprocessing, model training, and the utilization of off-the-shelf pre-trained weights are discussed. Information on the datasets used for evaluation is also provided, including action recognition datasets like Kinetics400 and Something-Something-V2, as well as the fine-grained action recognition dataset EPIC-KITCHENS-100. The data augmentation techniques employed to diversify the training data and ensure robustness in model performance are detailed.

Additional Quantitative Analysis: Supplementary quantitative analysis is offered in this subsection to complement the main paper. It covers the generality of CAST with different ViT architectures and pre-trained weights, along with the effect of B-CAST-specific positional embeddings on performance improvement. By showcasing the versatility of CAST across different architectures and pre-training datasets, the researchers demonstrate its adaptability and effectiveness in various settings.

Comparison with State-of-the-Art: To provide comprehensive information, tables are augmented for comparison with state-of-the-art methods. Additional details such as the number of frames per clip, computation complexity, and learnable parameters are included, enhancing the understanding of CAST's performance in relation to existing approaches. By juxtaposing CAST with state-of-the-art methods, valuable insights into its competitive edge are offered, highlighting its strengths in achieving balanced spatio-temporal understanding.

Class-Wise Performance Comparison: Class-wise F1 score improvement of CAST over spatial and temporal experts is presented here, offering insights into the model's performance across different action classes. By analyzing the improvement in F1 scores for individual classes, a nuanced understanding of CAST's efficacy in capturing fine-grained action details and improving classification accuracy is provided.

Limitations: Acknowledgment of the limitations of CAST, including computational complexity and challenges related to employing two models with significantly different architectures, is provided. Resource limitations and the inability to experiment with various input video lengths and model sizes are also highlighted. Transparent discussion of these limitations provides valuable insights for future research directions and potential improvements to the CAST architecture. Broader Impacts: Discussion of the broader impacts of the work is included, emphasizing potential applications such as surveillance and consumer-oriented video search and tagging, while also addressing privacy concerns associated with such technologies. Consideration of the societal implications of the research underscores the importance of ethical considerations in the development and deployment of spatio-temporal understanding systems.

IX. Overall Outcomes

In this section, the experimental findings that address several research questions that are outlined in the paper are shown:

Datasets: The researchers evaluated CAST on two prominent public datasets for conventional action recognition: Something-Something-V2 (SSV2) and Kinetics-400 (K400). SSV2 prioritizes temporal reasoning, while K400 exhibits static biases. Additionally, CAST was assessed on the fine-grained action recognition task using EPIC-KITCHENS-100 (EK100), where actions are defined as a combination of verbs and nouns, posing a more challenging scenario compared to conventional action recognition.

Balanced Spatio-Temporal Understanding: An analysis of the performance of existing methods across various tasks and datasets revealed a significant imbalance in spatial and temporal understanding among these methods. To address this, the researchers introduced two expert models: a spatial expert (CLIP) and a temporal expert (VideoMAE), aiming to leverage their complementary strengths for achieving balanced spatio-temporal understanding.

Analysis on Fine-Grained Action Recognition: A detailed analysis of CAST's performance in fine-grained action recognition tasks using the EK100 dataset was conducted. The results showcased significant improvements in both noun and verb prediction tasks compared to individual expert models. Qualitative analysis further demonstrated the effectiveness of CAST in achieving balanced spatio-temporal understanding, crucial for fine-grained action recognition.

Ablation Study on CAST Architecture: Comprehensive ablation studies were conducted to examine various design choices within the CAST architecture. These studies validated the effectiveness of information exchange mechanisms, the design of the B-CAST module, and the importance of optimal assignment of expert roles. Furthermore, the experiments highlighted the significance of bi-directional cross-attention for enhancing spatio-temporal understanding.