Supplementary material for ReGen: A good Generative zero-shot video classifier should be Rewarded

Adrian Bulat[1,2], Enrique Sanchez[1], Brais Martinez[1], Georgios Tzimiropoulos[1,3]
[1]Samsung AI Cambridge [2]Technical University of Iasi [3]Queen Mary University of London

A. Additional experimental details

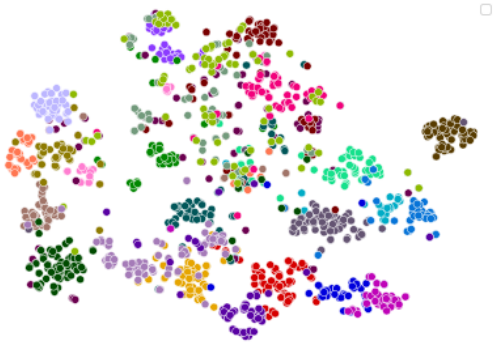
A.1. Datasets

Kinetics-400 [3] is a large scale video-recognition dataset consisting of short clips, collected from YouTube and annotated with one of the 400 defined classes, covering diverse human activities. Although some of the original videos are no longer available, the datasets consist of around 0.241M training and 20,000 validation videos.
Kinetics-600 [1] is an extension of the former to 600 classes. For evaluation purposes, we use the (three) splits defined in [2]. Each split consists of 160 novel classes, not present in Kinetics-400, covering 220 classes in total across the 3 splits. To ensure there is no overlap, the classes were renamed from 600 to 620. We refer to this evaluation subset as Kinetics-220 in the zero-shot case and as 620 in the generalized zero-shot one.
HMDB-51 [4] is a temporal-context sensitive dataset consisting of 6849 clips divided into 51 action categories, each containing a minimum of 101 clips. The dataset is divided into 3 pairs of train-test subsets.
UCF-101 [6] is an action recognition data set containing 13320 videos from 101 action categories. Similarly with HMDB-51, the dataset is split into 3 test-train partitions.
Vatex [7] is a bilingual (English and Chinese) video captioning dataset containing 41,250 videos and 825,000 captions. We used its 6000 videos public test to quantitatively ensure that our model produces text that remains coherent. No training is performed on it.

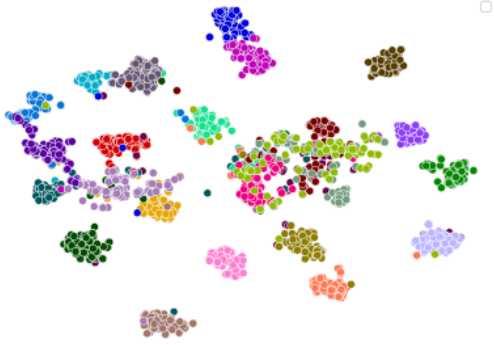
A.2. Training hyperparameters

We use the same data augmentation transformations during ReGen RL training and the few-shot fine-tuning, mainly: Random Flip (0.5), Multi-scale crop (0.66, 0.75, 0.875, 1.0), Color jitter (0.8) and Gray scale (0.2). In terms of training hyperparameters, for ReGen RL training we use: AdamW optimizer ($\beta_1 = 0.9, \beta_2 = 0.98$), a learning rate of $2.5e-6$, weight decay of 0.01, a cosine learning rate scheduler with the final rate of $2.5e-8$ and a total training duration of 10 epochs. We use the same hyperparameters for the few shot fine-tuning, except that the number of epochs is adjusted based on the number of samples per class, K: $\max(400/K, 30)$.

B. Further ablations



(a) t-SNE: REST



(b) t-SNE: ReGen

Figure 1: t-SNE plot of the generated pseudo-captions in CLIP text embedding space for a subset of randomly chosen novel classes for (a) REST, and (b) ReGen. Our method results in much more discriminative features.

t-SNE of ReGen vs REST: To illustrate the effect of ReGen training in terms of producing more discriminative captions compared to REST baseline, Fig. 1 shows the t-SNE plot of the generated pseudo-captions in CLIP text embedding

space for a subset of randomly chosen novel classes for both methods. Clearly, ReGen produces significantly more clustered and hence more discriminative features.

Similarity between the predictions and the video: The `CLIP-R` reward encourages the predicted text to reflect the content of the input video sample. To demonstrate this in practice, in Fig. 2, we plot the cosine similarities for a model trained with `CLS-R` only vs a model trained with all rewards (i.e. ReGen). As expected, the addition of the `CLIP-R` reward increases the cosine score.

Impact of the reward loss type of CLS-R: Beyond the Cross-Entropy (CE) loss considered in the main manuscript, herein we ablate its formulation, considering the following additional cases: a direct L_2 loss between the text features and a cosine similarity loss between the produced textual features t and the ground truth ones t_y obtained using the class names. As the results from Table 1 show, the CE loss outperforms both variants.

LM	HMDB-51	UCF-101
CE	55.1	76.4
cos(t_y, t)	52.2	73.1
L_2	50.2	72.4

Table 1: Zero-shot classification in terms of Top-1 (%) accuracy on HDMB-51 and UCF-101.

Autoregressive vs MLM for GRAMMAR-R: In addition to the MLM used to compute the GRAMMAR-R reward, herein we also consider autoregressive architectures, mainly a pre-trained GPT2 [5] model. Results are reported in Table 2. For the task at hand (inferring how likely the generated text is correct), both models perform very similarly. Note that both models have similar sizes.

LM	HMDB-51	UCF-101
BERT-B	53.6	74.6
GPT2	53.2	74.5

Table 2: Zero-shot classification in terms of Top-1 (%) accuracy on HDMB-51 and UCF-101.

C. Qualitative examples

Additional qualitative examples can be seen in Fig. 3 where we compare our approach with REST.

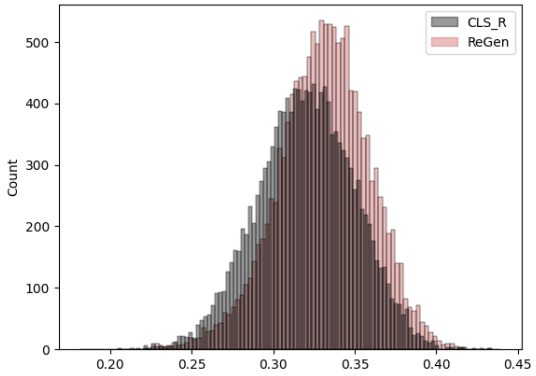


Figure 2: Cosine similarity between the videos and the generated caption in the CLIP embedding space for a model trained with CLS-R only vs a model trained with all rewards (i.e. ReGen).

Label: building sandcastle
REST: a little girl playing in the sand
ReGen (Ours): a little girl making a sand castle out of sand on the beach

Label: waving hand
REST: a man in a black shirt is shown in the middle of the screen
ReGen (Ours): two men waving and talking to the camera

Label: playing hand clapping games
REST: two girls playing a game of rock paper scissors
ReGen (Ours): two girls giving each other a high five to each other's hands

Label: putting on lipstick
REST: a woman's face with red lipstick
ReGen (Ours): a woman with red lips applying lipstick on her lips and looking at the camera

Figure 3: Additional examples of captions produced by our approach and REST for a set of videos from Kinetics-220 (*i.e.* zero-shot setting).

References

[1] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 1

[2] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *IEEE International Conference on Computer Vision*, 2021. 1

[3] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The Kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[4] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: a large video database for human motion recognition. In *IEEE International Conference on Computer Vision*, 2011. 1

[5] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019. 2

[6] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[7] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research. In *IEEE International Conference on Computer Vision*, 2019. 1