A Multimodal Benchmark and Improved Architecture for Zero Shot Learning

Keval Doshi, Amanmeet Garg, Burak Uzkent, Xiaolong Wang, Mohamed Omar
Amazon Prime Video
{kcdos, amanmega, burauzke, xiaowanf, omarmk}@amazon.com

Abstract

In this work, we demonstrate that due to the inadequacies in the existing evaluation protocols and datasets, there is a need to revisit and comprehensively examine the multimodal Zero-Shot Learning (MZSL) problem formulation. Specifically, we address two major challenges faced by current MZSL approaches: (1) Established baselines are frequently incomparable and occasionally even flawed since existing evaluation datasets often have some overlap with the training dataset, thus violating the zero-shot paradigm; (2) Most existing methods are biased towards seen classes, which significantly reduces the performance when evaluated on both seen and unseen classes. To address these challenges, we first introduce a new multimodal dataset for zero-shot evaluation called MZSL-50 with 4462 videos from 50 widely diversified classes and no overlap with the training data. Further, we propose a novel multimodal zero-shot transformer (MZST) architecture that leverages attention bottlenecks for multimodal fusion. Our model directly predicts the semantic representation and is superior at reducing the bias towards seen classes. We conduct extensive ablation studies, and achieve state-of-the-art results on three benchmark datasets and our novel MZSL-50 dataset. Specifically, we improve the conventional MZSL performance by a margin of 2.1%, 9.81% and 8.68% on VGG-Sound, UCF-101 and ActivityNet, respectively. Finally, we expect the introduction of the MZSL-50 dataset will promote the future in-depth research on multimodal zero-shot learning in the community.

1. Introduction

In existing literature, multimodal zero-shot learning (MZSL) can be broadly classified into two settings, the conventional zero-shot setup which assumes only previously unseen classes are available at test time, and the generalized zero-shot setup where the test samples belong to both seen and unseen classes. To address practical challenges such as domain adaptation and out-of-distribution examples, recent works employ off-the-shelf pre-trained action recognition models to extract video and audio features [3, 11, 18–20, 23, 31, 34, 35, 51]. Moreover, due to the lack of a dataset specifically designed for MZSL, most existing works evaluate the performance on small scale datasets such as UCF-101 [43], HMDB-51 [28], VGG-Sound [6] and ActivityNet [12]. However, this setup has several limitations. As shown in [3, 9], classes that are considered as unseen, are also present in the training set, which clearly violates the conventional zero-shot paradigm. To circumvent this problem, recent research works propose to remove the overlapping classes from large scale datasets such as Kinetics-400/600/700, and train action recognition models on these modified datasets [3, 34, 35]. Considering the lack of general consensus on a fair zero-shot setup, there have been several formulations proposed leading to multiple evaluation setups, as shown in Table 1. To tackle these problems and provide a consistent framework for zero-shot evaluation, we introduce a novel dataset called MZSL-50, which consists of 4462 videos from 50 classes. MZSL-50 does not overlap with any of existing benchmark datasets a shown in Fig. 1, thus eliminating the need for creating dataset splits.

The proposed dataset will be released in public domain for future research use upon publication of the manuscript.
Table 1. Multiple ZSL splits proposed in existing works. \( R \) indicates the dataset classes were randomly divided to generate train/test splits, whereas \( D \) represents a deterministic split based on class overlap. In our proposed formulation (\( O \)), all classes from the training dataset are considered as seen, and MZSL-50 classes \( (1) \) are considered as unseen.

<table>
<thead>
<tr>
<th>Formulation</th>
<th>Method</th>
<th>Multimodal Split</th>
<th>No Overlap with Kinetics 400/600/700</th>
<th>Type</th>
<th>UCF (Seen/Unseen)</th>
<th>ActivityNet (Seen/Unseen)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Remove training classes</td>
<td>EZL [4]</td>
<td>✔</td>
<td>✔</td>
<td>( R )</td>
<td>50/51</td>
<td>100/100</td>
</tr>
<tr>
<td></td>
<td>VSEIT [9]</td>
<td>✔</td>
<td>✔</td>
<td>( R )</td>
<td>50/51</td>
<td>100/100</td>
</tr>
<tr>
<td></td>
<td>CondDE [11]</td>
<td>✔</td>
<td>✔</td>
<td>( R )</td>
<td>50/51</td>
<td>100/100</td>
</tr>
<tr>
<td>Remove evaluation classes</td>
<td>ToteZe [21]</td>
<td>✔</td>
<td>✔</td>
<td>( D )</td>
<td>-71</td>
<td>-135</td>
</tr>
<tr>
<td></td>
<td>Malten [10]</td>
<td>✔</td>
<td>✔</td>
<td>( D )</td>
<td>-78</td>
<td>-135</td>
</tr>
<tr>
<td></td>
<td>VSST [9]</td>
<td>✔</td>
<td>✔</td>
<td>( D )</td>
<td>-8</td>
<td>-19</td>
</tr>
<tr>
<td></td>
<td>AVCA [30]</td>
<td>✔</td>
<td>✔</td>
<td>( D )</td>
<td>42/9</td>
<td>150/50</td>
</tr>
<tr>
<td></td>
<td>TCAP [35]</td>
<td>✔</td>
<td>✔</td>
<td>( D )</td>
<td>45/6</td>
<td>152/48</td>
</tr>
<tr>
<td>Proposed</td>
<td>Ours</td>
<td>✔</td>
<td>✔</td>
<td>( O )</td>
<td>101/76</td>
<td>206/70</td>
</tr>
</tbody>
</table>

2 We will release our novel dataset MZSL-50 publicly upon publication of our paper.

Figure 2. Comparison between existing and proposed approach towards multimodal zero shot learning (MZST). Unlike existing methods, our model performs end-to-end training for MZST.

2. Related Works

Zero Shot Learning: Action recognition has been extensively studied over the past several years [5, 14–16, 42, 47]. In contrast, ZSL for action recognition has only recently started gaining attention. Broadly, ZSL can be classified into the conventional setting [3, 7, 9, 11, 22, 31], where approaches are evaluated only on the unseen classes, and the generalized setting [27, 33–35, 40].

Multimodal Representation Learning: Action recognition research works [30, 42, 46] have mostly focused on learning representations from video frames. Only recently the interest in the multimodal domain learning for action recognition has picked up [1, 4]. Recently, transformer based architectures have shown strong performance in learning from multiple modalities. In [37], an attention bottleneck architecture is proposed to fuse feature tokens from all modalities. Similarly, there has been recent interest in the zero-shot action recognition community to leverage multiple modalities. Specifically, Mercea et al. [34, 35] jointly learn audio and video features using cross-attention blocks, whereas Mazumder et al. [33] propose using a composite triplet loss for aligning audio and video embeddings. In [31], a cross-modal approach is proposed to jointly learn
the visual and textual features using a transformer architecture, similar to BERT [8].

**Visual Feature Extraction:** To extract the visual features, most recent approaches [2, 3, 11, 18–20, 23, 48–51] propose using a 3D-CNN, which takes 16 frames sampled from a video as input. In [3], Brattoli et al. propose training a C3D [44] and a R(2+1)D model [45] in an end-to-end fashion for ZSL. On the other hand, Gowda et al. [11] propose a reinforcement learning based clustering approach, which uses a two-stream I3D [5] model for learning visual features. Similarly, it is well understood that hierarchical representations capture scale variation in the data and learn concepts that vary with scale. Recent works learn multi scale feature representation from video [30] and show strong performance in video classification.

**Zero Shot Evaluation:** Several approaches have extended the work of Roitberg et al. [41] to formulate a novel evaluation protocol that satisfies the ZSL paradigm. Particularly, Brattoli et al. [3] proposes removing certain classes from the training set which overlap with the test set by using semantic embedding matching. However, Doshi et al. [9] show that such an approach fails to remove all the overlapping classes, and propose a new evaluation split called Fair ZSL, composed of non-overlapping classes from benchmark datasets. Alternatively, Gowda et al. [22] also propose a TruZe split for the UCF-101 [43] and HMDB-51 [28] datasets, by manually removing all classes which overlap with the Kinetics-400 dataset. Unfortunately, these works fail to reach a common consensus regarding the evaluation splits, leading to multiple evaluation setups as shown in Table 1.

### 3. The MZSL-50 dataset

In this section, we present MZSL-50, a multimodal dataset composed of carefully chosen action classes that has no overlap with existing benchmark datasets. It is, to the best of our knowledge, the first dataset that has been specifically tailored for evaluation in multimodal zero-shot learning. We begin by discussing the motivation for proposing a new dataset, followed by a thorough analysis of the annotation protocol and dataset statistics. Finally, we present the evaluation procedure for both conventional and generalized zero-shot setups.

**Motivation:** The purpose of MZSL-50 is to establish a unified evaluation benchmark for multimodal zero-shot learning, such that it does not violate the zero-shot paradigm. As shown in Table 1, existing approaches lack a general consensus regarding classes that overlap with benchmark training datasets, which has lead to several proposed splits, increasing ambiguity in comparing related works. It is also important to take into account, in case of an unseen class, completely unrelated to any other seen classes (e.g. cooking omelette vs. playing tennis). Intuitively, even humans struggle to understand novel activities when they include unfamiliar relationships and objects. Thus, we also consider the semantic relatedness of the classes in the proposed dataset with respect to the 4 benchmark training datasets used in existing works. We hope that the straightforward way in which we have formulated the training and evaluation protocols will make it simple for future researchers to assess and compare their MZSL performance.

**Design:** To ensure a realistic MZSL setting and avoid an overlap with seen classes, we specifically take benchmark datasets used in the research community as our reference. Specifically, we consider the actions included in Kinetics-400/600/700, VGG-Sound, UCF-101 and ActivityNet as the seen classes, and collect videos for classes that are semantically related, but not identical to these seen classes. Formally, we define the semantic relatedness (SR) score as:

\[
SR(class) = \min D_{cos}(\phi(class), \phi(X^S)),
\]

where \(D_{cos}\) is the cosine distance, \(\phi(class)\) is the semantic embedding of an unseen class and \(\phi(X^S)\) is the set of semantic embeddings of all the classes in the benchmark datasets. We use Word2Vec [36] to parameterize function \(\phi\) for extracting the embeddings. The semantic similarity of classes in the training and evaluation sets can directly impact the model performance on individual classes. In order to quantify this impact, we assign a SR score to all the classes in the proposed evaluation set. A low SR value implies high semantic similarity, thus easy for the model to recognize, similarly, a high SR would imply a difficult case for the model to predict on. Empirically, we see that SR should be higher than 0.1 to avoid including identical classes, and less than 0.8 to avoid classes that are significantly different from the seen classes. Hence, we divide our classes into 3 sets based on the SR score: easy \((0.1 < SR \leq 0.33)\), medium \((0.33 < SR \leq 0.66)\) and hard \((0.66 < SR \leq 0.8)\) classes.

**Statistics:** The final dataset consists of 4462 videos for a total duration of 405 hours covering 50 classes. Following our dataset design, we further divide the videos into easy, medium, and hard sub-groups based on the semantic relatedness score. The Table 2 outlines the data in each subclass.

<table>
<thead>
<tr>
<th>Difficulty</th>
<th># Classes</th>
<th># Videos</th>
</tr>
</thead>
<tbody>
<tr>
<td>Easy</td>
<td>9</td>
<td>926</td>
</tr>
<tr>
<td>Medium</td>
<td>31</td>
<td>2811</td>
</tr>
<tr>
<td>Hard</td>
<td>10</td>
<td>725</td>
</tr>
<tr>
<td>Total</td>
<td>50</td>
<td>4462</td>
</tr>
</tbody>
</table>

Table 2. Statistics for difficulty levels based on semantic relatedness score (SR) on MZSL-50.

**Annotation protocol:** Given a video clip, the goal of an-
Figure 3. The sample distribution of all classes in MZSL-50 dataset. The color of the bars represent the Semantic Relatedness score ranging from 0.1 (blue, slapping) to 0.8 (red, surgery) across the 50 classes. Semantic Relatedness represents the maximal similarity value of an unseen class to a seen class in the training dataset.

Figure 4. Example filmstrips and statistics visualization for the dataset.

notators is to validate the presence of a sequence of frames related to the class of interest. Hence, to create this new evaluation dataset, we source public videos from top 300 results for each class from YouTube. In order to improve the annotation efficiency, we create a filmstrip using frames of a video sampled at 1 fps arranged into a 10x10 grid. The annotators are instructed to visually check the filmstrip and assert the presence of frames representing the action category of the class. For example, in making juice class (Fig. 4a) the presence of row 6-8 with juice frames would qualify this video as a positive sample. The videos were annotated by multiple reviewers to avoid negative examples, and we remove videos which had mixed annotations.

MZSL-50 Evaluation: For conventional and generalized zero-shot setups, we present a unified training and evaluation protocol. Instead of splitting the dataset and training on a subset of classes, we train the model on all classes from the benchmark datasets, such as (VGG-Sound, UCF-101 and ActivityNet). The model is then evaluated solely on MZSL-50 using the standard zero-shot setup. The final metric on the MZSL-50 dataset is reported as the weighted average of the performance in the easy, medium and hard subsets (weighted on number of classes in each subset). We evaluate the generalized zero-shot setup on both MZSL-50 and the respective benchmark dataset.

4. Multimodal Zero Shot Transformer

4.1. Preliminaries

We can formally define MZSL as a classification problem, where given a tuple of video, audio and text class labels \((V^*, A^*, T)\) as training data from \(S\) seen classes \(\{(v_1^*, a_1^*, t_1), \ldots, (v_N^*, a_N^*, t_N)\}\), we aim to accurately classify video and audio \(X^v = \{(v_1^*, a_1^*), \ldots, (v_M^*, a_M^*)\}\) from previously unseen classes \(U\), where \(N\) and \(M\) are the number of training and testing videos respectively. Ideally, to satisfy the conventional zero-shot learning paradigm, there should be no overlap between the seen and unseen classes, i.e., \((S \cap U) = \emptyset\). On the other hand, in the generalized zero-shot setup, a model is evaluated on both seen and unseen classes, which requires a broad generalization capacity. Conventionally, semantic embeddings are used as
Figure 5. The framework overview of Multiscale Zero-Shot Transformer, MZST, where multimodal inputs are frame sequence $V_i$ and audio spectrogram $A_i$ from the $i$-th video. We leverage the recently proposed MViT-v2 and AST architectures along with an audio-video-text (AVT) fusion network to directly predict the semantic embedding.

4.2. Proposed Model Architecture

**MViT & AST Architecture:** Hierarchical representations often outperform their single scale counterparts in recognition tasks [30]. We take a pretrained MViT [29] model as our backbone to provide multiscale embeddings for the video frames. Inspired by the recent strong performance by the MViT model [30], we extend the model architecture to learn hierarchical representations for the underlying data for video inputs. Recently, Audio Spectrogram Transformer (AST) [21] has utilized audio spectrograms as images to extract patch based embedding tokens from patches of the spectrograms. We leverage the AST architecture to obtain token embeddings for each audio sample. Specifically, we extract audio tokens $[E_1^A, \ldots, E_M^A]$ using $M$ non-overlapping patches from the input spectrogram and $N$ video tokens $[E_1^V, \ldots, E_N^V]$ from the input video.

**AVT Fusion Network:** Previous cross-modality transformers either simply concatenated multimodal representations [1], or exchanged the key and value matrices between the two modalities in the attention block [24]. On the other hand, inspired by the cross modality fusion between audio and image transformers [37], we construct an audio-video-text (AVT) fusion network by leveraging bottleneck transformers, which handles varied lengths of modality tokens efficiently as illustrated in Fig. 5.

Let $\{E_i^F, \ldots, E_L^F\}$ be the $L$ initial multimodal fusion tokens. During training, the fusion tokens are input to the transformer block alternatively with the joint video-audio tokens and text tokens. The model is forced to learn the fusion tokens with attention based token update along all three modalities. For example, for the joint video-audio token update we perform following operations formulated as

$$
\hat{E}_{AVF} = \text{FFN}(\text{LN}(\tilde{E}_{AVF})) + \tilde{E}_{AVF}, \\
\tilde{E}_{AVF} = \text{MSA}(\text{LN}(E_{AVF})) + E_{AVF},
$$

(2)

where $E_{AVF} = [E_{CS}^V, E_1^V, \ldots, E_N^V, E_1^A, \ldots, E_M^A, E_1^F, \ldots, E_L^F]$ whereas FFN represents the Feed Forward Neural Network and LN represents the layer norm.
We repeat the above operations for the text tokens such that $E^{TF} = [E^{LS}_1, E^{TF}_1, \ldots, E^{TF}_M, E^{AV}_1, \ldots, E^{AV}_F]$. Given modality-specific tokens, multimodal tokens can be updated by averaging the multimodal tokens along the AVT bottleneck blocks. The AVT fusion network consists of $K$ stacked blocks.

4.3. Loss Function

**Masked Language Modeling Loss:** Inspired by the Masked Language Modeling (MLM) task of BERT transformer [8], we apply the MLM loss to the discrete token sequence for input text. We randomly mask the input text tokens so as to force the bottleneck architecture to predict these masked tokens ($E^{TF}_o$) based on their surrounding word tokens ($E^{TF}_r$) and joint audio-visual feature tokens $E^{AV}F$ by minimizing the negative log-likelihood:

$$L_{MLM} = -\frac{1}{n} \sum_{i=1}^{n} \log p_i^{AV} (E^{TF}_o | E^{TF}_r, E^{AV}F). \tag{3}$$

The purpose of using MLM loss is to align and learn the dependencies between visual and audio content and semantic concepts.

**Semantic Embedding Loss:** The supervised semantic embedding loss is formulated as:

$$L_{SE} = \| f(x^i_t) - (\phi(s^i_t) + \phi(s^i_t)) \|^2, \tag{4}$$

where $f(x^i_t)$ is the output of the FFN and $\phi(s^i_t), \phi(s^i_t)$ are the semantic embedding for the training video $x^i_t$ from the class label of the pretrained video model and the ground truth label, respectively, extracted using the Word2Vec model [38].

**Task Loss:** We perform multimodal classification by passing the joint video-audio representation $[E^{AV}]$ through a fully connected layer. We use cross entropy loss, $L_{TASK}$, for this objective.

**Combined Loss Function:** Finally, we combine all the objectives linearly as

$$L = L_{TASK} + L_{SE} + \sigma \cdot L_{MLM} \tag{5}$$

where $\sigma \in [0, 1]$. We provide the sensitivity analysis for $\sigma$ in the supplementary material.

5. Experiments

We evaluate the proposed approach and compare with other state-of-the-arts both on three popular benchmarks and newly collected MZSL-50. In this section, we first give the description of dataset, then provide the implementation details. After that, we list the experimental results and ablation studies.

5.1. Experimental Setup

**Benchmark Datasets:** We first evaluate our proposed approach on our new MZSL-50 dataset and three benchmark datasets, namely VGGSound [6], UCF-101 [43] and ActivityNet [12] datasets. **VGGSound** is a large scale action recognition dataset, which consists of about 200K 10-second clips and 309 categories ranging from human actions and sound-emitting objects to human-object interactions. Like other YouTube datasets, e.g., K400 [26], some clips are no longer available. After removing invalid clips, we collect 159,223 valid training multimodal videos and 12,790 valid test multimodal videos. **UCF-101** consists of over 13k videos in 101 classes. We only consider classes which include the audio modality, leading to 6,816 videos from 51 classes. **ActivityNet** consists of videos from 200 classes related to daily activities and is considerably more comprehensive, consisting of 27,801 videos from 200 classes.

For a fair comparison with existing state-of-the-art approaches, we use the same training and evaluation splits as proposed in [34].

**Comparing with State-of-the-Art:** We compare our proposed approach to recent state-of-the-art multimodal ZSL approaches TCaF [35], AVCA [34], AVGZSLNet [33] and CIME [39]. AVCA leverages cross-attention mechanism and combines information from video and audio modalities by temporally averaging them. TCaF builds upon AVCA and applies an improved cross-attention mechanism which introduces temporal attention in addition to spatial attention. Furthermore, we also compare to image based approaches Attention Fusion [13], Perceiver [25] and DeViSE [17], which are adapted for multimodal inputs by [34, 35].

**Implementation Details:** We employ 16 frames for multiscale video Transformer [29] along with 3 spatial crops and 4 ensemble views during inference. We are able to train the model using a batch size of 64 on 8 NVIDIA A100 GPUs, each with 40 GB of memory. AdamW [32] is used in the backpropagation and the learning rate is set as 0.0001. The number of epochs is set as 100. We set $\lambda_1$ and $\lambda_2$ as 0.25, 0.25 for the first 20 epochs, 0.1, 0.1 from the 21- to 40-th epochs, and 0.05, 0.05 after the 40-th epochs. These hyperparameters are generally set to tune the loss values into the same scale. We sample audio clips at 16kHz and convert them to mono channel. We extract log mel spectrograms with a frequency dimension of 128. The AST model is initialized with ImageNet weights. The bottleneck tokens are initialized using a Gaussian distribution of mean 0 and standard deviation of 0.02. More details are available in the supplementary material.

**Evaluation metrics:** We follow the evaluation protocol discussed in [34], and propose to evaluate all models using the mean class accuracy. For generalized MZSL, we eval-
evaluate the proposed model on both seen and unseen classes, and report the harmonic mean. The harmonic mean is given by:

$$HM = \frac{2 \times S \times U}{S + U}$$

where $S, U$ are the mean class accuracies on the seen and unseen classes, respectively.

5.2. Results on MZSL-50

In Table 3, we show the performance of our approach based on the semantic relatedness of the classes, specifically the easy, medium and hard cases. We can clearly observe that the semantic relatedness metric is an excellent indicator for estimating the performance of a model on a certain class, since all approaches notice a drop in performance as the classes become more difficult. This shows that the semantic relatedness can be used as a reliable metric for designing datasets which have classes between a certain range, unlike existing datasets such as UCF-101 which has 80% of the classes with a SR score of less than 0.2.

<table>
<thead>
<tr>
<th>Model</th>
<th>Modality</th>
<th>VGG-Sound/MZSL-50</th>
<th>ActivityNet/MZSL-50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>S M H Avg</td>
<td>S M H Avg</td>
</tr>
<tr>
<td>AVCA [40]</td>
<td>A,V</td>
<td>32.13 15.29 4.78 17.4</td>
<td>31.78 14.12 5.03 16.97</td>
</tr>
</tbody>
</table>

Table 3. MZSL performance on the MZSL-50 dataset when using audio and visual features as inputs on the VGG-Sound and ActivityNet datasets. We report the performance on easy (E), medium (M) and hard (H) classes, as well as the average (Avg) class accuracy. We can clearly see that MZST outperforms the recent SOTA methods by a wide margin.

Moreover, we also compare our proposed framework to recent approaches under the conventional zero-shot setup. As shown in Fig. 2, under the conventional setup, we finetune our model on all the classes of the benchmark dataset and evaluate on the MZSL-50 dataset. We only finetune on classes from VGG-Sound and ActivityNet since very few classes in UCF-101 contain both video and audio. We observe that the proposed approach achieves a significantly higher performance (Table 4) across all settings. We were unable to compare our model performance to TCaF [35], since it required features extracted at separate temporal intervals, whereas all the other recent approaches use averaged features.

<table>
<thead>
<tr>
<th>Model</th>
<th>Modality</th>
<th>VGG-Sound/MZSL-50</th>
<th>ActivityNet/MZSL-50</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>S U HM ZS</td>
<td>S U HM ZS</td>
</tr>
<tr>
<td>AVCA</td>
<td>A,M</td>
<td>18.27 9.15 12.19 11.37</td>
<td>8.05 7.25 7.66 9.03</td>
</tr>
</tbody>
</table>

Table 4. Audio-visual MZSL results under the generalized zero-shot setting when using audio and visual features as inputs on the VGG-Sound, UCF and ActivityNet datasets.

5.3. Results on Benchmark Dataset

We compare our proposed framework to recent approaches under the conventional zero-shot setup in Table 5 and under the generalized zero-shot setup in Table 6. For a fair comparison, we use the same splits proposed in [34]. The multiscale video model is trained on 595 classes from Kinetics, after removing the overlapping classes, as proposed in [3]. We observe that the proposed approach achieves a significantly higher performance, by a margin of 18.03%, 20.23% and 20.58% as compared to [34]. To have a fair comparison with recent state-of-the-art approaches, we also evaluate the proposed approach using the features extracted using [34], and yet have a significant performance gap. Specifically, we outperform recent approaches by a margin of 2.1%, 9.81% and 8.68%, which shows the efficacy of our approach.

5.4. Ablation Studies

**Impact of pretraining datasets:** We experiment with various backbones to study the impact of pretraining on the zero-shot performance. As shown in Table 5, we see that pretraining the video model on K595 significantly increases
Table 5. Audio-visual MZSL results under the zero-shot setting on benchmark datasets. VGG† pretraining implies self-supervised SeLAVi features used by state-of-the-art approaches [34, 35]

Table 6. Audio-visual MZSL results under the generalized zero-shot setting when using audio and visual features as inputs on the VGG-Sound, UCF, and ActivityNet datasets.

Table 7. Results on comparing the impact of different fusion techniques on the performance in the MZSL task.

the performance as compared to using self-supervised pretraining on VGG-Sound. Moreover, it is interesting to observe that audio model initialized with imagenet weights performs better than AVCA pretrained on VGG-Sound, which further demonstrates the efficacy of our proposed model (Table 7).

Evaluating fusion mechanisms: Next, we investigate the effect of using our AVT fusion network and loss functions in Table 7. To obtain results without audio-video fusion, each branch is optimised individually. For evaluation, we simply concatenate the video and audio embeddings. We observe that except for UCF-101, audio-video fusion consistently performs better than feature concatenation. The reason for feature concatenation performing better on UCF-101 can be attributed to the dataset being spatially heavy and video being the dominant modality.

Impact of variation in semantic embedding: As illustrated in Fig. 7, we compare and visualize the projected semantic embedding of the proposed approach to the Word2Vec semantic embedding of the recent SOTA approach. We see that normalizing the semantic feature representation deteriorates its inherent relatedness. For example, before normalizing the embeddings, Shaving Beard is semantically similar to Haircut and Brushing Teeth, but normalizing the embeddings cause it to be similar to Javelin Throw. We can also see that there is a noticeable improvement in the performance (Table 8) when non-normalized semantic embeddings are used, which ascents our conjecture.

6. Conclusion

We first propose a novel dataset called MZSL-50 for multimodal zero-shot action recognition. The MZSL-50 allows for a unified formulation and proposes a comprehensive evaluation protocol that strictly adheres to the zero-shot premise. Additionally, we propose an end-to-end multimodal transformer called MZST that outperforms existing approaches by a wide margin on both existing datasets and on MZSL-50. Specifically, we outperform the state-of-the-art approaches by a wide margin on both existing datasets and on MZSL-50. This demonstrates the superiority of our proposed method.
References


Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. arXiv preprint arXiv:2112.01526, 2021. 5, 6
Otniel-Bogdan Mercea et al. Audio-visual generalised zero-shot learning with cross-modal attention and language. In CVPR, 2022. 1, 2, 6, 7, 8
Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems, 26, 2013. 3