HEART-MET Assistive Robot Challenge with VideoMAE Transformers

Olmer Garcia-Bedoya
Ekumen Inc.
Bogota, Colombia
ORCID 0000-0002-6964-3034
olmerg@ekumenlabs.com

Jose Tomas Lorente
Ekumen Inc.
Buenos Aires, Argentina
jtlorente@ekumenlabs.com

Sebastian Murcia
Ekumen Inc.
Bogota, Colombia
sebastian.murcia@ekumenlabs.com

***Abstract*—Robotics in healthcare has been potential to help in the quality of life of many people. This article present the methodology used during METRICS HEART-MET Assistive Robot Challenge at ICSR 2022 to recognition of a set of activities in videos. The methodology consists of a fine-tuning process from the VideoMAE model based in the data of the challenge. The results give a 0.7 of accuracy over the validation and test data set which were evaluated by codelab platform. We found that some challenges is that the data is unbalanced and some videos could have more than one categories, which should be taken account before to try improve the results.**

***Index Terms*—Robotics, Vision Transformer, human activity, deep learning**

I. Introduction

This report presents the methodology and some conclusions learned during the icsr-2022 HEART-MET Assistive Robot Challenge. The task in the challenge is to recognize human activities from videos. The videos are recorded from robots operating in a domestic environment and include activities such as reading a book, drinking water, falling on the floor, etc. HEART-MET is one of the competitions in the METRICS project, which has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 871252. The competition aims to benchmark assistive robots performing healthcare-related tasks in unstructured domestic environments.

Ekumen Inc is an international engineering boutique, provider of advanced software development services and technology. We specialize in bridging the gap between scientific research and deployable software products, with experience in open source projects like ROS. We specialize in the following areas: Robotics Software Applications, Web and Mobile Technology, Embedded Systems and augmented reality applications.

After getting a lot of data successfully in natural language processing (NLP) by self-supervised learning. The solutions, based on autoregressive language modeling in GPT [1] and masked autoencoding in BERT [2] , are conceptually simple: they remove a portion of the data and learn to predict the removed content. Application of transformer in images start with [3], which outperform the current state-of-the-art (CNN) by almost x4 in terms of computational efficiency and accuracy [4], where Encoding: key,value(position in image of subimage, subimage). After this concept Facebook proposed [5] which is the base of [6] [7] the model used during this challenge. In the next section is presented some insight about the methodology and some conclusions

II. Methodology

In hugginFaces [8] videoMAE is a transformer with video classification, like other transformers it is trained in two stages. The first stage called pre-taining is created by an encoder and a decoder which require a lot computational power. The input receives a pipeline of masking random cubes take from video composed by 16 frames by 16 fixed-size patches of each image frame over a video of resolution 224x224. The output of the decoder receives the complete video with the idea of "challenging self-supervisory tasks that require holistic understanding" [6]. The second stage, called fine tuning, consists in a dense layer with input the output of the encoder and the output the number of classification classes for the video.

In huggingFace, videoMAE is a PyTorch torch.nn.Module subclass. The hugging face data and model come from the authors repository, but in our case we start from the fine tuning model present in HuggingFace of Multimedia Computing Group-Nanjing University. Specifically, we start making the fine tuning process of the last layer of the model *videomae-base-finetuned-kinetics [9]*. In the next subsections we present our conclusion about the methodology.

A. Data exploration

The data exploration was motivated by the need to improve the performance of the detection of some specific labels. Consequently, we proceeded to analyze the training data by counting the total number of videos for each label [Figure 1].

Figure 1 shows that there are seven labels (Drinking water, Eating food with a fork, reading a book, talking on the phone, using a computer, Brushing teeth and writing) with a notably higher number of videos than the other labels, resulting in the average number of videos per label in the training data being equal to 39.6 and the standard deviation equal to 25.53 videos, leading to the conclusion that the training data were highly

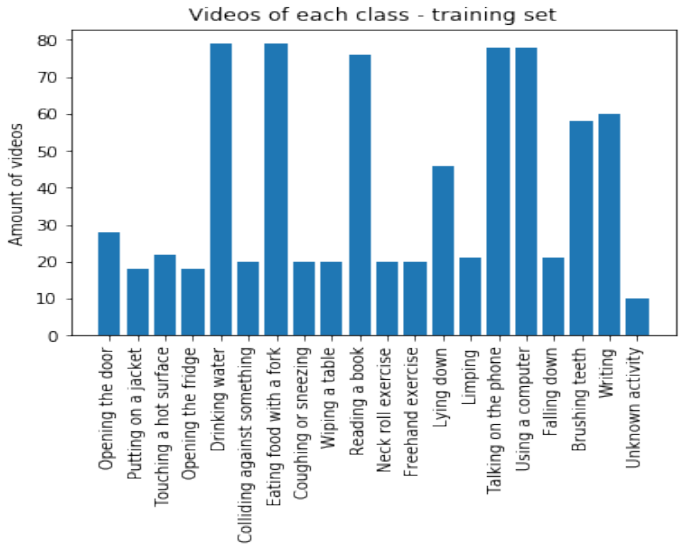

Fig. 1. **Amount of videos for each Label - training set**

unbalanced. Additionally, some conclusion found in this stage was that the number of frames is different in the videos.

B. *Sampling the video*

The hugging face examples suggest to sample 16 frames sampling each 4 frames, with the initial frame a random number. Although it was our first attempt, the fine tuning was not successful. The second approach (which gave the best result) was to take 16 frames with a sampling determined by the length of the video over 16.

Between the sampling method tested , we try to start randomly over the first x frames , try to change a small random change in the sampling period, this give similar results although we increase the number of epochs.

Other approach we tried, which ended up yielding similar results, but less dependant on randomly starting on the right frame, was dividing the total frame of the video in 16, and then slicing the video into frames/16 smaller videos, with frames/16 index jumps in between, starting from the frame 0 to the frames/16 frame. Resulting in 16 frames lengthed videos, all with frames equally distant from each other.

C. *Fine Tuning training*

Hugging face does not have examples of fine tuning of Video Transformer so we adapted an example of fine tuning of Vision Transformer. Initially, we had problems training the model because the batch size of 2 videos required more than 8GB of RAM, which made it difficult to train with our computers. We found that we had not frozen the encoder layers to execute the training (86.242.580 parameters), so after this change we can increase to batches of 16 videos with around 8GB of ram required in the GPU, which let us to run more than one test in the same time in our server.

Here we test different hyperparameters without any significant change. We also test some modification in the model like :

- Freezing the fully connected normalization layer before the classifier layer gave better results than not freezing it.
- After augmenting the dataset by dividing the videos into smaller ones, a smaller learning rate and no weight decay made training more stable and allowed the training cost and the validation cost to move similarly

D. *Validate*

During analysis of results we use scikit learn metrics to generate confusion matrix and identify problems in the classification. We conclude that we require more data in categories , which we are planning to synthesize with video augmentation strategies. Review specific videos and try to understand why it is challenging this category, maybe try to balance the data or augment the data.

III. Discussion

The final result during the challenge was 0.7 of accuracy over the test and validation data in the Codalab platform [10]. The most interesting characteristic of VideoMAE approach was the speed in the fine tuning part, because with less than 30 epoch (which takes around 50 minutes in a NVIDIA 3090TI) we get an 0.68 over the validation dataset. This contrasts with the results that we obtain with the base model (based in the fine-tuning RGB Charades model of [11]) which take around this time by each epoch. Although we tested different approaches to improve the results we did not find any solution, we think that the principal problem comes from the data which is unbalanced , and many videos could be classified in different classes. Adding some sort of memory neuron to the model, or some sort of information about like optical flow, to convey more information about the video than only 16 frames could have also improved the result.

We also tested X-CLIP [12] through hugging Faces (**X-CLIP HuggingFace, X-CLIP Github**) but gave 36% accuracy over the training dataset without any training process. We think this could be an interesting approach increasing the description of the categories taking into account classes of the kinetics, because it can give an initial classification for labeling the dataset.

Next steps, after try to balance the data , include start the pretraining from the large model of kinetics-400 or from another pre-trained model available (Something-Something V2 or AVA 2.2), change from half precision to double precision.

IV. Acknowledge

We acknowledge the support of Michel Hidalgo like head of the R&D group of the company during the Challenge.

References

[1] L. Floridi and M. Chiriatti, "Gpt-3: Its nature, scope, limits, and consequences," *Minds and Machines*, vol. 30, no. 4, pp. 681–694, 2020.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[4] G. Boesch, "Vision transformers (vit) in image recognition - 2022 guide," Aug 2022. [Online]. Available: https://viso.ai/deep-learning/vision-transformer-vit/

[5] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 16 000–16 009.

[6] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *arXiv preprint arXiv:2203.12602*, 2022.

[7] ——, "VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training," in *Advances in Neural Information Processing Systems*, 2022.

[8] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Huggingface's transformers: State-of-the-art natural language processing," 2019. [Online]. Available: https://arxiv.org/abs/1910.03771

[9] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev *et al.*, "The kinetics human action video dataset," *arXiv preprint arXiv:1705.06950*, 2017.

[10] A. Pavao, I. Guyon, A.-C. Letournel, X. Baró, H. Escalante, S. Escalera, T. Thomas, and Z. Xu, "Codalab competitions: An open source platform to organize scientific challenges," *Technical report*, 2022. [Online]. Available: https://hal.inria.fr/hal-03629462v1

[11] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," 2017. [Online]. Available: https://arxiv.org/abs/1705.07750

[12] B. Ni, H. Peng, M. Chen, S. Zhang, G. Meng, J. Fu, S. Xiang, and H. Ling, "Expanding language-image pretrained models for general video recognition," 2022.