FitCLIP: Refining Large-Scale Pretrained Image-Text Models for Zero-Shot Video Understanding Tasks

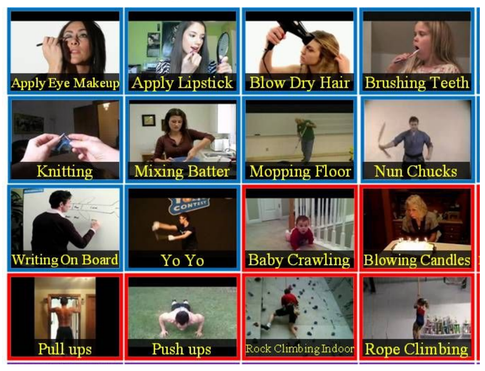
Santiago Castro and Fabian Caba Heilbron

sacastro@umich.edu

Zero-shot Video Understanding

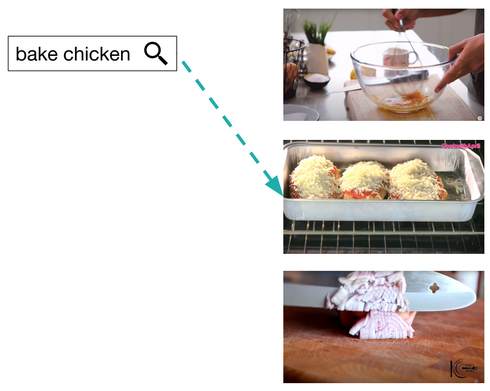
Zero-shot learning: evaluate a model on a dataset different from what it was trained on.

Video Classification (e.g., UCF101)



… and 85 more action classes

Text-to-Video Retrieval (e.g., YouCook2)



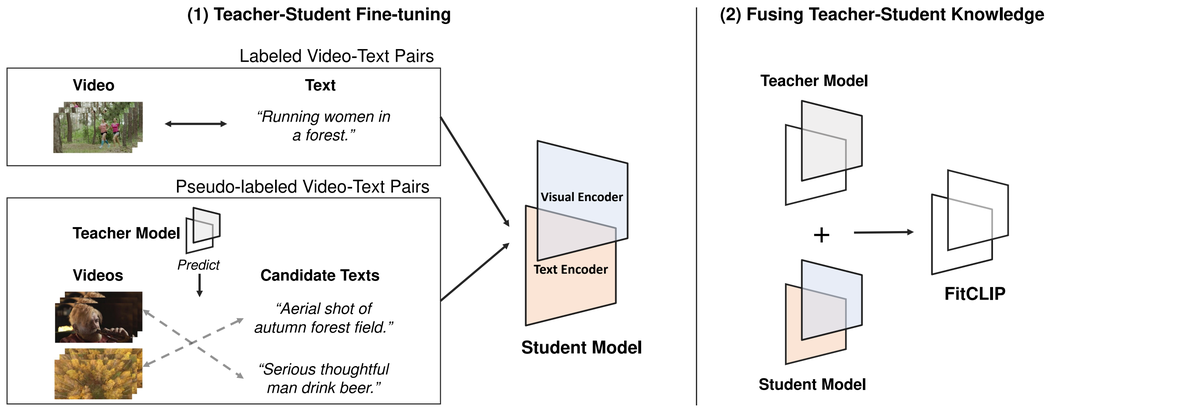
…

Motivation

Strong image-text models do great on videos. For example, 64% of the times, the correct video is recovered within the top 10 with CLIP on MSR-VTT given a text query, without ever seeing any video from it. This is because of its robustness.

But they aren't made for video. Can we adapt CLIP but at the same time keep the robustness?

Our Proposed Method: FitCLIP



$$\mathcal{L} = \lambda(\mathcal{L}_{distill,v2t}+\mathcal{L}_{distill,t2v})+(1-\lambda)(\mathcal{L}_{v2t}+\mathcal{L}_{t2v}) \qquad w_i = \alpha w_i^{(student)} + (1-\alpha) w_i^{(teacher)}$$

Dataset: WebVid-4.5k

Benchmarks

Video Classification

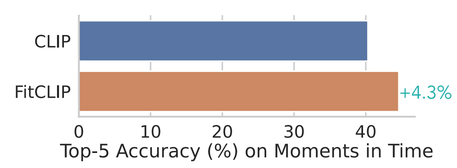
dataset	# classes	# samples
Moments in Time (MiT)	339	33,900
UCF101	101	1,794

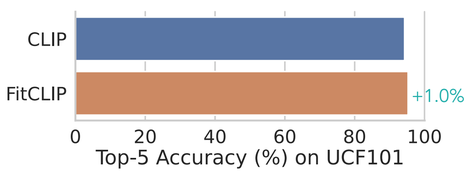
Text-to-Video Retrieval

dataset	# samples	genre
MSR-VTT	1,000	user-generated
YouCook2	3,305	cooking
DiDeMo	4,021	user-generated

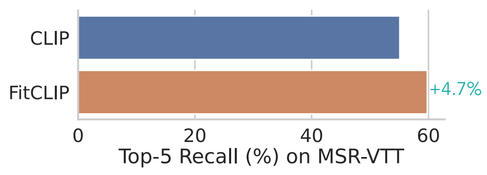
Main Results

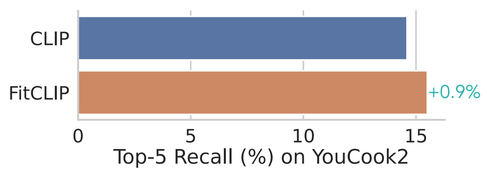
Video Classification

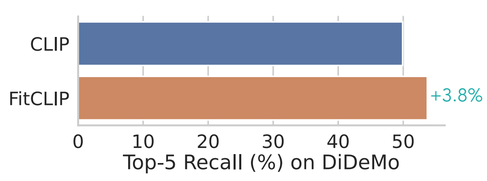




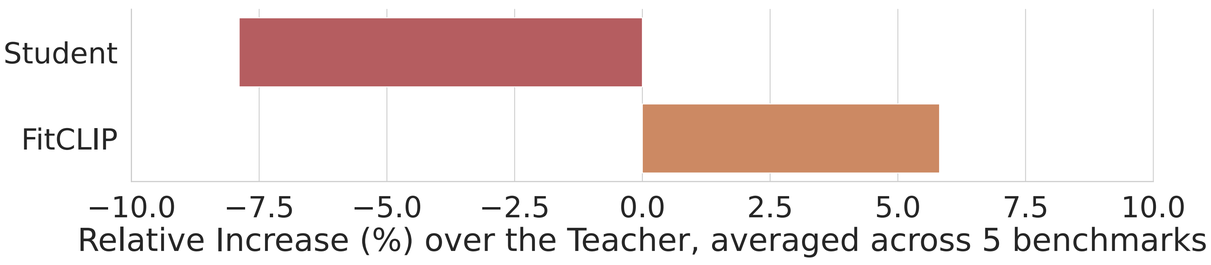
Text-to-Video Retrieval



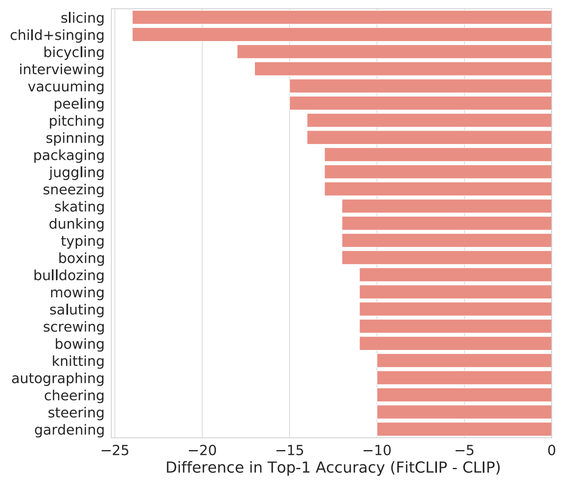


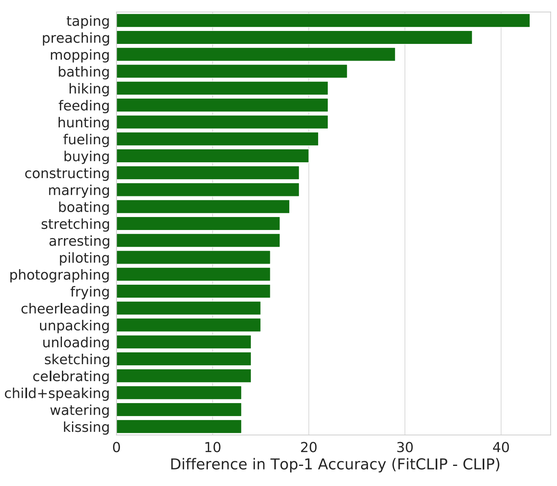


Importance of Fusing the Teacher and the Student

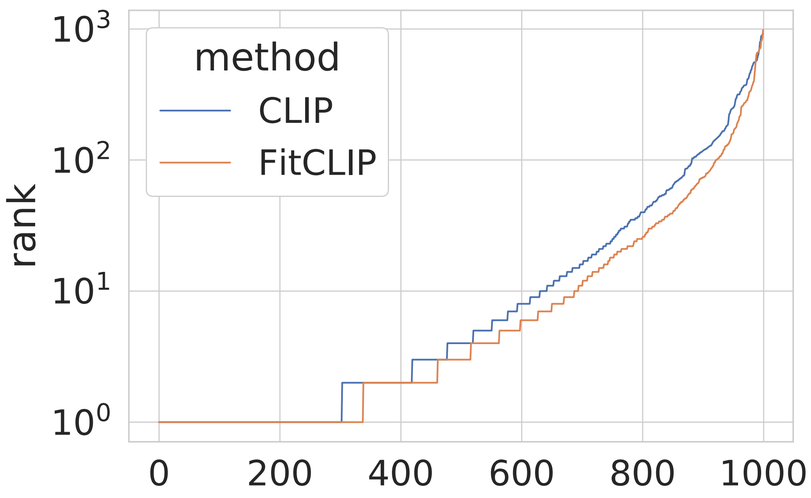


FitCLIP vs. CLIP per-class improvements (MiT)

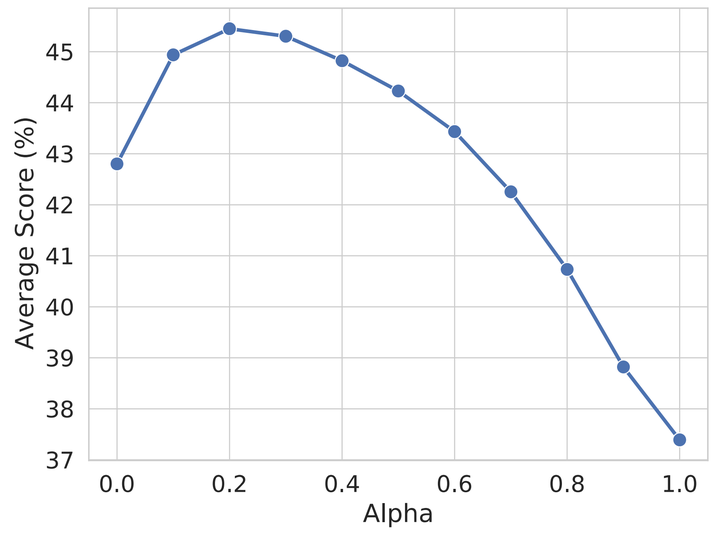




FitCLIP vs. CLIP distribution of Text-to-Video Retrieval rankings (MSR-VTT)



Varying the Fusion Factor (alpha)



Takeaways

- An effective strategy to adapt Vision-Language models to Video.
- We show how to prevent knowledge drifting by fusing teacher-student knowledge.

Data + Code:

github.com/bryant1410/fitclip

SCAN ME

BMVC 2022