Skeleton based Action Recognition using Learnable Textual Inputs

Masters Thesis - Research Project Proposal

1. Description

Action Recognition is one of the fundamental tasks in video understanding. A number of works have focused on exploring different modalities for learning discriminative features such as optical flows [7], RGB frames [2], and human skeletons [8]. In the recent years, representing the human body as a skeleton for recognizing an action has received increasing attention as human skeletons are in-general mildly affected by changes in appearance / lighting conditions. In practice, one can represent human skeletons in a video as a sequence of inter-connected joints (spatial and temporal connections); resulting in a compact and well-connected graph; which are immune to contextual changes such as background lighting conditions.

One natural and plausible implementation of such an interconnected skeleton is to represent the skeletons as a graph with the nodes representing the detected 2D key-points and the edges representing the connection between neighboring nodes both in the spatial and temporal domain respectively. This 2D graph is further processed using a Graph Convolutional Network (GCN), which attempts to holistically learn a representation as a form of an embedding; that encapsulates the changes both in the spatial and temporal domain. A number of works have been undertaken in this direction [9].

However, none of the aforementioned methods have use any available textual knowledge in order to improve the generalization ability of the underlying networks. We believe the that textual prompts certainly consist of valuable contextual information that can be imbibed into the graph based GCNs to enhance its discrimination and generalization ability; such that it can attempt to assimilate more contextual features from the individual frames in accordance with the textual prompts for improved action recognition.

Motivation: Therefore, in this project our major goal is to improve the generalization ability of the Action Recognition systems by incorporating contextual knowledge using image-text captions. In order to achieve the above stated goal, we attempt to use textual prompts which will be processed using a pre-trained text encoder [4]. We hypothesize this by incorporating such textual knowledge distillation so that one can successfully learn contextual information from the input videos in order to recognize an action with an enhanced discrimination ability. We will thus follow the learning protocol of [10, 11] and incorporate learnable textual prompts to facilitate the learning of important cues for recognizing an action.

2. Tackling the problem step by step:

Steps **1** to **3** are necessary steps for successful completion of the project; Step **4** is a possible extension provided we have achieved the objectives of Steps **1** to **3**.

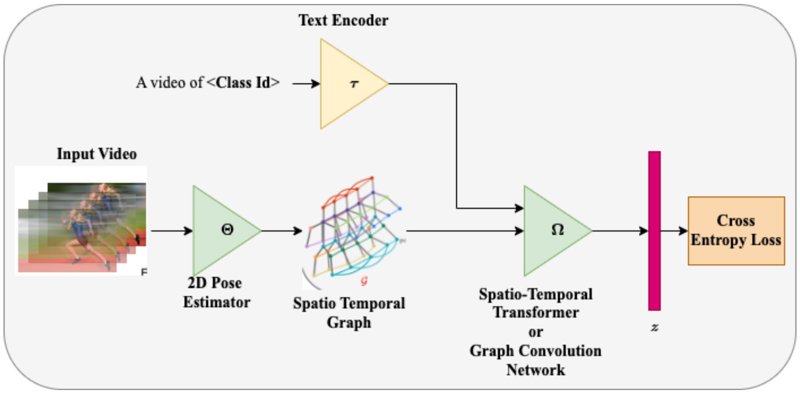
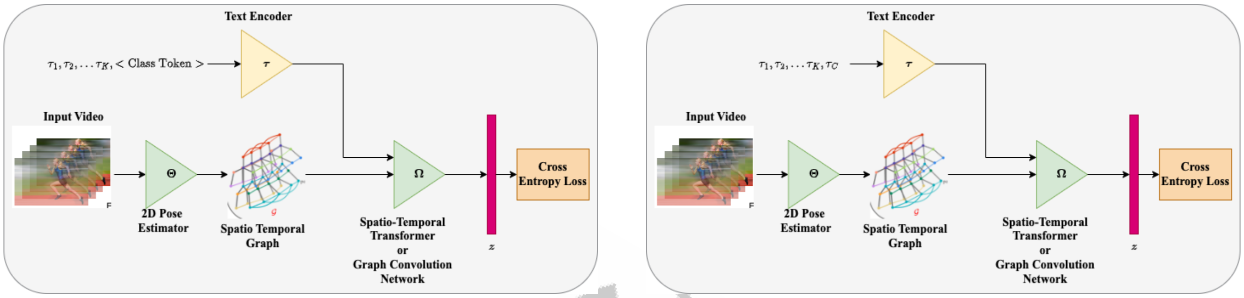


Figure 1. A overview of the proposed method for Step 2.



(a) Learnable "*Contextual Vectors*". (b) Learnable "*Contextual and Class Specific Vectors*".

Figure 2. A overview of the proposed method with learnable contextual vectors $\tau_1, \tau_2, ...\tau_K$ (Step 3).

1. Obtain 2D joints (or 2D poses) on 2D frames of the video using any off-the-shelf 2D pose estimator Net_Θ [3,6]. These 2D skeletons can be grouped into a spatial-temporal graph where each joint is connected to its spatial and temporal neighbors. This graph will be processed by a spatio-temporal GCN [9] or a Temporal Transformer Network (Net_Ω), which will obtain the final embedding of the entire image sequence, which will be further converted in classification probabilities using a standard classifier layer. Training loss will be the cross entropy loss. Other losses such as OTAM [1] can also be investigated.

2. We dive into using the textual prompt by using the following textual prompt: "A video of <Class Token>", <Class Token> denotes the class label of the action. This textual prompt will be processed by a text encoder Net_Γ to produce the final textual embedding output $\boldsymbol{t}$. We will investigate the use of $\boldsymbol{t}$ as an conditional vector to generate the embedding features using Net_Ω [4]. One simple yet reliable method of conditioning is using the cross-attention module proposed in [5] to obtain the graph embeddings conditioned on $\boldsymbol{t}$. A overview of the proposed work is shown in Fig. 1.

3. Replace the class tokens with K different learnable prompt vectors $\boldsymbol{\tau} = \{\tau\}_1^K$ similar to [10, 11] in Step 2, in order to learn to assimilate the contextual knowledge using $\boldsymbol{\tau}$. **Successful implementation of this step is the key to the completion of the masters thesis project.** We will try three different variants of this step:

  - We first begin with the prompt of the following order $\{\tau_1, \tau_2 \cdots \tau_K, <\text{ Class Token }>\}$, where the <Class Token> is fixed while $\boldsymbol{\tau}$ is learnt (Refer to Figure 2a for an overview).
  - Next, we will try to learn the class specific token vectors instead of the fixed token <Class Token>. The prompt will be of the form $\{\tau_1, \tau_2 \cdots \tau_K, \tau_c\}$, where c denotes the class of action performed in the video (Refer to Figure 2b for an overview).
  - Thereafter, we will look into the possibility of using a prompt of the following form $\{\tau_1, \tau_2 \cdots \tau_K,$ "a type of action."$\}$ into the learning framework; which removes the dependence on classification labels. Thereby making the overall algorithm a form of **un-supervised** learning (We will replace the $<$ Class Token $>$ in Figure 2a with the tokens of "a type of action."; and an appropriate pretext task for unsupervised learning will be looked into).

4. Possible extension to Zero/Few shot action recognition regimes (where the need of contextual information through textual inputs deems to be beneficial).

Note: We will be first freezing the parameters of the networks $\text{Net}_{\boldsymbol{\Theta}}$ and $\text{Net}_{\boldsymbol{\Gamma}}$ and only learn the parameters of $\text{Net}_{\boldsymbol{\Omega}}$ in step **1** and the learnable textual-contextual vectors $\boldsymbol{\tau}$ in step **3**.

3. Prerequisites

- This project is offered to Master's students for their **thesis** project.
- The candidate should have Python programming experience.
- Previous experience with deep learning and PyTorch, is recommended.
- Knowledge of Human Action Recognition or (and) Learning with Vision Language Models is a **huge** plus.

Contact:

Soumava Kumar Roy, EPFL Switzerland (soumava.roy@epfl.ch).

References

[1] Kaidi Cao, Jingwei Ji, Zhangjie Cao, Chien-Yi Chang, and Juan Carlos Niebles. Few-shot video classification via temporal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10618–10627, 2020. 2

[2] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 1

[3] Yilun Chen, Zhicheng Wang, Yuxiang Peng, Zhiqiang Zhang, Gang Yu, and Jian Sun. Cascaded pyramid network for multi-person pose estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7103–7112, 2018. 2

[4] Bo Dai and Dahua Lin. Contrastive learning for image captioning. *Advances in Neural Information Processing Systems*, 30, 2017. 1, 2

[5] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 2

[6] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, pages 483–499. Springer, 2016. 2

[7] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014. 1

[8] Philippe Weinzaepfel and Grégory Rogez. Mimetics: Towards understanding human actions out of context. *International Journal of Computer Vision*, 129(5):1675–1690, 2021. 1

[9] Sijie Yan, Yuanjun Xiong, and Dahua Lin. Spatial temporal graph convolutional networks for skeleton-based action recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018. 1, 2

[10] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2

[11] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision (IJCV)*, 2022. 1, 2