CEZSAR: A Contrastive Embedding Method for Zero-Shot Action Recognition

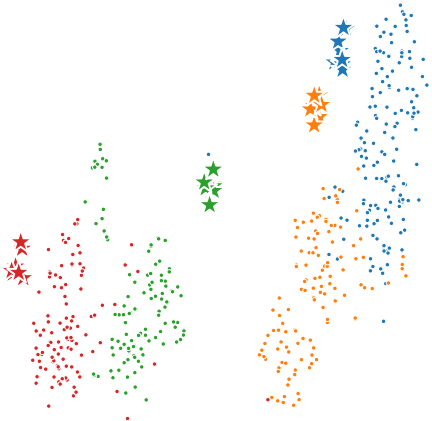
Valter Estevam[1,2], Rayson Laroca[3,2], Hélio Pedrini[4], and David Menotti[2]

[1]Federal Institute of Paraná, Irati, Brazil
[2]Federal University of Paraná, Curitiba, Brazil
[3]Pontifical Catholic University of Paraná, Curitiba, Brazil
[4]University of Campinas, Campinas, Brazil

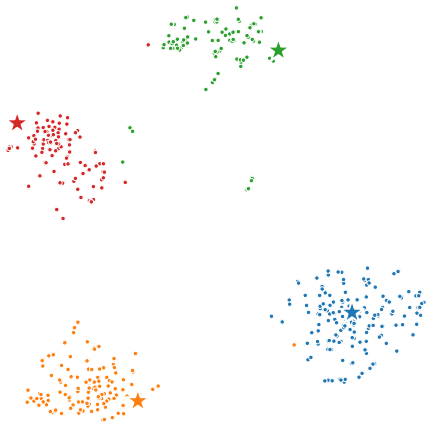
Problem Statement

- ZSAR: Recognition of human actions never seen during training.
- Knowledge transfer through semantic representations (videos and labels).
- Main challenges: semantic gap and domain shift.
  - Semantic gap: rich descriptions for videos and prototypes instead of only class labels.
  - Domain shift: training guided by textual descriptions.

Semantic Gap



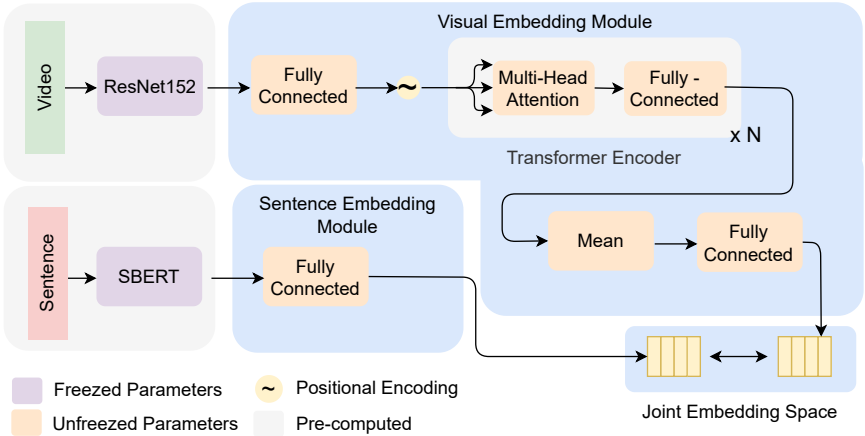
ZSARCap [5].



Ours.

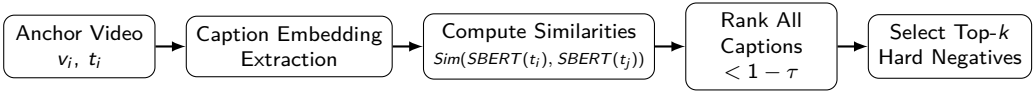
Horse Riding (blue); Horse Race (orange); Pommel Horse (green); Balance Beam (red);

Proposed Method



Contrastive multimodal embedding model.

Proposed Method - Hard Negative Mining



v_i (positive video)
t_i (positive sentence)

t_j (negative candidate sentence)
t_{j+1}
t_n
SBERT = Sentence-BERT encoding

$\tau = 0.8$

Negative samples are selected at the video clip description level rather than based on differences in the visual space, which we hypothesize helps mitigate the domain shift problem.

Experimental Setup

- ActivityNet Captions:
  - 20,000 untrimmed videos (12,000*).
  - One sentence per segment.
  - Average of 1 minute per segment.
  - No class labeling.
- Semantic class descriptions generated with Google Gemini 3 (Pro).

- UCF-101 benchmark.
  - 101 classes (13,320 videos).
  - $\approx$ 7 seconds per video.
  - ~~50%~~/50% - 50 runs; ~~80%~~/20% - 50 runs, and ~~0~~/100% - 1 run.
- Kinetics-400 benchmark.
  - 400 classes (at least 400 videos per class).
  - we obtained only 242,658 ($\approx$ 80%)
  - 10 seconds per video.
  - 25 classes - 50 runs, 100 classes - 50 runs, and 400 classes - 1 run

Results: UCF-101

Accuracy (%) under different numbers of test classes. VE = visual features; O = objects; C = captions.

Model	Classes	UCF-101 – Test classes		
		101	50	20
Mettes and Snoek [13]	–	32.8	40.4 ± 1.0	51.2 ± 5.0
Mettes *et al.* [14]	–	36.3	47.3	61.1
Kim *et al.* [11]	51	–	48.9 ± 5.8	–
Chen and Huang [3]	51	–	51.8 ± 2.9	–
Brattoli *et al.* [1]	664	39.8	48	–
Huang *et al.* [8]	51	–	46.4 ± 3.1	–
Kerrigan *et al.* [10]	664	40.1	49.2	–
Doshi *et al.* [4]	595	45.0	–	–
Huang *et al.* [9]	51	–	45.9 ± 3.4	–
Estevam *et al.* [5]	–	–	49.0 ± 3.5	–
Lin *et al.* [12]	664	–	58.7 ± 3.3	–
Lin *et al.* [12]	605	46.7	55.9	–
Gowda *et al.* [7]	51	–	53.9 ± 2.5	–
Estevam *et al.* [6] (O)	–	39.8	49.4 ± 4.0	60.0 ± 8.5
Estevam *et al.* [6] (O + C)	–	40.9	53.1 ± 3.9	63.7 ± 8.3
Ours (VE)	–	49.8	59.1 ± 2.7	72.6 ± 5.9
Ours (VE + O)	–	**50.9**	**59.4 ± 3.6**	**72.9 ± 5.9**

Discussion: UCF-101

- Consistent improvements across all evaluation protocols.
- Significant gains over recent methods:
  - $+4.2\%$ over Lin et al. (CVPR'22 – 605 classes in traning phase)
- Object cues improve discrimination.

Our method consistently reduces the semantic gap on the UCF-101 benchmark.

Results: Kinetics-400

Accuracy (%) under different numbers of test classes. VE = visual features; O = objects; C = captions.

Model	Kinetics-400 – Test classes		
	400	100	25
Mettes and Snoek [13]	6.0	10.8 ± 1.0	21.8 ± 3.5
Mettes *et al.* [14]	6.4	11.1 ± 0.8	21.9 ± 3.8
Bretti and Mettes [2]	9.8	18.0 ± 1.1	29.7 ± 5.0
Estevam *et al.* [6] (O)	20.4	32.4 ± 2.4	49.3 ± 6.8
Estevam *et al.* [6] (O + C)	19.4	35.1 ± 2.4	54.6 ± 6.1
Ours (VE)	21.5	38.4 ± 2.0	58.1 ± 4.6
Ours (VE + O)	**23.8**	**40.4 ± 1.7**	**58.7 ± 5.9**

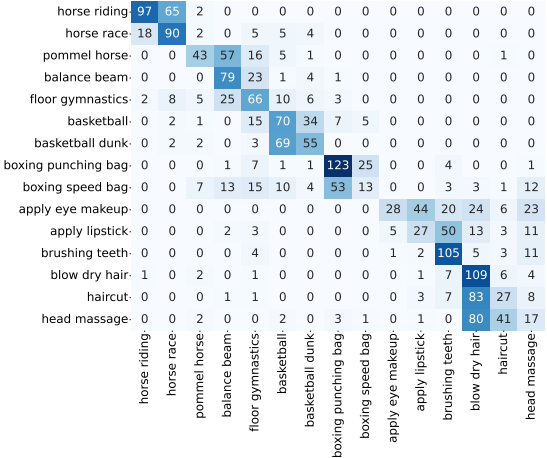
Discussion: Kinetics-400

- Kinetics-400 presents substantially larger semantic diversity.
- CEZSAR consistently improves over previous methods.
- Rich semantic prototypes generated by Gemini 3.
- Object representations provide complementary information.

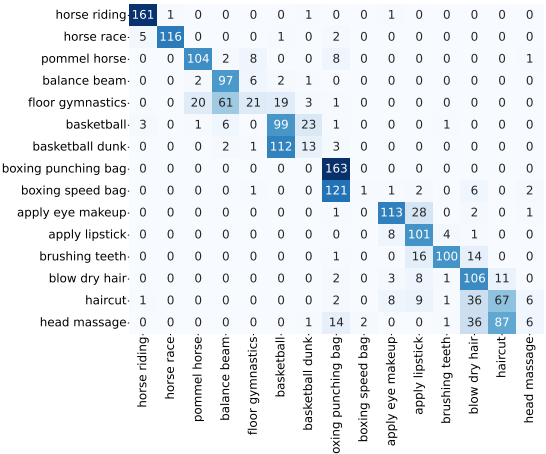
Our method remains effective even in large-scale ZSAR scenarios with hundreds of unseen actions.

Evaluation on Hard Action Classes - UCF-101

1. **Horses**
   - Horse Riding
   - Horse Race
2. **Gymnastics**
   - Pommel Horse
   - Balance Beam
   - Floor Gymnastics
3. **Basketball**
   - Basketball
   - Basketball Dunk
4. **Boxing**
   - Punching Bag
   - Speed Bag
5. **Face-related**
   - Apply Eye Makeup
   - Apply Lipstick
   - Brushing Teeth
6. **Hair-related**
   - Blow Dry Hair
   - Haircut
   - Head Massage



ZSARCap [5].



Ours.

Conclusions

- Contrastive learning effectively bridges the semantic gap
  - Significant improvements across two ZSAR benchmarks.
  - Easily incorporates semantic information without retraining (ex.: object descriptions).
- Textual information mitigates domain shift
  - Text prototypes consistently improve visual representations.
  - Samples are naturally drawn closer to their corresponding classes.
- Automatic hard negative sampling is practical and scalable
  - Improves representation learning.
  - Training in a few hours on a single commercial GPU.
- We use only 0.5% of the number of triplets employed by contrastive models such as CLIP [15].

References

[1] B. Brattoli, J. Tighe, F. Zhdanov, P. Perona, and K. Chalupka. Rethinking zero-shot video classification: End-to-end training for realistic applications. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4613–4623, June 2020.

[2] C. Bretti and P. Mettes. Zero-shot action recognition from diverse object-scene compositions. In *British Machine Vision Conference (BMVC)*, pages 1–14, Nov. 2021.

[3] S. Chen and D. Huang. Elaborative rehearsal for zero-shot action recognition. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13638–13647, Oct. 2021.

[4] K. Doshi et al. A Multimodal Benchmark and Improved Architecture for Zero Shot Learning . In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2010–2019, 2024.

[5] V. Estevam et al. Tell me what you see: A zero-shot action recognition method based on natural language descriptions. *Multimedia Tools and Applications*, 83:28147–28173, 2024.

[6] V. Estevam, R. Laroca, H. Pedrini, and D. Menotti. Global semantic descriptors for zero-shot action recognition. *IEEE Signal Processing Letters*, 29:1843–1847, 2022.

[7] S. N. Gowda, L. Sevilla-Lara, F. Keller, and M. Rohrbach. CLASTER: Clustering with reinforcement learning for zero-shot action recognition. In *European Conf. on Computer Vision (ECCV)*, pages 187–203, 2022.

[8] K. Huang, L. Miralles-Pechuán., and S. Mckeever. Combining text and image knowledge with GANs for zero-shot action recognition in videos. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 623–631, 2022.

References

[9] K. Huang, L. Miralles-Pechuán, and S. Mckeever. Enhancing zero-shot action recognition in videos by combining gans with text and images. *SN Computer Science*, 4(4):375, 2023.

[10] A. Kerrigan, K. Duarte, Y. Rawat, and M. Shah. Reformulating zero-shot action recognition for multi-label actions. In *International Conference on Neural Information Processing Systems (NeurIPS)*, volume 34, pages 25566–25577, 2021.

[11] T. S. Kim et al. DASZL: Dynamic action signatures for zero-shot learning. In *AAAI Conference on Artificial Intelligence*, 2021.

[12] C.-C. Lin et al. Cross-modal representation learning for zero-shot action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19978–19988, June 2022.

[13] P. Mettes and C. G. M. Snoek. Spatial-aware object embeddings for zero-shot localization and classification of actions. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4453–4462, 2017.

[14] P. Mettes, W. Thong, and C. Snoek. Object priors for classifying and localizing unseen actions. *International Journal of Computer Vision*, 129:1954–1971, 2021.

[15] A. Radford et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763, 2021.

Thank you!

https://github.com/valterlej/cezsar