Zero-Shot Action Recognition on UCF101 Using Vision-Language Models Project Report

Airlangga Parikesit Wibowo

I. Model Used

VideoCLIP is a Vision-Language model used for zero-shot understanding of the relationship between video and text, so it can transfer to various video-text understanding tasks without manual labeling. VideoCLIP uses a contrastive learning approach by training transformers for video and text. The model distinguishes temporally overlapping positive video-text pairs from negative pairs retrieved from nearest neighbor retrieval [1].

II. Experimental Setup

Environment: Kaggle Notebook
Dataset: UCF101 - Action Recognition
(https://www.kaggle.com/datasets/matthewjansen/ucf101-action-recognition)

- Install necessary Libraries

```
!pip install pandas numpy opencv-python pillow matplotlib ftfy torch
```

- Install VideoCLIP model files from Huggingface

```
!git lfs install

!git clone https://huggingface.co/alibaba-pai/VideoCLIP-XL
!git clone https://huggingface.co/alibaba-pai/VideoCLIP-XL-v2
```

- Move model files in VideoCLIP-XL file to root directory

```
!mv VideoCLIP-XL/* /kaggle/working
```

III. Accuracy and example outputs

VideoCLIP model achieved good results when used for zero-shot learning on 10 classes with each having 1 video and/or frame from UCF101 dataset. The model correctly classified all video or frame samples with high confidence scores, which demonstrate a strong generalization capability even without fine-tuning or few-shot learning. This suggests that for these 10 examples, a single frame was enough for the model to do a correct classification.

Based on the comparison between the result of video clips input and frame input, video clips give slightly higher confidence scores, but not much. Only ‘CricketShot’ and ‘TennisSwing’ classes show a noticeable improvement,

sequentially by 2,07% and 1,41%. On the other hand, ‘PlayingDhol’, ‘ShavingBeard’, and ‘BoxingPunchingBag’ show nearly identical scores between video clip and frame, indicating that a single frame contains enough context.

Table 1. Zero-shot learning result of video clip

Ground Truth	Prediction	Probability Score (%)	Ground Truth	Prediction	Probability Score (%)
Basketball Video: v_Basketball_g12_c02.avi	Basketball	98,78	Drumming Video: v_Drumming_g07_c07.avi	Drumming	96,8
	ShavingBeard	0,07		PlayingDhol	2,01
	IceDancing	0,03		ShavingBeard	0,56
	CricketShot	0,03		PlayingCello	0,43
	TennisSwing	0,02		TennisSwing	0,07
ShavingBeard Video: v_ShavingBeard_g16_c02.avi	ShavingBeard	99.95	PlayingDhol Video: v_PlayingDhol_g05_c06.avi	PlayingDhol	63,39
	CricketShot	0.01		Drumming	35,33
	PlayingCello	0.01		PlayingCello	0,15
	Drumming	0.01		CricketShot	0.07
	IceDancing	0.01		BoxingPunchingBag	0.04
HorseRiding Video: v_HorseRiding_g09_c03.avi	HorseRiding	99.74	TennisSwing Video: v_TennisSwing_g25_c02.avi	TennisSwing	99.68
	IceDancing	0.18		Basketball	0.13
	TennisSwing	0.02		CricketShot	0.06
	BoxingPunchingBag	0.02		PlayingCello	0.04
	ShavingBeard	0.01		IceDancing	0.03
PlayingCello Video: v_PlayingC	PlayingCello	99.73	BoxingPunchingBag Video:	BoxingPunchingBag	99.96

ello_g09_c06.avi	TennisSwing	0.16	v_BoxingPunchingBag_g18_c06.avi	TennisSwing	0.01
	Drumming	0.06		PlayingDhol	0.01
	PlayingDhol	0.02		CricketShot	0.01
	Basketball	0.01		Drumming	0.01
CricketShot Video: v_CricketShot_g03_c07.avi	CricketShot	99.45	IceDancing Video: v_IceDancing_g02_c02.avi	IceDancing	99.93
	TennisSwing	1,31		TennisSwing	0.03
	Drumming	0.06		HorseRiding	0.01
	Basketball	0.05		PlayingCello	0.01
	IceDancing	0.03		ShavingBeard	0.0

Table 2. Zero-shot learning result of single frame

Ground Truth	Prediction	Probability Score (%)	Ground Truth	Prediction	Probability Score (%)
Basketball Video: v_Basketball_g12_c02.avi	Basketball	98,71	Drumming Video: v_Drumming_g07_c07.avi	Drumming	97,72
	CricketShot	0,07		PlayingDhol	1,71
	TennisSwing	0,07		PlayingCello	0,33
	BoxingPunchingBag	0,04		TennisSwing	0,06
	ShavingBeard	0,03		ShavingBeard	0,05
ShavingBeard Video: v_ShavingBeard_g16_c02.avi	ShavingBeard	99.9	PlayingDhol Video: v_PlayingDhol_g05_c06.avi	PlayingDhol	64,4
	Drumming	0.02		Drumming	35,41
	PlayingCello	0.02		PlayingCello	0,09
	Basketball	0.01		BoxingPunchingBag	0.05
	CricketShot	0.01		CricketShot	0.03
HorseRiding	HorseRiding	99.64	TennisSwin	TennisSwing	99.27

Video: v_HorseRiding_g09_c03.avi	IceDancing	0.11	g Video: v_TennisSwing_g25_c02.avi	Basketball	1,15
	TennisSwing	0.1		IceDancing	0.19
	PlayingDhol	0.04		CricketShot	0.14
	PlayingCello	0.04		Drumming	0.06
PlayingCello Video: v_PlayingCello_g09_c06.avi	PlayingCello	99.27	BoxingPunchingBag Video: v_BoxingPunchingBag_g18_c06.avi	BoxingPunchingBag	99.53
	TennisSwing	0.5		PlayingDhol	0.16
	Drumming	0.1		TennisSwing	0.12
	Basketball	0.03		Drumming	0.06
	PlayingDhol	0.03		CricketShot	0.05
CricketShot Video: v_CricketShot_g03_c07.avi	CricketShot	96,38	IceDancing Video: v_IceDancing_g02_c02.avi	IceDancing	99.68
	TennisSwing	3,27		TennisSwing	0.12
	Drumming	0.11		Basketball	0.04
	Basketball	0.07		Drumming	0.04
	IceDancing	0.04		PlayingDhol	0.03

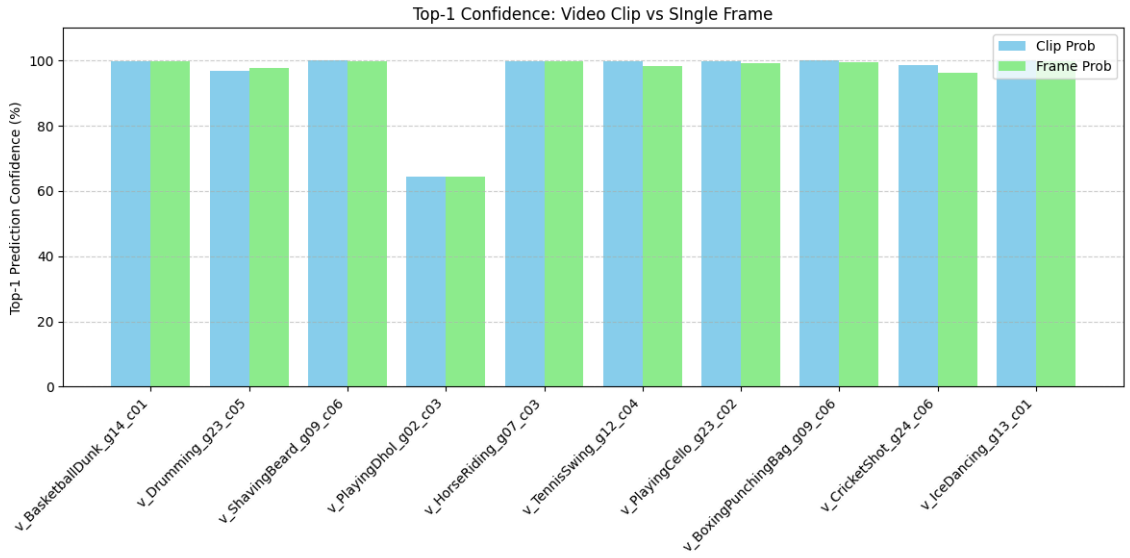

Image 1. Comparison of top-1 confidence score between video clip and single frame

The cosine similarity analysis between text and visual features reveals the performance of zero-shot action recognition using both video clips (multiple frames) and single-frame inputs. The cosine similarity heat map illustrates that each video-text and frame-text pair corresponding to the correct class has the highest cosine similarity value. This result indicates that VideoCLIP effectively aligns video visual and textual features in a zero-shot learning. However, there is one similarity other than the correct classes that are notably high, that is PlayingDhol and Drumming which have visually similar actions.

When compared, there's a little difference between the result of video-text and frame-text cosine similarity. On average, the cosine similarity for the correct class is slightly higher for video clips, but the difference is not significant. Single-frame inputs perform surprisingly well, achieving comparable similarity scores for many classes and offering a lightweight alternative suitable for applications with limited computational resources. This suggests the model performs well even with a single frame for most actions.

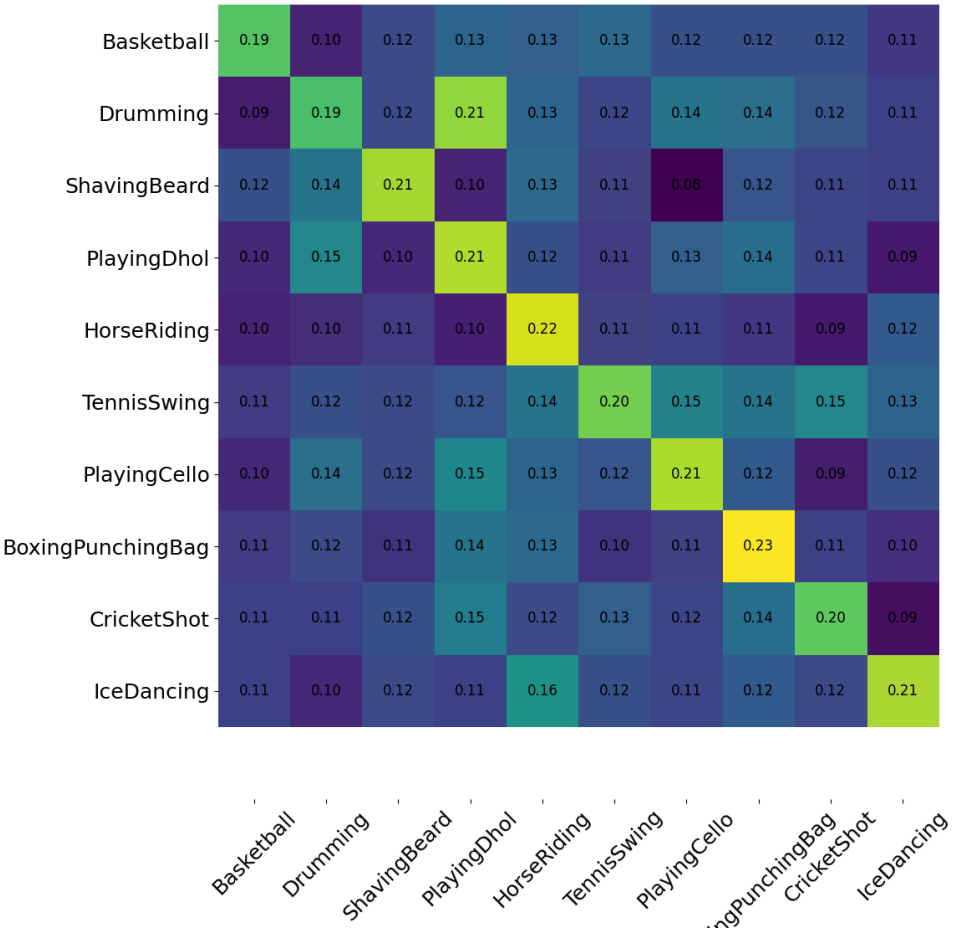

Image 2. Cosine similarity heatmap of video clip and text

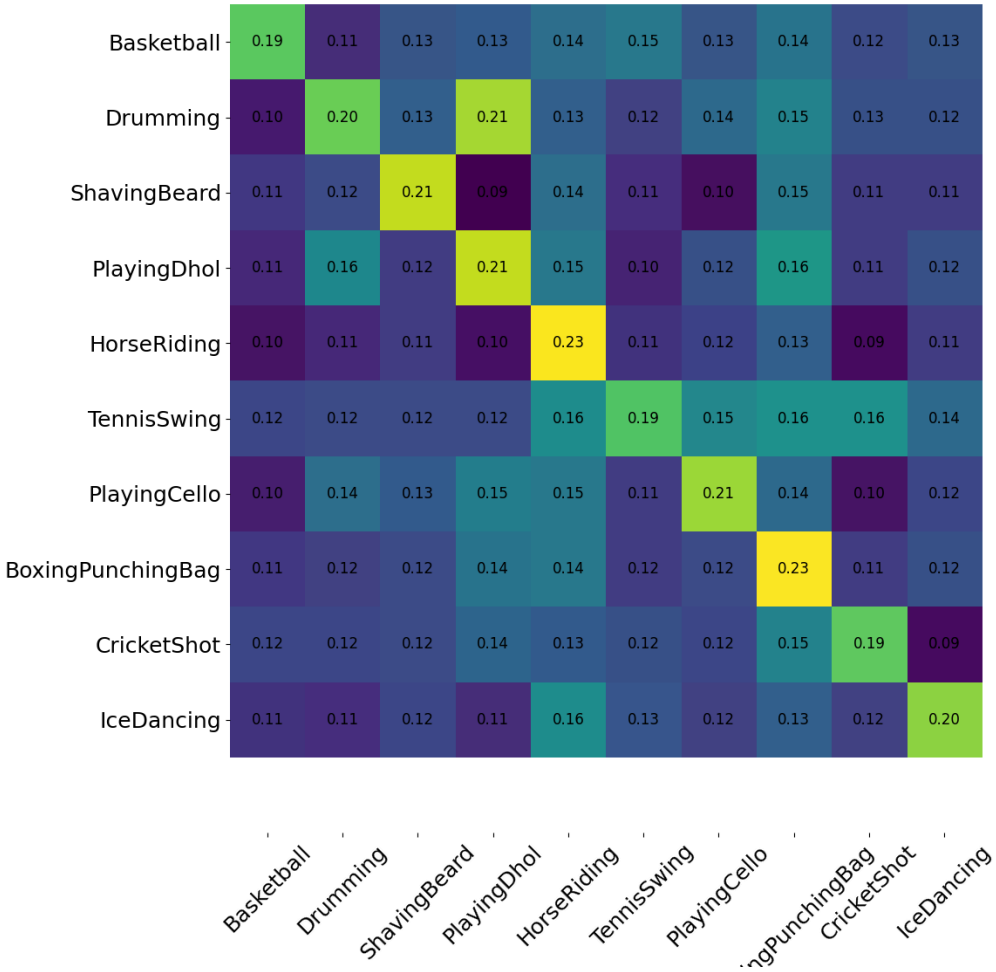

Image 3. Cosine similarity heatmap of frame and text

IV. Limitations

While the VideoCLIP model delivers impressive zero-shot performance on UCF101 dataset, achieving probabilities above 90%, it still struggles to distinguish actions that share nearly identical visual characteristics. In particular, the model's confidence at recognizing 'PlayingDhol' action are relatively low, with only a 64,39% confidence score using multiple frames and 64,4% using a single frame. This indicates a confusion with the 'Drumming' class, where PlayingDhol action was recognized as Drumming with 35,33% score using video clips and 35,41% using single frame. At glance, 'PlayingDhol' and 'Drumming' appear visually similar, both involving hand movements with similar instruments.

Reference

Xu H, Ghosh G, Huang P, Okhonko D, Aghajanyan A, Metze F, Zettlemoyer L, Feichtenhofer C. 2021. VideoCLIP: Contrastive Pre-training for Zero-shot

Video-Text Understanding. https://arxiv.org/abs/2109.14084.