An evaluation of large pre-trained models for gesture recognition using synthetic videos

Arun Reddy[a,b], Ketul Shah[b], Corban Rivera[a], William Paul[a], Celso M. De Melo[c], and Rama Chellappa[b]

[a]Johns Hopkins University Applied Physics Laboratory, Laurel, MD, USA
[b]Johns Hopkins University, Baltimore, MD, USA
[c]Army Research Laboratory, Los Angeles, CA, USA

ABSTRACT

In this work, we explore the possibility of using synthetically generated data for video-based gesture recognition with large pre-trained models. We consider whether these models have sufficiently robust and expressive representation spaces to enable "training-free" classification. Specifically, we utilize various state-of-the-art video encoders to extract features for use in k-nearest neighbors classification, where the training data points are derived from synthetic videos only. We compare these results with another training-free approach— zero-shot classification using text descriptions of each gesture. In our experiments with the RoCoG-v2 dataset, we find that using synthetic training videos yields significantly lower classification accuracy on real test videos compared to using a relatively small number of real training videos. We also observe that video backbones that were fine-tuned on classification tasks serve as superior feature extractors, and that the choice of fine-tuning data has a substantial impact on k-nearest neighbors performance. Lastly, we find that zero-shot text-based classification performs poorly on the gesture recognition task, as gestures are not easily described through natural language.

Keywords: gesture recognition, action recognition, video classification, video-language, synthetic data

1. INTRODUCTION

One of the promises of synthetic data is the possibility of reducing reliance on real data for training deep learning models, which can introduce practical challenges and ethical concerns. As such, there has been a growing interest in using synthetically generated video data to train models for various video-related tasks. However, previous work has shown the existence of a large domain gap between real and synthetic video data, resulting in sub-optimal performance when naively applying a synthetically-trained model to data from the real domain.[1–4] This issue has spurred the development of various approaches for video domain adaptation,[3,5,6] which can be computationally expensive and difficult to implement in practice.

Here, we consider whether state-of-the-art video backbones, given the scale of their pre-training, are capable of extracting domain-invariant representations to enable video classification without needing any real data for the task. Specifically, we experiment with two different "training-free" approaches. The first uses modern video backbones as feature extractors for K-nearest neighbors (KNN) classification, where the training samples are derived from synthetic videos. In the second, we rely on text descriptions of the classes and attempt to perform zero-shot classification using similarity between video features and text features from each class description. We use video-based gesture recognition as our video classification task in this study.

2. EXPERIMENTS

We perform two sets of experiments. In the first set, we leverage large pre-trained video models as feature extractors and perform KNN classification using these features. In the second, we perform zero-shot classification using textual descriptions of gestures. These two methods are described in detail below. All experiments are performed on ground viewpoint videos from the RoCoG-v2 dataset,[1] examples of which are shown in Figure 1.

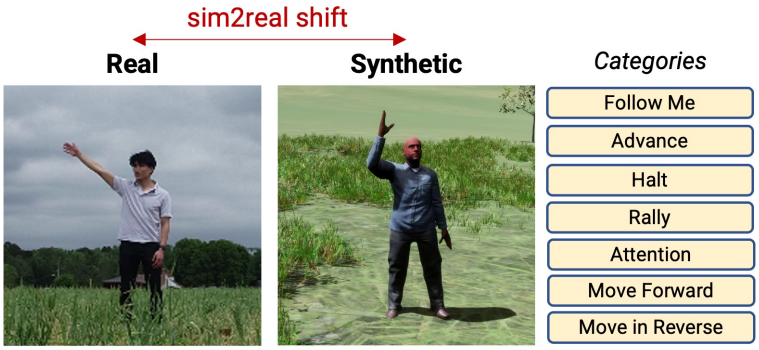


Figure 1: Examples of a real (left) and synthetic (right) video from RoCoG-v2. The datset consists of 7 gesture categories.

2.1 KNN Classification

When using synthetic training videos to specify the gesture recognition task, we perform KNN classification on features extracted from large pre-trained models. More specifically, the synthetic training data (44K videos) is embedded in the feature space of a video encoder, and for a given real test video, the majority vote of the K nearest training data points is used for classification, based on L2 distance. We also experiment with a scenario where we have a small real dataset (203 videos) available for training. We set K=3 for all experiments. We consider three different types of video encoders based on pre-training (and fine-tuning) strategies, as described below. We choose a ViT-B/16 model for all experiments, and also study the effect of a larger ViT-L/16 model for the best setting.

Self-Supervised Pre-Training. We consider Unmasked Teacher (UMT),[7] which is a state-of-the-art video self-supervised learning approach. This approach masks out most of the video tokens and enforces alignment between the representations of unmasked patches and the corresponding ones from a teacher model (CLIP[8]). We use the UMT model pre-trained on K710 videos (a union of K400,[9] K600[10] and K700[11]) for our experiments. Eight frames are sampled from each video using the TSN[12] frame-sampling strategy. The entire video is divided into eight segments, and one frame is selected at random from each segment. The input frames to the network are resized to 224 $\times$ 224 resolution.

Vision-Language Pre-Training. In contrast to UMT which uses only video data for pre-training, here we consider a vision-language pre-training approach of ViCLIP.[13] ViCLIP uses a video-language contrastive objective similar to CLIP,[8] while also masking videos for efficient pre-training. We use the model pre-trained on a filtered version of the InternVid[13] dataset that has 10M video-text pairs. Eight frames of 224 $\times$ 224 resolution are used as input to the network.

Self-Supervised Pre-Training + Supervised Fine-Tuning. Here, we use self-supervised models which were further fine-tuned for video classification in a supervised manner. We consider two pre-training methods, UMT[7] and VideoMAE.[14] VideoMAE is a powerful self-supervised pre-training approach which works by encoding partially masked inputs and reconstructing the masked out regions. The VideoMAE models are pre-trained on a larger (1.35M) UnlabeledHybrid[15] dataset, whereas UMT models are pre-trained on the K710 dataset (650K). These models are either fine-tuned on Kinetics[9] (K710, K400, K600, K700) or the more motion-centric Something-Something-v2 (SSv2)[16] dataset. For the VideoMAE models, we sample sixteen frames from the input video, with a tubelet size of two frames, resulting in the same number of tokens as using eight frames with a tubelet size of one frame, as in the above scenarios.

2.2 Zero-Shot Text-Based Classification

The gesture recognition task can alternatively be specified by providing the textual descriptions of each activity. In our second set of experiments, shown in Table 1, we use text descriptions to perform zero-shot classification using pre-trained vision-language models. Specifically, we use the pre-trained ViCLIP[13] text and video encoders,

Type	Backbone	Pre-Training Method	Pre-Training Data	Fine-Tuning Data	Real Test KNN Acc. (%)	
					Synthetic Train	Real Train
Self-Supervised Pre-Training	ViT-B/16	UMT	K710	-	18.2	31.2
Vision-Language Pre-Training	ViT-B/16	ViCLIP	InternVid FLT-10M	-	19.2	40.4
Self-Supervised Pre-Training + Supervised Fine-Tuning	ViT-B/16	UMT	K710	K710	42.4	49.5
	ViT-B/16	UMT	K710	K710 + K400	38.4	45.5
	ViT-B/16	UMT	K710	K710 + K600	33.3	49.5
	ViT-B/16	UMT	K710	K710 + K700	35.4	51.5
	ViT-B/16	VideoMAE	UnlabeledHybrid	K710	32.3	60.6
	ViT-B/16	VideoMAE	UnlabeledHybrid	SSv2	43.4	68.7
	ViT-L/16	VideoMAE	UnlabeledHybrid	SSv2	64.6	71.7

Table 1: K-nearest neighbor classification results on RoCoG-v2 ground videos using a variety of video backbones.

where classification is performed based on similarity between video features and text embeddings of descriptions of all classes. We use two kinds of text descriptions of the gestures, original and transformed, as follows:

Original. These are the instructions associated with each gesture as they appear in the US Army Field Manual.[17] These were provided to the performers of the gestures in the RoCoG-v2 dataset.

Transformed. We use GPT-3.5 to turn the original text instructions into gesture descriptions by prompting the model with the following text: "*Can you summarize the description below of an activity instruction and start with "A person"*".

Model	Training Data	Text Description	Test Accuracy (%)
ViCLIP-B	InternVid FLT-10M	Original	25.3
ViCLIP-B	InternVid FLT-10M	Transformed	26.3

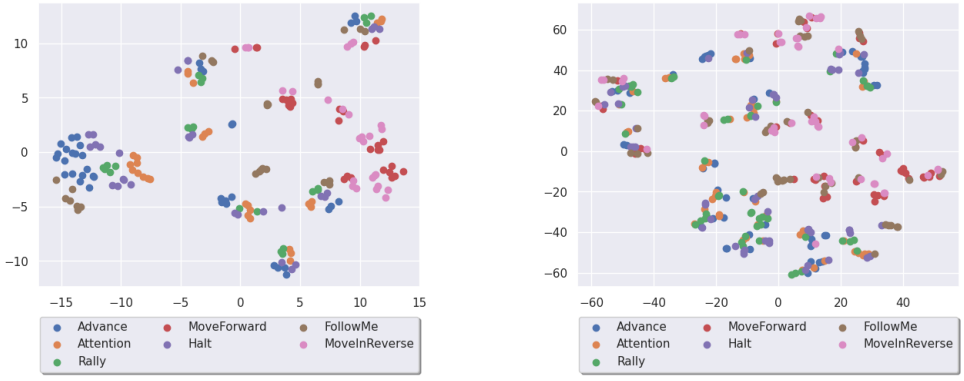
Table 2: Zero-shot classification on RoCoG-v2 ground videos using two forms of text descriptions.

3. DISCUSSION

Several observations can be made from the KNN classification results in Table 1. First, we can clearly see that, in all cases, using synthetic training data results in lower accuracy on real videos than using real training data. Despite the large quantity of synthetic training data (roughly 200× that of real training data), we find it is not as effective at defining the gesture classes in the RoCoG-v2 dataset as real data. This is indicative of the synthetic-to-real domain gap that has been observed previously, which is still far from solved. Figure 2 also illustrates this domain gap, as real video features appear noticeably more clustered by gesture class than synthetic video features.

Next, we can see from comparing the first two rows in Table 1 that large-scale vision-language pre-training seems to confer some benefit over smaller-scale masked pre-training, particularly when using real videos for KNN classification. However, we find that backbones that have undergone supervised fine-tuning vastly outperform both the self-supervised and vision-language pre-trained backbones. Interestingly, as the backbone is fine-tuned on more real videos, we see a drop in KNN accuracy using synthetic training data while accuracy using real training data increases. This suggests that video transformer backbones become less robust to the synthetic-to-real domain shift the more they are trained on real videos.

We find that the choice of fine-tuning data has a substantial impact on KNN classification. We can see that backbones fine-tuned on the SSv2 dataset[16] perform much better than those trained on Kinetics videos. SSv2 is a temporally-heavy dataset, where modeling of motion is critical for solving the classification task. In contrast, the action categories in Kinetics videos exhibit a high degree of object and scene bias. Because the gesture recognition task in RoCoG-v2 is also motion-focused, SSv2 serves as an effective source of fine-tuning data. Notably, scaling up from the ViT-B model to ViT-L results in significant boosts in KNN accuracy, particularly when using synthetic training data. Future experiments should investigate whether, in general, larger models might be more robust to synthetic-to-real shifts.



(a) t-SNE plot for real data features.

(b) t-SNE plot for synthetic data features.

Figure 2: t-SNE plots for real and synthetic data. Real data is more meaningfully clustered compared to synthetic data, as the features are extracted using a ViT-B/16 model pre-trained on K710 and fine-tuned on K710 and K400. For (a), we use all real data whereas for (b), we use 50 samples per class chosen at random from the synthetic dataset.

Finally, we observe in Table 2 that zero-shot classification using ViCLIP[13] performs poorly on the gesture recognition task, regardless of the type of text description used. This is likely because the gesture recognition task involves fine-grained motion differences that are not easily expressed through natural language.

ACKNOWLEDGMENTS

This research was sponsored by the Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-21-2-0211. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] Reddy, A. V., Shah, K., Paul, W., Mocharla, R., Hoffman, J., Katyal, K. D., Manocha, D., de Melo, C. M., and Chellappa, R., "Synthetic-to-Real Domain Adaptation for Action Recognition: A Dataset and Baseline Performances," in [*2023 IEEE International Conference on Robotics and Automation (ICRA)*], 11374–11381 (2023).

[2] Shah, K., Shah, A., Lau, C. P., de Melo, C. M., and Chellappa, R., "Multi-view action recognition using contrastive learning," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 3381–3391 (2023).

[3] Da Costa, V. G. T., Zara, G., Rota, P., Oliveira-Santos, T., Sebe, N., Murino, V., and Ricci, E., "Dual-head contrastive domain adaptation for video action recognition," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*], 1181–1190 (2022).

[4] Katyal, K., Chellappa, R., Shah, K., Reddy, A., Hoffman, J., Paul, W., Mocharla, R., Handelman, D., and de Melo, C., "Leveraging synthetic data for robust gesture recognition," in [*Synthetic Data for Artificial Intelligence and Machine Learning: Tools, Techniques, and Applications*], **12529**, 238–241 (2023).

[5] Reddy, A., Paul, W., Rivera, C., Shah, K., de Melo, C. M., and Chellappa, R., "Unsupervised Video Domain Adaptation with Masked Pre-Training and Collaborative Self-Training," (2024).

[6] Wei, P., Kong, L., Qu, X., Ren, Y., Xu, Z., Jiang, J., and Yin, X., "Unsupervised video domain adaptation for action recognition: A disentanglement perspective," in [*Advances in Neural Information Processing Systems*], (2023).

[7] Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., and Qiao, Y., "Unmasked Teacher: Towards Training-Efficient Video Foundation Models," in [*ICCV*], (2023).

[8] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al., "Learning transferable visual models from natural language supervision," in [*International conference on machine learning*], 8748–8763, PMLR (2021).

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

[10] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., and Zisserman, A., "A Short Note about Kinetics-600," (2018).

[11] Carreira, J., Noland, E., Hillier, C., and Zisserman, A., "A short note on the kinetics-700 human action dataset," *arXiv preprint arXiv:1907.06987* (2019).

[12] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., and Van Gool, L., "Temporal segment networks for action recognition in videos," *IEEE transactions on pattern analysis and machine intelligence* **41**(11), 2740–2755 (2018).

[13] Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Chen, X., Wang, Y., Luo, P., Liu, Z., Wang, Y., Wang, L., and Qiao, Y., "InternVid: A Large-scale Video-Text Dataset for Multimodal Understanding and Generation," in [*ICLR*], (2024).

[14] Tong, Z., Song, Y., Wang, J., and Wang, L., "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training," in [*NeurIPS*], (2022).

[15] Wang, L., Huang, B., Zhao, Z., Tong, Z., He, Y., Wang, Y., Wang, Y., and Qiao, Y., "Videomae v2: Scaling video masked autoencoders with dual masking," in [*Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*], 14549–14560 (2023).

[16] Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., et al., "The" something something" video database for learning and evaluating visual common sense," in [*Proceedings of the IEEE international conference on computer vision*], 5842–5850 (2017).

[17] "Visual signals: Field manual 21-60," (1987).