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

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

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

Abstract. This paper proposes a novel Zero-Shot Action Recognition (ZSAR) method based on contrastive learning. In ZSAR, we aim to classify examples from classes that were missing during training. Two well-known problems remain in ZSAR: the semantic gap and the domain shift. A semantic gap occurs because label representations come from the textual domain (i.e., language models) and must be associated with visual representations (i.e., CNNs, RNNs, transformer-based). This multimodal nature implies that the semantic properties of the two spaces are not identical. On the other hand, the domain shift arises from differences between the training and test sets and is inherent to ZSAR once the test set is unknown. One of the most promising methods to address both issues is learning joint embedding spaces. Therefore, we propose a new model that encodes videos and sentences in a joint embedding space, trained by aligning videos with their natural-language descriptions. We design an automatic negative sampling procedure to augment the training dataset and generate unpaired data, i.e., visual appearance and unrelated descriptions. Our results are state-of-the-art on the UCF-101 and Kinetics-400 datasets under several split configurations. Our code is available at https://github.com/valterlej/cezsar.

Keywords: Semantic gap · Language-video representation · Contrastive learning · Zero-shot learning.

1 Introduction

Zero-shot learning is a well-established problem in computer vision that aims to classify instances belonging to classes that were not available for training the models, usually called unknown or unseen classes [26]. Nowadays, there are zero-shot approaches for objects [6,23,32], human actions [9,13,14,17], and many other domains [29]. This work focuses on Zero-Shot Action Recognition (ZSAR) in videos, i.e., in classifying instances (short video clips up to 10 s duration) of unknown action classes. This particular problem has attracted the attention of the computer vision community in the last decade [11].

Accepted at ICPR 2026. The final published version will be available on *Springer*.

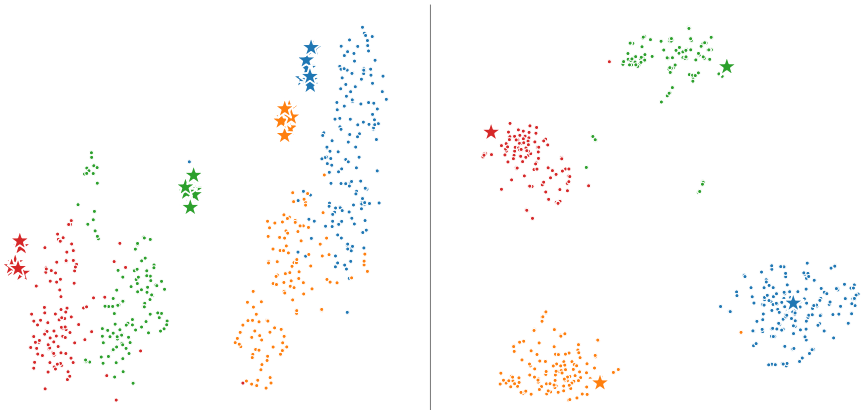
The most popular human action recognition approaches employ supervised learning, requiring a massive set of annotated videos for training, and keeping these models up to date is extremely challenging because new actions emerge every day as new objects, techniques, and forms of human interaction appear. Moreover, new actions are rare and unavailable on YouTube or other large-scale sources. Even when available, the inclusion of new classes requires retraining existing models, demanding extensive computational resources, energy, and human labor to annotate instances with appropriate labels [12].

In ZSAR, on the other hand, the need for annotations is transferred from the instances to the action classes. It takes a lot less work to annotate classes (a few hundred annotations) than it does to annotate tens or hundreds of thousands of instances. Hence, several pioneer works considered a set of attributes defined by humans as semantic information [25]. This representation is called prototype and ideally represents the archetype for each class. Nevertheless, even such an approach requires a lot of human effort and is not scalable, being replaced by an automatic procedure called label embedding, which uses word embedding methods [34] or sentence embedding methods [5, 12] to define the prototypes. Recently, the most promising ZSAR methods relate visual appearance (e.g., given by some neural network) with semantic class information associated with their label projecting them into a joint embedding space. Due to the multi-modal nature, two crucial problems remains in these approaches: the domain shift and the semantic gap between the modalities.

The semantic gap is the information difference for each modality used by the methods, i.e., the distribution of instances in visual space is often distinct from that of their underlying semantics in semantic space [34]. For example, in Fig. 1a, we demonstrate that this problem occurs even in joint embedding-based models such as ZSARCAP [12] or our proposed method. The dots in the figure represent the video embeddings, and the stars the label embeddings. The lack of information and the challenges in relating them are the origins of this problem. For instance, *Pommel Horse* (green) and *Balance Beam* (red) are usually performed in gymnasiums. Therefore, they present similar frames in which the scene structure is similar, only differing in the artistic gymnastic equipment and some specific motions.

A strategy to mitigate the semantic gap on the visual side is to provide temporal information to learn a motion signature [34]. Another strategy is to explore the relationships among actions and objects [9, 27]. These relationships occur in videos and texts. Thus, it is possible to recognize objects in scenes and infer the action because the same information is used. This last approach is robust in visual-semantic representation but fails in temporal modeling, which is essential to recognize actions independent of scenarios or objects (e.g., *run, turn, punch,* and *head massage*).

The semantic gap is also present in label encoding. Methods extensively used, such as Word2Vec or GloVe, fail to capture fine-grained differences because they project similar concepts (e.g., *Horse Riding* and *Horse Race*) close and, in some cases, also dissimilar ones (e.g., *Pommel Horse and Horse Riding*). Moreover,



(a) ZSARCAP [12] (Acc. 67.2%) (b) Ours (Acc. 95.2%)

Fig. 1. T-SNE visualization for a subset with the classes Horse Riding (blue), Horse Race (orange), Pommel Horse (green), and Balance Beam (red). Dots are videos, and stars are label prototypes.

the label encoding process usually produces one array[5] for which we assume all required semantic information is encoded. Strategies to alleviate the semantic gap in label encoding include adding more descriptive texts and using better encoders, such as LLM-based encoders. [5,9,12]. As shown in Fig. 1b, our method generates a better separation among the classes (for both videos and prototypes) and a lower distance between prototypes and their corresponding videos.

Even though we have good descriptors for videos and texts, the domain shift problem remains unsolved. It corresponds to the differences in the probability distribution for the patterns in the training set compared to the test set [34]. We believe that textual semantics is much less affected by domain shift than visual. Hence, learning a joint embedding space for these modalities, conditioned by textual descriptions, should alleviate the domain shift problem for visual patterns and reduce the semantic gap between information modalities. Taking this into account, we propose a new method for ZSAR, called Contrastive Embedding Method for Zero-Shot Action Recognition (CEZSAR). It consists of a joint projection method trained with an additional dataset containing untrimmed videos paired with human-generated sentences describing what is occurring in the videos.

As illustrated in Fig. 2, our proposed model is a neural network with two modules. The first, called Visual Embedding Module (VEM), is responsible for encoding visual information given by a pre-trained Convolutional Neural Network (CNN) (e.g., ResNet-152). In this module, the videos are sampled at 1 frame per second (FPS) and passed through the CNN, resulting in a feature stack. There is also a fully connected layer responsible for reducing the stack dimen-

[5] This array represents the class prototype.

sionality and feeding a Transformer Encoder. This encoder uses self-attention to model temporal information for the videos. Therefore, we have two dense representations for which we expect to be close if the text describes the video and distant otherwise.

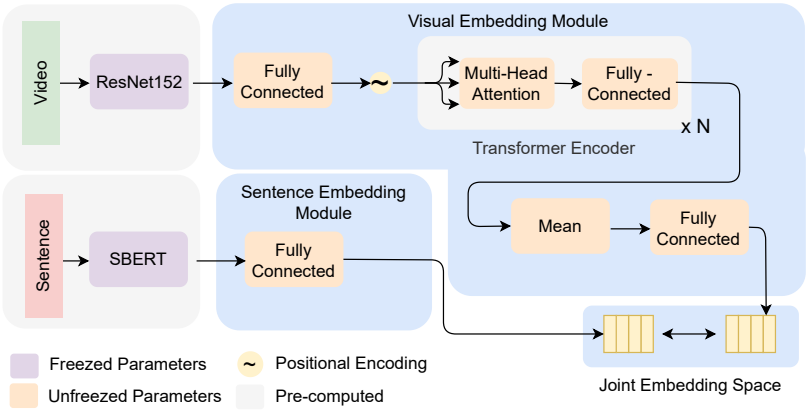


Fig. 2. Our method is composed of the Visual Embedding and Sentence Embedding modules. Each module produces a dense representation that is expected to be close if the sentence describes the video and far otherwise.

For training the model, we propose a hard negative sampling method. This method seeks negative alignments between videos and texts without human supervision. Thus, we can generate triplets (video, positive description, and negative description) and employ a triplet loss function. Our training process does not require a closed set of classes, but only enough pairs of videos and descriptions in natural language, and it can be completed in just a few hours on a standard Graphics Processing Unit (GPU).

In summary, the main contributions of this work are: (i) we introduce a new cross-modal contrastive learning method that effectively associates visual features and sentence descriptions with a reduced semantic gap; (ii) our model enables projecting videos and descriptions with two distinct sub-networks. Hence, we can include additional information such as texts, images, or even videos; and (iii) the robustness of our joint semantic space is demonstrated by reaching state-of-the-art results on the UCF-101 and Kinetics-400 datasets.

2 Related Work

This section briefly discusses joint embedding learning employing sentences and contrastive learning.

2.1 Joint Embedding Learning for ZSAR using Sentences

Estevam et al. [12] proposed a method to represent both sides, videos and labels, with descriptive sentences. They trained video captioning models [10] that produce a single sentence per video. The video captioning method models temporal information in videos to infer probabilities over a vocabulary and generate a sentence. Although the results obtained with this technique were promising, there is much room for improvement in video captioning to effectively establish stronger associations between visual and textual patterns, which we believe would improve the performance of ZSAR. Subsequently, the same research group [9] proposed to enrich the captioning sentences with textual descriptions given by objects recognized in the scenes, providing a robust set of semantic information that can be incorporated into our model.

2.2 Contrastive Learning for Zero-Shot Learning

Contrastive learning is a self-supervised learning technique that aims to learn a dense representation given label-visual pairs. In learned space, similar pairs stay close together, and dissimilar pairs stay far apart. Chopra et al. [7] were among the pioneers to propose a loss function for this problem. Recently, Han et al. [16] employed contrastive learning for generalized zero-shot learning, i.e., a sub-variant of the zero-shot learning problem that assumes the presence of seen and unseen classes in the test set. Although promising, their method was proposed for and evaluated on datasets that use manually annotated attributes to represent classes.

A benefit of contrastive learning is its robustness in preventing deep networks from overfitting noisy labels [37]. This property is critical for us because we deal with natural language descriptions that are intrinsically noisy due to ambiguities and annotators' perceptions of what should be described. In addition, language-image pre-trained models such as CLIP [29] have attracted increasing attention from the research community [22, 35, 36]. These models have shown impressive results in zero-shot experiments, but they rely on extensive training infrastructure (e.g., clusters with up to 596 Tesla V100 GPUs used for 18 [29]). Moreover, the dataset containing 400 million image-text pairs is not available for download, leading us to the following question: what would the results be for ER [5] or ZSARCAP [12] trained with comparable infrastructure and a similar amount of data? Our proposed method, for example, has 1000× fewer visual representations and 100× fewer data pairs, is trained with 5× less time on a single GPU, and achieves superior performance compared to non-clip-based methods.

3 Classification Model

3.1 Problem Definition

ZSAR can be stated as classifying a set of unseen action categories $Z_u = \{z_1, ..., z_{u_n}\}$ (i.e., never seen before by the model). It can be achieved by using

a set of seen categories $Z_s = \{z_1, ..., z_{s_n}\}$ so that $Z_u \cap Z_s = \emptyset$, or by transferring knowledge from other models trained without class labels, as in the proposed method. As mentioned earlier, our model consists of a neural network compounded by two modules fed with pre-computed features for both modalities, visual and semantic description. These modules are described in Section 3.2. As explained in Section 3.3, the model is trained in a contrastive way leveraging the proposed Hard Negative Sampling method, which is covered in Section 3.4. Finally, we present the ZSAR procedure in Section 3.5.

3.2 Joint Embedder Model

Initially, we explain the Visual Embedding Module (VEM). Given a video clip v with t seconds duration, we encode the frames at a rate of 1 FPS using a pre-trained CNN. Then, we got $v_c \in \mathbb{R}^{t \times d_c}$, where v_c is the feature stack for the video and d_c is the CNN output dimension (e.g., using the ResNet-152 model $d_c = 4.096$). This stack is fed to a fully connected layer aiming to reduce the dimensionality

$$v_r = \text{ReLU}(v_c W + b)\,, \tag{1}$$

where ReLU is a usual Rectified Linear Unit, W is an internal weight matrix, b is a bias vector, and v_r is the video stack projection into a lower dimensional space. This stack is fed to a Transformer encoder, and the position of each feature is encoded with sine and cosine at different frequencies, as proposed by Vaswani et al. [33]. Then, these representations are passed through a multi-head attention layer that employs the scaled dot-product, defined in terms of queries (Q), keys (K), and values (V) as

$$Att(Q, K, V) = softmax(\frac{Q.K^T}{\sqrt{d_k}})V\,. \tag{2}$$

The multi-head attention layer is a concatenation of several heads (1 to h) of self-attention ($Q = K = V = v_r^{PE}$) applied to the input projections as

$$\text{MHAtt}(v_r^{PE}, v_r^{PE}, v_r^{PE}) = [head_1, ..., head_h]W^0\,, \tag{3}$$

where $head_i = Att(v_r^{PE}W_i^{v_r^{PE}}, v_r^{PE}W_i^{v_r^{PE}}, v_r^{PE}W_i^{v_r^{PE}})$, v_r^{PE} is the v_r positional encoded, and [ ] is a concatenation operator.

Afterward, a fully connected feed-forward network FFN($\cdot$) is applied to each position separately and identically

$$\text{FFN}(u) = \max(0, uW_1 + b_1)W_2 + b_2\,, \tag{4}$$

resulting in v_r^{FFN}. These features are averaged and fed to a fully connected layer responsible for projecting the result onto the joint semantic space, with d_{emb} dimensions[6], as

$$v_{emb} = \text{ReLU}(\overline{v_r^{\text{FFN}}}W + b)\,. \tag{5}$$

[6] In our experiments, $d_{emb} = 128$.

The Sentence Embedding Module (SEM) takes a sentence s and computes their Sentence-BERT (SBERT) [30] representation SBERT($\cdot$), resulting in an array of 768 dimensions. This representation is fed to a fully connected layer to project onto the joint semantic space with d_{emb} dimensions

$$s_{emb} = \text{ReLU}(\text{SBERT}(s)W + b). \tag{6}$$

3.3 Contrastive Learning and Loss Function

We train our model using contrastive learning. Our goal is to learn representations for which the video and its positive description are close to each other, and the video and its negative description are far apart. Therefore, we employ the triplet loss [1] defined as

$$\max(\parallel v_{emb} - s_{emb_p} \parallel - \parallel v_{emb} - s_{emb_n} \parallel + \epsilon, 0)\,, \tag{7}$$

where v_{emb} is the output of our VEM, s_{emb_p} and s_{emb_n} are positive and negative sentence embeddings produced by our Sentence Embedding Module (SEM), $||\cdot||$ is a distance metric, and ϵ is a margin ensuring that s_{emb_p} is at least ϵ closer to v_{emb} than s_{emb_n}.

The positive description is annotated by humans, using natural language sentences. A complete description on how these annotations were made is available in [21]. As the dataset does not provide human-annotated negative samples, we design an automatic hard negative sampling procedure, described in the next section.

3.4 Hard Negative Sampling

Negative sampling is a straightforward procedure when the samples are class annotated. We need to select samples from any other class randomly. Similar samples can come from different classes, but human judgment is the ground truth. In our case, we have pairs of videos and descriptions, and using human judgment to evaluate the similarity degree of descriptions is infeasible. Therefore, we employ a neural network — pre-trained in the paraphrasing task (i.e., the SBERT model) — to evaluate if two different sentences have the same semantics. We consider similar sentences if $Sim(\text{SBERT}(x_1), \text{SBERT}(x_2)) > 1 - \tau$[7], where Sim is the cosine similarity. We can find n negative descriptions for each pair using this rule[8].

To improve our search for negative samples and augment the dataset, we evaluate the similarity of detected objects. First, we filter two descriptive sentences for the detected objects most similar (using the rule previously defined) to the human-annotated sentence. We then select a negative candidate for each positive description that is sufficiently different from each of these three positive descriptions (i.e., one from human annotation and two from object descriptions).

[7] In our experiments, we set $\tau = 0.8$.
[8] We set $n = 10$.

Finally, for each temporal segment in the untrimmed videos, we randomly select three segments of up to 10 seconds. This augments the dataset by generating different positive pairs. Using these strategies, we obtained about three million triplets (video, positive description, and negative description).

3.5 ZSAR Classification

Our classification consists of mapping both videos, including all semantic information available (i.e., visual and object definitions) and class semantic information (i.e., prototypes given by sentence class descriptions[9]) into a joint embedding space. Then, the classification is performed with the nearest neighbor rule under some similarity function, such as

$$z_{u_{pred}} = \underset{z_{u_{prot}} \in \mathcal{Z}_{u_{prots}}}{\arg\max} \; \text{Sim}(\text{SE}(z_{u_{prot}}), \text{VidE}(v)) \,, \tag{8}$$

in which Sim is the cosine similarity; v is a video, $z_{u_{prot}}$ is a sentence for each class, $\text{SE}(\cdot)$ is a sentence embedding function defined in Eq. (6), and $\text{VidE}(\cdot)$ is the video embedding function defined as

$$\text{VidE}(v) = \alpha \text{VE}(v) + \beta \text{SE}(\text{O}(v)) \,, \tag{9}$$

where $\text{VE}(\cdot)$ is the visual embedding that uses the visual embedding module to encode the raw frames, $O(\cdot)$ is responsible for encoding objects recognized in scenes with their definitions from WordNet (as in [9]). The object classification can be performed directly from the ResNet-152 pre-computed features. Finally, α and β control the importance of each semantic feature.

4 Datasets, Protocol and Implementation Details

Our ZSAR experiments were carried out on the well-known UCF-101 [31] and Kinetics-400 [4] datasets. UCF-101 has 13,320 videos from 101 action classes, with an average duration of 7.2 seconds sampled at 25 FPS. Kinetics-400 is much larger, comprising 306,245 videos from 400 action classes with at least 400 clips each. All videos have a duration of 10 seconds and were collected from YouTube. It should be noted that we obtained only 242,658 clips (i.e., $\approx 80\%$) of the original dataset because many videos are unavailable[10].

The joint embedding model is learned with the ActivityNet Captions dataset [21]. It is a large-scale collection of YouTube videos, with temporal segments annotated and described by humans, each segment receiving one sentence of description. There are 20,000 untrimmed videos, divided into training, validation, and test sets, with 50/25/25% of the videos. We obtained $\approx$ 12,000 videos from the training and validation subsets in this work for the same reasons as in Kinetics-400.

[9] More details in Section 4.
[10] Removed or unavailable in our region.

The model was trained using an AdamW optimizer with lr $= 1e-4$, $\beta_1 = 0.9$, and $\beta_2 = 0.99$, with a weight decay of $1e-5$. A batch size of 128 was employed. Each video was encoded by obtaining a ResNet-152 feature per second (generating stacks with up to 15 features of 4096-d)[11]. Starting points for frame sampling were randomly selected. The visual branch of the model employs a Transformer encoder with d_model = 512, layers = 1, and heads = 2. The training was performed with 25 epochs and early stopping set to 10. The ZSAR classification is performed by projecting videos and labels onto a common space where the nearest neighbor classifier with $k = 1$ is used.

We evaluate our model on the UCF-101 dataset using the traditional protocol that randomly splits the dataset into seen and unseen classes (50%/50% - 50 runs; 80%/20% - 50 runs, and 0/100% - 1 run). We take only the test split because our joint embedding model is pre-trained on ActivityNet Captions, as described before. Therefore, in our case, the splits are 0%/50%; 0%/20%, and 0/100%. Considering the Kinetics-400 dataset, we evaluate the performance adopting the same number of random classes from [3,9,27,28] (i.e., 25 - 50 runs, 100 - 50 runs, and 400 - 1 run).

The labels were represented by descriptive sentences generated using Google Gemini 3 (Pro version). The model was asked to create descriptions defining the action, the objects involved, and the locations where it usually occurs[12]. All experiments were conducted on a computer with an AMD Ryzen 9 7950X 4.5 GHz CPU, 64 GB of RAM, and two NVIDIA RTX 5070 Ti GPUs (16 GB each).

5 Results and Discussion

Table 1 shows the results for the UCF-101 dataset. We highlight three sections in the table: the first, with a list of recent works; the second, with the performances of [9] using only objects (O) and using objects and captions (O+C). This model was chosen because it also constructs a joint embedding space, but employing SBERT exclusively; finally, we include our results using visual features (i.e., considering VidE(v) = VE(v) in Eq. (9)) and using our complete model (considering $\alpha = 0.8$, $\beta = 0.2$ in Eq. (9)). These values were defined based on the dominance of visual features.

Under the 0/101 configuration, we observe an expressive increment of 10.0 percentage points (p.p.) in accuracy compared to Estevam et al. [9] and 4.2 p.p. compared to Lin et al. [24]. Even when using only visual features, our model performs better than the one presented in [24]. It is worth noting that Lin et al. [24] use 50% of the UCF-101 classes, in addition to 605 classes from Kinetics-700, thereby significantly increasing the availability of training data.

[11] The features are available at https://1drv.ms/f/c/099341b05c7977d7/IgDTNPrqsdDQowglu0TNlvmAWx3reSAAOiXI2-PvW4Oeio?e=4F52DG

[12] The prompt used was: "Write a descriptive sentence for each human action in the following list. Include a definition of the action as well as information about the objects and places where it is usually performed. Return the results in JSON format. The list of human actions is: «dataset labels»".

Table 1. Results on the UCF-101 dataset reporting accuracy (%) under different numbers of test classes. No classes were used for training. The best results are highlighted. VE = visual features; O = objects; C = captions.

Model	Classes	UCF-101 – Test classes		
		101	50	20
Mettes and Snoek [28]	–	32.8	40.4 ± 1.0	51.2 ± 5.0
Mettes *et al.* [27]	–	36.3	47.3	61.1
Kim *et al.* [20]	51	–	48.9 ± 5.8	–
Chen and Huang [5]	51	–	51.8 ± 2.9	–
Brattoli *et al.* [2]	664	39.8	48	–
Huang *et al.* [17]	51	–	46.4 ± 3.1	–
Kerrigan *et al.* [19]	664	40.1	49.2	–
Doshi *et al.* [8]	595	45.0	–	–
Huang *et al.* [18]	51	–	45.9 ± 3.4	–
Estevam *et al.* [12]	–	–	49.0 ± 3.5	–
Lin *et al.* [24]	664	–	58.7 ± 3.3	–
Lin *et al.* [24]	605	46.7	55.9	–
Gowda *et al.* [15]	51	–	53.9 ± 2.5	–
Estevam *et al.* [9] (O)	–	39.8	49.4 ± 4.0	60.0 ± 8.5
Estevam *et al.* [9] (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**

Our results on UCF-101 have marginally improved with the inclusion of object-level semantic information.

Considering the experiments in the Kinetics-400 dataset shown in Table 2, we reached better results than the state of the art (SOTA) under all configurations. Semantic information was responsible for consistent improvements, strongly suggesting that the semantic gap has been reduced. Comparing VE against VE + O in the 0/25 configuration, we do not observe a real gain in mean accuracy from including objects, and the standard deviation has increased compared to results using only visual features. On the other hand, under the 0/400 configuration, the increase of 2.3 p.p. is significant due to the higher amount of unknown classes and high intra-class similarity in this dataset (e.g., eating: *burger, cake, carrots, chips, doughnuts, hotdog, ice cream, spaghetti,* and *eating watermelon*).

We investigated the impact of adjusting the parameter τ in the hard negative sampling procedure. This parameter regulates the criterion for determining whether a sentence is a positive or negative example. Essentially, τ is used in the unsupervised identification of pairs of negative sentences. Higher values of τ make it easier for the model to find negative examples. We found that it is challenging to mine negative examples when $\tau \leq 0.7$ and that the model fails to find a sufficient number of negative examples when $\tau \leq 0.6$. Consequently, we restricted our analysis to the threshold values of 0.7, 0.8, and 0.9 for distinguishing positive and negative examples. The models were evaluated using the UCF-101 dataset with all its classes and employing Eq. (9) to encode the videos. We observed that using a more relaxed criterion ($\tau = 0.7$) to determine negative examples did not lead to better ZSAR classification results (47.8% accuracy).

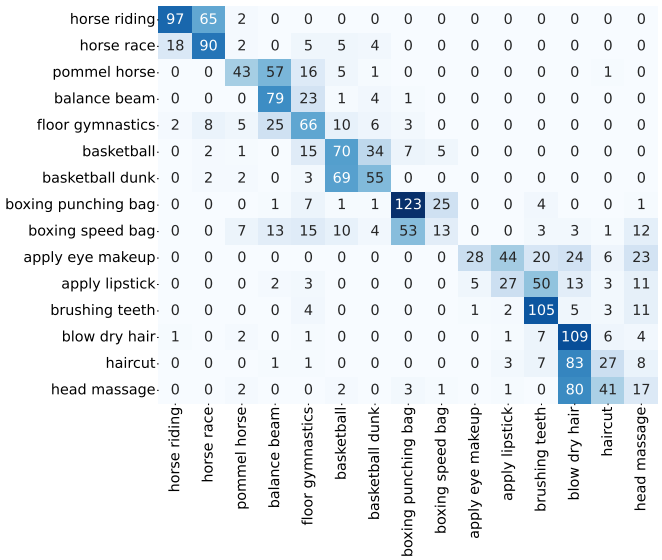
Table 2. Results on Kinetics-400 reporting accuracy (%) under different numbers of test classes. No classes were used for training. The best results are highlighted. VE = visual features; O = objects; C = captions.

Model	Kinetics-400 – Test classes		
	400	100	25
Mettes and Snoek [28]	6.0	10.8 ± 1.0	21.8 ± 3.5
Mettes *et al.* [27]	6.4	11.1 ± 0.8	21.9 ± 3.8
Bretti and Mettes [3]	9.8	18.0 ± 1.1	29.7 ± 5.0
Estevam *et al.* [9] (O)	20.4	32.4 ± 2.4	49.3 ± 6.8
Estevam *et al.* [9] (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**

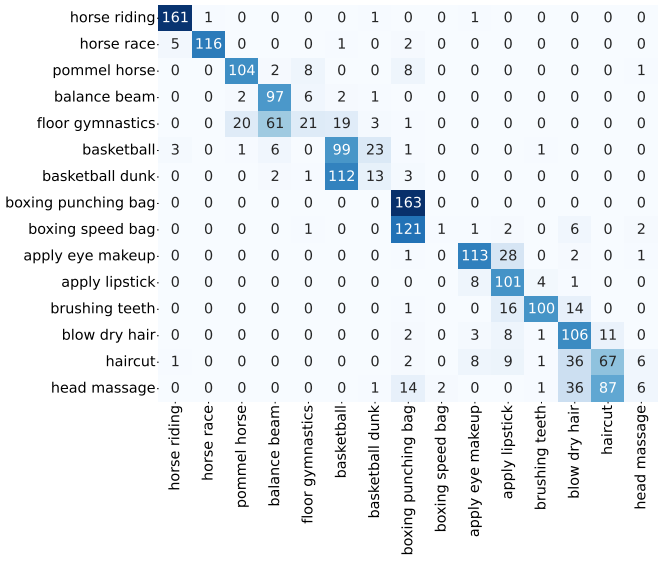
Moreover, restricting the definition of similar examples ($\tau = 0.9$) did not lead to higher ZSAR performance (48.3% accuracy). Therefore, we ultimately adopted the threshold of 0.8 in our experiments.

Finally, we evaluated our model on its ability to classify hard samples. We compare our model against ZSARCAP [12] because both use the same set of features as input and the same training dataset (i.e., ActivityNet Captions). We observe an elevated increment in accuracy arising from improving the visual descriptor. This is an excellent indication that, as shown in Fig. 1, our joint space can approximate visual features of their semantic descriptions, narrowing the semantic gap and making the visual features less subject to domain shift. To investigate in more detail the relationship between the information modalities and the semantic gap, we choose a subset of 15 classes from UCF-101 that are hard examples due to their high intra-class similarity. These classes can be divided into six groups: (1 – using horses) *horse riding*, and *horse race*; (2 – performing gymnastics) *pommel horse*, *balance beam*, and *floor gymnastics*; (3 – using basketballs) *basketball* and *basketball dunk*; (4 – boxing) *boxing punching bag* and *boxing speed bag*; (5 – involving the face) *apply eye makeup*, *apply lipstick*, and *brushing teeth*; (6 – involving the hair) *blow dry hair*, *haircut*, and *head massage*. This subset is particularly hard because 30 random runs of 15 classes get $70.4 \pm 6.3\%$ of accuracy against 63.3% (our model with the 15 selected classes). Fig. 3 shows the confusion matrices for this subset.

When comparing each group's results for ZSARCAP and our method, we observed a reduction in confusion for all groups except 3 and 4. In group 3, our model was prone to classify all samples as basketball. On the other hand, in group 4, our model tends to classify *boxing* videos as *boxing punching bag*. We believe the inclusion of object semantics was not beneficial in these cases, although, overall, the results have been considerably better (63.3% against 47.4%). Demonstrating that the most significant performance gain came from an effective reduction in the semantic gap and not just from the inclusion of more semantic information.



(a) Acc. 47.4%



(b) Acc. 63.3%

Fig. 3. (a) ZSARCAP [12] results encoded with SBERT; (b) CEZSAR ($VE + O$).

6 Conclusions

Our conclusions are threefold: (i) contrastive learning is a straightforward yet effective approach for bridging the semantic gap between different information modalities in ZSAR. Comparing our approach against other architectures or training schemes, our results demonstrated a significant reduction in the semantic gap while allowing for easy inclusion of semantic information from other approaches without the need to retrain the contrastive model; (ii) conditioning the learning of visual features to a modality that is less impacted by the problem, such as texts, naturally reduces the domain shift problem. In all evaluated scenarios, the inclusion of text, even from a different domain, was beneficial to ZSAR performance and brought the samples closer to their corresponding prototypes; and (iii) automatic negative sampling is a practical method for augmenting a dataset without severely increasing the time required for the pre-computation of features, thus enabling training to be completed in just a few hours on a single commercial GPU. In future works, we intend to investigate the influence of the pre-training dataset size on contrastive learning performance and the impact of including images as label prototypes on the semantic gap.

Acknowledgments

This study was supported in part by the *Programa de Excelência Acadêmica (PROEX) da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* (CAPES), and by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico* (CNPq) under grants # 315409/2023-1 and # 304836/2022-2. We also gratefully acknowledge the *Pontifícia Universidade Católica do Paraná* and *Fundação Araucária* for their financial support, which enabled conference participation. We further thank PROEQ-IFPR for providing the equipment used in this research.

References

1. Balntas, V., Riba, E., Ponsa, D., Mikolajczyk, K.: Learning local feature descriptors with triplets and shallow convolutional neural networks. In: British Machine Vision Conference (BMVC). pp. 1–11 (2016). https://doi.org/10.5244/C.30.119
2. Brattoli, B., Tighe, J., Zhdanov, F., Perona, P., Chalupka, K.: Rethinking zero-shot video classification: End-to-end training for realistic applications. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4613–4623 (Jun 2020). https://doi.org/10.1109/CVPR42600.2020.00467
3. Bretti, C., Mettes, P.: Zero-shot action recognition from diverse object-scene compositions. In: British Machine Vision Conference (BMVC). pp. 1–14 (Nov 2021)
4. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the kinetics dataset. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4724–4733 (Jul 2017). https://doi.org/10.1109/CVPR.2017.502
5. Chen, S., Huang, D.: Elaborative rehearsal for zero-shot action recognition. In: IEEE/CVF International Conference on Computer Vision (ICCV). pp. 13638–13647 (Oct 2021)

6. Chen, X., et al.: AnyDoor: Zero-shot object-level image customization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6593–6602 (2024). https://doi.org/10.1109/CVPR52733.2024.00630
7. Chopra, S., Hadsell, R., LeCun, Y.: Learning a similarity metric discriminatively, with application to face verification. In: IEEE Computer Vision and Pattern Recognition (CVPR). pp. 539–546 (2005). https://doi.org/10.1109/CVPR.2005.202
8. Doshi, K., et al.: A Multimodal Benchmark and Improved Architecture for Zero Shot Learning . In: IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 2010–2019 (2024). https://doi.org/10.1109/WACV57701.2024.00202
9. Estevam, V., Laroca, R., Pedrini, H., Menotti, D.: Global semantic descriptors for zero-shot action recognition. IEEE Signal Processing Letters **29**, 1843–1847 (2022). https://doi.org/10.1109/LSP.2022.3200605
10. Estevam, V., Laroca, R., Pedrini, H., Menotti, D.: Dense video captioning using unsupervised semantic information. Journal of Visual Communication and Image Representation **107**, 104385 (2025). https://doi.org/10.1016/j.jvcir.2024.104385
11. Estevam, V., Pedrini, H., Menotti, D.: Zero-shot action recognition in videos: A survey. Neurocomputing **439**, 159–175 (2021). https://doi.org/10.1016/j.neucom.2021.01.036
12. Estevam, V., 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). https://doi.org/10.1007/s11042-023-16566-5
13. Gowda, S.N.: Synthetic sample selection for generalized zero-shot learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 58–67 (2023). https://doi.org/10.1109/CVPRW59228.2023.00011
14. Gowda, S.N., Moltisanti, D., Sevilla-Lara, L.: Continual learning improves zero-shot action recognition. In: Asian Conference on Computer Vision (ACCV). p. 403–421 (2024). https://doi.org/10.1007/978-981-96-0908-6_23
15. Gowda, S.N., Sevilla-Lara, L., Keller, F., Rohrbach, M.: CLASTER: Clustering with reinforcement learning for zero-shot action recognition. In: European Conf. on Computer Vision (ECCV). pp. 187–203 (2022)
16. Han, Z., Fu, Z., Chen, S., Yang, J.: Contrastive embedding for generalized zero-shot learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2371–2381 (2021). https://doi.org/10.1109/CVPR46437.2021.00240
17. Huang, K., Miralles-Pechuán., L., Mckeever., S.: Combining text and image knowledge with GANs for zero-shot action recognition in videos. In: International Conference on Computer Vision Theory and Applications (VISAPP). pp. 623–631 (2022). https://doi.org/10.5220/0010903100003124
18. Huang, K., Miralles-Pechuán, L., Mckeever, S.: Enhancing zero-shot action recognition in videos by combining gans with text and images. SN Computer Science **4**(4), 375 (2023). https://doi.org/10.1007/s42979-023-01803-3
19. Kerrigan, A., Duarte, K., Rawat, Y., Shah, M.: Reformulating zero-shot action recognition for multi-label actions. In: International Conference on Neural Information Processing Systems (NeurIPS). vol. 34, pp. 25566–25577 (2021)
20. Kim, T.S., et al.: DASZL: Dynamic action signatures for zero-shot learning. In: AAAI Conference on Artificial Intelligence (2021)
21. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Niebles, J.C.: Dense-captioning events in videos. In: IEEE International Conference on Computer Vision (ICCV). pp. 706–715 (2017). https://doi.org/10.1109/ICCV.2017.83

22. Lee, J.C., Lee, D.G.: ESC-ZSAR: Expanded semantics from categories with cross-attention for zero-shot action recognition. Expert Systems with Applications **255**, 124786 (2024). https://doi.org/10.1016/j.eswa.2024.124786
23. Li, X., Yang, X., Wei, K., Deng, C., Yang, M.: Siamese contrastive embedding network for compositional zero-shot learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 9326–9335 (June 2022)
24. Lin, C.C., et al.: Cross-modal representation learning for zero-shot action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19978–19988 (June 2022)
25. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3337–3344 (2011). https://doi.org/10.1109/CVPR.2011.5995353
26. Ma, P., Lu, H., Yang, B., Ran, W.: GAN-MVAE: A discriminative latent feature generation framework for generalized zero-shot learning. Pattern Recognition Letters **155**, 77–83 (2022). https://doi.org/10.1016/j.patrec.2022.02.002
27. Mettes, P., Thong, W., Snoek, C.: Object priors for classifying and localizing unseen actions. International Journal of Computer Vision **129**, 1954–1971 (2021). https://doi.org/10.1007/s11263-021-01454-y
28. Mettes, P., Snoek, C.G.M.: Spatial-aware object embeddings for zero-shot localization and classification of actions. In: IEEE International Conference on Computer Vision (ICCV). pp. 4453–4462 (2017). https://doi.org/10.1109/ICCV.2017.476
29. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). vol. 139, pp. 8748–8763 (2021)
30. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 3982–3992 (2019)
31. Soomro, K., Zamir, A.R., Shah, M.: UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint **arXiv:1212.0402**, 1–6 (2012)
32. Sun, S., et al.: CLIP as RNN: Segment countless visual concepts without training endeavor. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13171–13182 (2024)
33. Vaswani, A., et al.: Attention is all you need. In: International Conf. on Neural Information Processing (NeurIPS). pp. 6000–6010 (2017)
34. Wang, Q., Chen, K.: Zero-shot visual recognition via bidirectional latent embedding. Intl. Journal of Computer Vision **124**(3), 356–383 (2017). https://doi.org/10.1007/s11263-017-1027-5
35. Wu, W., et al.: Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6620–6630 (2023). https://doi.org/10.1109/CVPR52729.2023.00640
36. Xu, H., et al.: VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In: Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 6787–6800 (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.544
37. Xue, Y., Whitecross, K., Mirzasoleiman, B.: Investigating why contrastive learning benefits robustness against label noise. In: International Conf. on Machine Learning (ICML). vol. 162, pp. 24851–24871 (2022)