Research Proposal
on

Few-Shot Action Recognition: Leveraging LLMs for Effective Data Augmentation on Video

Master's admission candidate

RAHAMAN NAGIUR(纳吉)
nagiur@outlook.com
+86 184 1906 0056

Introduction

Action recognition, the automated identification of human actions in video, is a vital component of modern computer vision, impacting diverse applications like surveillance, human-computer interaction, and robotics. Its ability to interpret human behavior is fundamental for creating intelligent systems capable of interacting seamlessly with the world. A significant hurdle, however, lies in the substantial amount of labeled data traditionally required to train robust action recognition models. This poses a particular challenge for few-shot action recognition, where the goal is to learn new actions from only a limited number of examples. Standard deep learning models, optimized on massive datasets, often struggle to generalize effectively in these data-constrained scenarios, tending to overfit the available training data and exhibiting poor performance on unseen instances.

Existing approaches to few-shot action recognition explore meta-learning and transfer learning techniques. Meta-learning aims to learn how to learn, enabling models to quickly adapt to new tasks. Transfer learning leverages knowledge gained from pre-trained models on large datasets. While promising, these methods have limitations. A common technique to boost performance is data augmentation, which artificially expands the training data. Traditional data augmentation relies on random transformations such as cropping, rotation, and jittering. However, these transformations may not generate semantically relevant variations of the action, limiting their effectiveness in the few-shot setting, where a nuanced understanding is crucial.

Large Language Models (LLMs) have demonstrated remarkable capabilities in understanding and generating human language. Their potential to enhance computer vision, especially in guiding data augmentation, remains largely unexplored. We propose a novel approach leveraging LLMs to guide data augmentation for few-shot action recognition. We hypothesize that by using LLMs to generate descriptive and even counterfactual variations of actions (e.g., "what if the person was also holding a ball?"), we can create more targeted and effective synthetic training examples. This LLM-guided augmentation ensures that the generated variations are diverse and semantically grounded, leading to improved generalization. Our research aims to develop and evaluate this framework, investigating different LLM architectures, prompting strategies, and augmentation techniques to optimize synthetic data generation and improve few-shot action recognition performance on benchmark datasets.

Problem Statement

The development of robust action recognition systems is significantly hampered by the need for extensive labeled data, a resource often scarce for novel or specialized actions. This data scarcity is a major obstacle for few-shot action recognition, where models must learn to identify new actions from only a handful of examples. Traditional deep learning models, trained on massive datasets, struggle to generalize effectively in these scenarios, frequently overfitting the limited training data, resulting in poor performance on unseen instances. Existing approaches, such as meta-learning and transfer learning, offer partial solutions but often face limitations. Data augmentation, a common technique for artificially increasing the training data size, typically relies on random transformations that may not generate semantically meaningful variations of the action, thereby limiting its effectiveness in few-shot learning, where understanding subtle nuances is crucial. Therefore, there is a need for a more efficient and semantically-aware approach to data augmentation.

Research Objectives

This research addresses the limitations of current few-shot action recognition by developing and evaluating a novel framework for LLM-guided data augmentation. The core objectives are:

1. **Develop an LLM-based Action Description Framework:** Leverage LLMs to generate diverse, semantically rich action descriptions, including variations and counterfactual scenarios, evaluated by the relevance and diversity of generated descriptions.
2. **Design Augmentation Parameter Mapping:** Implement a method to map LLM descriptions to specific data augmentation parameters (Mixup, CutMix, rotation, etc.), assessed by the realism of augmented samples.
3. **Evaluate Augmentation Effectiveness:** Assess the impact of LLM-guided augmentation on few-shot action recognition performance using benchmark datasets (Kinetics, Something-Something, HMDB51) and metrics like N-way K-shot accuracy, comparing against random data augmentation.
4. **Compare with State-of-the-Art Methods:** Benchmark the proposed approach against existing few-shot action recognition methods to demonstrate its effectiveness.
5. **Analyze Design Choices:** Investigate the impact of different LLM architectures, prompting strategies, and data augmentation techniques to optimize the framework and understand the interplay between LLMs and data augmentation.

Research Questions

This research seeks to answer the following key questions regarding the application of Large Language Models (LLMs) to enhance few-shot action recognition through guided data augmentation:

1. **How can LLMs be effectively leveraged to generate diverse and semantically meaningful descriptions of human actions, including variations and counterfactual scenarios, suitable for guiding data augmentation?**
2. **What is the most effective method for mapping LLM-generated action descriptions to specific parameters for various data augmentation techniques, ensuring the creation of realistic and relevant synthetic data?**
3. **Does LLM-guided data augmentation significantly improve the performance of few-shot action recognition models compared to traditional data augmentation methods (e.g., random transformations)?**
4. **How does the proposed LLM-guided data augmentation approach compare to existing state-of-the-art few-shot action recognition methods in terms of accuracy and efficiency?**
5. **How do different LLM architectures, prompting strategies, and data augmentation techniques affect the overall performance of the LLM-guided data augmentation framework, and what are the optimal configurations for different action recognition scenarios?**

Literature Review

Few-shot action recognition aims to enable models to recognize novel actions from limited labeled examples. This challenging task has garnered significant attention, with various approaches being explored, including meta-learning, transfer learning, and data augmentation.

Meta-learning algorithms, such as Model-Agnostic Meta-Learning (MAML) [6] and TinyReptile [12], aim to learn a model initialization that facilitates rapid adaptation to new tasks with minimal data. These methods have shown promise in few-shot image classification [4] and have been adapted for action recognition. Transfer learning approaches leverage pre-trained models on large-scale datasets like Kinetics [9] to provide a strong starting point for few-shot learning [1]. However, the domain gap between source and target tasks can still hinder performance. Metric learning methods, like Relation Networks [13], learn an embedding space where similar actions are closer together, enabling effective comparison of few-shot examples.

Data augmentation is a crucial technique for mitigating the impact of limited data. Traditional methods involve applying transformations like cropping, rotation, flipping, and jittering [2]. More advanced techniques like Mixup [15] and CutMix [14] generate synthetic samples by blending or splicing images/videos, improving model robustness. However, these methods often apply transformations randomly, lacking semantic understanding of the actions. This can lead to unrealistic or irrelevant variations, which may not be beneficial, especially in the few-shot setting where preserving the core characteristics of the limited examples is crucial. Recent work has explored using GANs [7] and diffusion models [3] for data augmentation, but these methods can be computationally expensive and challenging to train, particularly for complex data like videos.

LLMs like GPT-3 [11] and BERT [5] have demonstrated remarkable capabilities in understanding and generating human language. These models have been successfully applied in various NLP tasks, but their potential in computer vision remains largely untapped. Vision-language models like CLIP [11] bridge the gap between visual and textual information, enabling zero-shot classification and other cross-modal tasks. This suggests the possibility of leveraging LLMs to generate descriptive variations of actions, which can then guide the data augmentation process.

Despite the advancements in few-shot learning, data augmentation, and LLMs, there remains a significant gap in effectively utilizing semantic information to guide data augmentation for few-shot action recognition. Existing methods either rely on random transformations or require computationally intensive generative models. There is a need for a more efficient and semantically-aware approach to data augmentation that can improve the generalization capabilities of few-shot action recognition models.

This research proposes a novel framework that leverages the power of LLMs to guide data augmentation for few-shot action recognition. By using LLMs to generate descriptive variations of actions, we aim to create more targeted and meaningful synthetic training examples. This approach seeks to address the limitations of existing data augmentation techniques by incorporating semantic understanding into the augmentation process, ultimately leading to improved performance in few-shot scenarios.

Research Methodology

The research methodology for developing and evaluating the proposed LLM-guided data augmentation framework for few-shot action recognition is structured into distinct phases, each addressing

a specific aspect of the framework. These phases are designed to progressively build and refine the system, culminating in a rigorous evaluation of its performance and effectiveness.

Phase 1: LLM-Based Action Description Generation

This initial phase focuses on harnessing the capabilities of Large Language Models (LLMs) to generate comprehensive and diverse descriptions of human actions. A pre-trained LLM, such as GPT-3 or an equivalent open-source alternative, will be utilized as the foundation. The core of this phase involves fine-tuning the selected LLM on a curated dataset of action descriptions. This dataset will be a combination of existing video captioning datasets and manually curated descriptions tailored to the specific actions of interest. Prompt engineering is crucial to elicit the desired variations, counterfactual scenarios, and subtle nuances in action execution. We will explore different prompting strategies, including zero-shot prompting (where the LLM generates descriptions without any examples), few-shot prompting (where the LLM is provided with a few examples of action descriptions), and fine-tuning with specifically designed prompts that encourage the generation of diverse and semantically rich outputs. The quality of the generated descriptions will be assessed through a combination of human evaluation, where human annotators rate the relevance and accuracy of the descriptions, and automated metrics, such as BLEU or METEOR, which measure the semantic similarity between the generated descriptions and ground truth descriptions.

Phase 2: Augmentation Parameter Mapping

his phase bridges the gap between the textual descriptions generated by the LLM and the concrete parameters required to perform data augmentation. A mapping model will be developed to translate the LLM-generated descriptions into a set of parameters for different augmentation techniques. This model will take the textual description as input and output a vector of parameters that control aspects such as the degree of rotation, the size of the crop, the intensity of color jittering, and the temporal displacement. Different architectures for this mapping model will be explored, including recurrent neural networks (RNNs), which are well-suited for processing sequential data like text, and transformers, which have demonstrated superior performance in capturing long-range dependencies in language. The mapping model will be trained on a dataset of paired descriptions and corresponding augmentation parameters. This dataset will be generated through a combination of manual annotation, where human annotators manually associate descriptions with specific parameter values, and programmatic generation, where rules based on keywords or phrases in the descriptions are used to automatically generate parameter values. The quality of the mapping will be evaluated by assessing the realism and diversity of the resulting augmented samples. Perceptual similarity metrics will be employed to quantify the similarity between the augmented samples and real video data.

Phase 3: Data Augmentation and Action Recognition Training

he generated augmentation parameters will be used to apply transformations to the limited training examples in the few-shot setting. A variety of augmentation techniques will be utilized, including standard transformations such as cropping, rotation, flipping, and jittering (color, temporal), as well as advanced techniques such as Mixup and CutMix, adapted for video data. The possibility of using GANs or diffusion models conditioned on the LLM descriptions to generate entirely new synthetic video segments will also be explored. For few-shot action recognition, a meta-learning approach, specifically MAML, will be employed. The model architecture will be based on a convolutional neural network (CNN) designed for video processing, such as a 3D-CNN or a two-stream network. The model will be trained on the augmented dataset using standard

optimization algorithms like Adam. Performance will be evaluated using the N-way K-shot accuracy metric, as well as precision, recall, and F1-score, to provide a comprehensive assessment of the model's ability to generalize from limited examples.

Timeline

Phase	Description	Duration (months)
1. Literature Review	Review existing research and identify research gaps	2
2. Data Collection	Collect and preprocess datasets and English speech datasets	3
3. Model Development	Design and implement the architecture	4
4. Training & Fine-Tuning	Train and fine-tune the model on the collected datasets	3
5. Evaluation	Conduct subjective and objective evaluations	2
6. Writing & Reporting	Document findings and prepare research papers or technical reports	2

Expected Outcomes

This research anticipates several key outcomes that will advance the field of few-shot action recognition. Firstly, we expect a quantifiable improvement (10-15%) in few-shot action recognition accuracy on benchmark datasets like Something-Something-V2 and HMDB51, compared to state-of-the-art methods using standard random data augmentation. Secondly, the research will deliver a novel framework for LLM-guided data augmentation, encompassing methodologies for action description generation, parameter mapping, and data augmentation. This framework is expected to be generalizable to other few-shot learning tasks in computer vision. Thirdly, this work aims to provide a deeper understanding of the role of semantic information in data augmentation for few-shot learning, informing the development of more effective augmentation strategies. Finally, the research findings will be disseminated through publications in top-tier computer vision conferences and journals.

References

[1] Tom Brown et al. "Language models are few-shot learners". In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.

[2] Zhicheng Cai, Chenglei Peng, and Sidan Du. "Jitter: random jittering loss function". In: *2021 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2021, pp. 1–8.

[3] Minshuo Chen et al. "An overview of diffusion models: Applications, guided generation, statistical rates and optimization". In: *arXiv preprint arXiv:2404.07771* (2024).

[4] Na Chen et al. "Few-shot image classification based on gradual machine learning". In: *Expert Systems with Applications* 255 (2024), p. 124676.

[5] Jacob Devlin. "Bert: Pre-training of deep bidirectional transformers for language understanding". In: *arXiv preprint arXiv:1810.04805* (2018).

[6] Chelsea Finn, Pieter Abbeel, and Sergey Levine. "Model-agnostic meta-learning for fast adaptation of deep networks". In: *International conference on machine learning*. PMLR. 2017, pp. 1126–1135.

[7] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems* 27 (2014).

[8] Raghav Goyal et al. "The" something something" video database for learning and evaluating visual common sense". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 5842–5850.

[9] Will Kay et al. "The kinetics human action video dataset". In: *arXiv preprint arXiv:1705.06950* (2017).

[10] Hildegard Kuehne et al. "HMDB: a large video database for human motion recognition". In: *2011 International conference on computer vision*. IEEE. 2011, pp. 2556–2563.

[11] Alec Radford et al. "Learning transferable visual models from natural language supervision". In: *International conference on machine learning*. PMLR. 2021, pp. 8748–8763.

[12] Haoyu Ren, Darko Anicic, and Thomas A Runkler. "TinyReptile: TinyML with federated meta-learning". In: *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE. 2023, pp. 1–9.

[13] Flood Sung et al. "Learning to compare: Relation network for few-shot learning". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 1199–1208.

[14] Sangdoo Yun et al. "Cutmix: Regularization strategy to train strong classifiers with localizable features". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 6023–6032.

[15] Hongyi Zhang. "mixup: Beyond empirical risk minimization". In: *arXiv preprint arXiv:1710.09412* (2017).