Temporal and cross-modal attention for audio-visual zero-shot learning

Otniel-Bogdan Mercea*[1], Thomas Hummel*[1], A. Sophia Koepke[1], and Zeynep Akata[1,2,3]

[1] University of Tübingen [2] MPI for Informatics [3] MPI for Intelligent Systems
{otniel-bogdan.mercea, thomas.hummel, a-sophia.koepke,zeynep.akata}@uni-tuebingen.de

Abstract. Audio-visual generalised zero-shot learning for video classification requires understanding the relations between the audio and visual information in order to be able to recognise samples from novel, previously unseen classes at test time. The natural semantic and temporal alignment between audio and visual data in video data can be exploited to learn powerful representations that generalise to unseen classes at test time. We propose a multi-modal and Temporal Cross-attention Framework (TCAF) for audio-visual generalised zero-shot learning. Its inputs are temporally aligned audio and visual features that are obtained from pre-trained networks. Encouraging the framework to focus on cross-modal correspondence across time instead of self-attention within the modalities boosts the performance significantly. We show that our proposed framework that ingests temporal features yields state-of-the-art performance on the UCF-GZSL^{cls}, VGGSound-GZSL^{cls}, and ActivityNet-GZSL^{cls} benchmarks for (generalised) zero-shot learning. Code for reproducing all results is available at https://github.com/ExplainableML/TCAF-GZSL.

Keywords: Zero-shot learning, Audio-visual learning

1 Introduction

Learning task-specific audio-visual representations commonly requires a great number of annotated data samples. However, annotated datasets are limited in size and in the labelled classes that they contain. If a model which was trained with supervision on such a dataset is applied in the real world, it encounters classes that it has never seen. To recognise those novel classes, it would not be feasible to train a new model from scratch. Therefore, it is essential to analyse the behaviour of a trained model in new settings. Ideally, a model should be able to transfer knowledge obtained from classes seen during training to previously unseen categories. This ability is probed in the zero-shot learning (ZSL) task. In addition to zero-shot capabilities, a model should retain the class-specific information from seen training classes. This is challenging and is investigated in

* Denotes equal contribution

the so-called generalised ZSL (GZSL) setting which considers the performance on both, seen and unseen classes.

Prior works [55,46,47] have proposed frameworks that address the (G)ZSL task for video classification using audio-visual inputs. Those methods learn a mapping from the audio-visual input data to textual label embeddings, enabling the classification of samples from unseen classes. At test time, the class whose word embedding is closest to the predicted audio-visual output embedding is selected. Similar to this, we use the textual label embedding space to allow for information transfer from training classes to previously unseen classes. However, [55,46,47] used temporally averaged features as inputs that were extracted from networks pre-trained on video data. The averaging disregarded the temporal dynamics in videos. We propose a Temporal Cross-attention Framework (TCAF) which builds on [47] and additionally exploits temporal information by using temporal audio and visual data as inputs. This gives a significant boost in performance for the audio-visual (G)ZSL task compared to using temporally averaged input features. Different from computationally expensive methods that operate directly on raw visual inputs [13,40,33], our TCAF uses features extracted from networks pre-trained for audio and video classification as inputs. This leads to an efficient setup that uses temporal information instead of averaging across time.

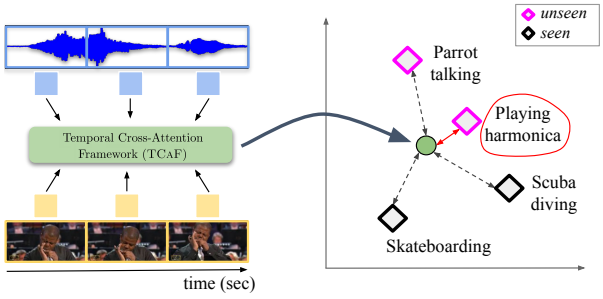


Fig. 1. Our temporal cross-attention framework for audio-visual (G)ZSL learns a multi-modal embedding (green circle) by exploiting the temporal alignment between audio and visual data in videos. Textual label embeddings (grey squares) are used to transfer information from seen training classes (black) to unseen test classes (pink). The correct class is playing harmonica (red).

The natural alignment between audio and visual information in videos, e.g. a frog being visible in a frame while the sound of a frog croaking is audible, provides a rich training signal for learning video representations. This can be attributed to the semantic and temporal correlation between the audio and visual information when comparing the two modalities. We encourage our TCAF to put special emphasis on the correlation across the two modalities by employing repeated cross-attention. This attention mechanism only allows attention to tokens from the other modality. This effectively acts as a bottleneck which results in cheaper computations and gives a boost in performance over using full self-attention across all tokens from both modalities.

We perform a detailed model ablation study to show the benefits of using temporal inputs and our proposed cross-attention. Furthermore, we confirm that our training objective is well-suited to the task at hand. We also analyse the learnt audio-visual embeddings with t-SNE visualisations which confirm that training our TCaF improves the class separation for both seen and unseen classes.

To summarise, our contributions are as follows: (1) We propose a temporal cross-attention framework TCaF for audio-visual (G)ZSL. (2) Our proposed model achieves state-of-the-art results on the UCF-GZSL^{cls}, VGGSound-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets, demonstrating that using temporal information is extremely beneficial for improving the (generalised) zero-shot classification accuracy compared to using temporally averaged features as model inputs. (3) We perform a detailed analysis of the use of enhanced cross-attention across modalities and time, demonstrating the benefits of our proposed model architecture and training setup.

2 Related work

Our work relates to several themes in the literature: audio-visual learning, ZSL with side information, audio-visual ZSL with side information, and multi-modal transformer architectures. We discuss those in more detail in the following.

Audio-visual learning. The temporal alignment between audio and visual data in videos is a strong learning signal which can be exploited for learning audio-visual representations. [53,54,7,56,37,10]. In addition to audio and video classification, numerous other tasks benefit from audio-visual inputs, such as the separation and localisation of sounds in video data [52,65,8,24,15,4,1], audio-driven synthesis of images [70,31], audio synthesis driven by visual information [77,25,36,35,59,23,50], and lip reading [3,2]. Some approaches use class-label supervision between modalities [20,16] which does not require the temporal alignment between the input modalities. In contrast to full class-label supervision, we train our model only on the subset of seen training classes.

ZSL with side information. Visual ZSL methods commonly map the visual inputs to class side information [21,6,5], e.g. word2vec [48] class label embeddings. This allows to determine the class with the side information that is closest at test time as the class prediction. Furthermore, attribute annotations have been used as side information [68,74,71,19]. Recent non-generative methods identify key visual attributes [76], use attention to find discriminative regions [75], or disambiguate class embeddings [43]. In contrast, feature generation methods train a classifier on generated and real features [73,51,78,72]. Unlike methods for ZSL with side information with unimodal (visual) inputs, our proposed framework uses multi-modal audio-visual inputs.

Audio-visual ZSL with side information. The task of GZSL from audio-visual data was introduced by [55,46] on the AudioSetZSL dataset [55] using class label word embeddings as side information. Recently, [47] proposed the AVCA framework which uses cross-attention to fuse information from the averaged audio and visual input features for audio-visual GZSL. Our proposed

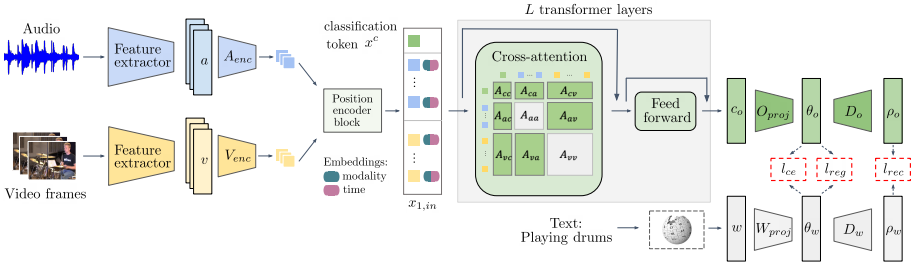


Fig. 2. TCaF takes audio and visual features extracted from video data as inputs. Those are embedded and equipped with modality and time embeddings before passing through a sequence of L transformer layers with cross-attention. The output classification token c_o is then projected to embedding spaces that are shared with the textual information. The loss functions operate on the joint embedding spaces. At test time, the class prediction c is obtained by determining the word label embedding θ_w^j that is closest to θ_o.

framework builds on [47], but instead of using temporally averaged features as inputs [47,55,46], we explore the benefits of using temporal cross-attention information. Unlike [47]'s two-stream architecture, we propose the fusion into a single output branch with a classification token that aggregates multi-modal information. Furthermore, we simplify the training objective, and show that the combination of using temporal inputs, our architecture, and training setup leads to superior zero-shot classification performance.

Multi-modal transformers. The success of transformer models in the language domain [67,17,57] has been translated to visual recognition tasks with the Vision Transformer [18]. Multi-modal vision-language representations have been obtained with a masked language modelling objective, and achieved state-of-the-art performance on several text-vision tasks [61,62,44,38,39,60,63]. In this work, we consider audio-visual multi-modality. Transformer-based models that operate on audio and visual inputs have recently been proposed for text-based video retrieval [22,42,69], dense video captioning [28], audio-visual event localization [41], and audio classification [12]. Different to vanilla transformer-based attention, our TCaF puts special emphasis on cross-attention between the audio and visual modalities in order to learn powerful representations for the (G)ZSL task.

3 TCaF Model

In this section, we describe the problem setting (Section 3.1), our proposed model architecture (Section 3.2), and the loss functions used to train TCaF (Section 3.3).

3.1 Problem setting

We address the task of (G)ZSL using audio-visual inputs. The aim of ZSL is to be able to generalise to previously unseen test classes at test time. For GZSL, the model should additionally preserve knowledge about seen training classes, since the GZSL test set contains samples from both, seen and unseen classes.

We denote an audio-visual dataset with N samples and K (seen and unseen) classes by $\mathcal{V} = \{\mathcal{X}_{\boldsymbol{a}[i]}, \mathcal{X}_{\boldsymbol{v}[i]}, y_{[i]}\}_{i=1}^{N}$, consisting of audio data $\mathcal{X}_{\boldsymbol{a}[i]}$, visual data $\mathcal{X}_{\boldsymbol{v}[i]}$, and ground-truth class labels $y_{[i]} \in \mathbb{R}^K$. Naturally, video data contains temporal information. In the following, we use T_a and T_v to denote the number of audio and visual segments in a video clip.

A pre-trained audio classification CNN is used to extract a sequence of audio features $\boldsymbol{a}_{[i]} = \{a_1, \ldots, a_t, \ldots, a_{T_a}\}_i$ to encode the audio information $\mathcal{X}_{\boldsymbol{a}[i]}$. The visual data $\mathcal{X}_{\boldsymbol{v}[i]}$ is encoded into a temporal sequence of features $\boldsymbol{v}_{[i]} = \{v_1, ..., v_t, ..., v_{T_v}\}_i$ by representing visual segments with features extracted from a pre-trained video classification network.

3.2 Model architecture

In the following, we describe the architecture of our proposed TCaF (see Fig. 2).
Embedding the inputs and position encoder block. TCaF takes pre-extracted audio and visual features $\boldsymbol{a}_{[i]}$ and $\boldsymbol{v}_{[i]}$ as inputs. For readability, we will drop the subscript i in the following which denotes the $i-$th sample. In order to project audio and visual features to the same feature dimension, $\boldsymbol{a}$ and $\boldsymbol{v}$ are passed through two modality-specific embedding blocks, giving embeddings:

$$\phi_a = A_{enc}(\boldsymbol{a}) \text{ and } \phi_v = V_{enc}(\boldsymbol{v}), \tag{1}$$

with $\phi_a \in \mathbb{R}^{T_a * d_{dim}}$ and $\phi_v \in \mathbb{R}^{T_v * d_{dim}}$. The embedding blocks are composed of two linear layers f_1^m, f_2^m for $m \in \{\boldsymbol{a}, \boldsymbol{v}\}$, where $f_1^m : \mathbb{R}^{T_m * d_{in_m}} \rightarrow \mathbb{R}^{T_m * d_{fhidd}}$ and $f_2^m : \mathbb{R}^{T_m * d_{fhidd}} \rightarrow \mathbb{R}^{T_m * d_{dim}}$. f_1^m, f_2^m are each followed by batch normalisation [29], a ReLU [49], and dropout [58] with dropout rate $drop_{enc}$.

The position encoder block adds learnt modality and temporal positional embeddings to the outputs of the modality-specific embedding blocks. We explain this in detail below. To handle different frame rates in the audio and visual modalities, we use Fourier features [64] $pos_t \in \mathbb{R}^{d_{pos}}$ for the temporal embeddings that encode the actual point in time in the video which corresponds to an audio or visual representation. This allows to capture the relative temporal position of the audio and visual features across the modalities.

For an audio embedding ϕ_{a_t} at time t, a linear map $g_a : \mathbb{R}^{d_{pos}+d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$, and a dropout layer g^D with dropout probability $drop_{prob,pos}$, we obtain position-aware audio feature tokens

$$a_t^p = g^D(g_a(concat(\phi_{a_t}, pos_{at}))) \qquad \text{with} \qquad pos_{at} = pos_a + pos_t, \tag{2}$$

with modality and temporal embeddings $pos_a, pos_t \in \mathbb{R}^{d_{pos}}$ respectively. Position-aware visual tokens v_t^p are obtained analogously.

Furthermore, we prepend a learnt classification token $x^c \in \mathbb{R}^{d_{dim}}$ to the sequence of feature tokens. The corresponding output classification token c_o is used by our output projection O_{proj} to obtain the final prediction.

Audio-visual transformer layers. TCAF contains L stacked audio-visual transformer layers that allow for enhanced cross-attention. Each of our transformer layers consists of an attention function $f_{l,Att}$, followed by a feed forward function $g_{l,FF}$. The output of the l-th transformer layer is given as

$$x_{l,out} = x_{l,ff} + x_{l,att} = g_{l,FF}(x_{l,att}) + x_{l,att}, \tag{3}$$

with

$$x_{l,att} = f_{l,Att}(x_{l,in}) + x_{l,in}, \tag{4}$$

where

$$x_{l,in} = \begin{cases} [x^c, a_1^p, \cdots, a_{T_a}^p, v_1^p, \cdots, v_{T_v}^p] & \text{if } l = 1, \\ x_{l-1,out} & \text{if } 2 \geq l \leq L. \end{cases}$$

We explain the cross-attention used in our transformer layers in the following.

Transformer cross-attention. TCAF primarily exploits cross-modal audio-visual attention to combine the information across the audio and visual modalities. All attention mechanisms in TCAF consist of multi-head attention [67] with H heads and a dimension of d_{head} per head.

We describe the first transformer layer $\mathcal{M}_1$, the transformer layer $\mathcal{M}_l$ operates analogously. We project the position-aware input features x^c, $\{a_t^p\}_{t\in[1,T_a]}$, $\{v_t^p\}_{t\in[1,T_v]}$ to queries, keys, and values with linear maps $g_s : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{head}H}$ for $s \in \{q, k, v\}$. We can then write the outputs of the projection as zero-padded query, key, and value features. We write those out for the queries below, the keys and values are padded in the same way:

$$\mathbf{q}_c = [g_q(x^c), 0, \cdots, 0], \tag{5}$$

$$\mathbf{q}_a = [0, \cdots, 0, g_q(a_1^p), \cdots, g_q(a_{T_a}^p), 0, \cdots, 0], \tag{6}$$

$$\mathbf{q}_v = [0, \cdots, 0, g_q(v_1^p), \cdots, g_q(v_{T_v}^p)]. \tag{7}$$

The full query, key, and value representations, $\mathbf{q}$, $\mathbf{k}$, and $\mathbf{v}$, are the sums of their modality-specific components

$$\mathbf{q} = \mathbf{q}_c + \mathbf{q}_a + \mathbf{q}_v, \qquad \mathbf{k} = \mathbf{k}_c + \mathbf{k}_a + \mathbf{k}_v, \qquad \text{and } \mathbf{v} = \mathbf{v}_c + \mathbf{v}_a + \mathbf{v}_v. \tag{8}$$

The output of the first attention block $x_{1,att}$ is the aggregation of the per-head attention with a linear mapping $g_h : \mathbb{R}^{d_{head}H} \rightarrow \mathbb{R}^{d_{dim}}$, g^{DL} dropout with dropout probability $drop_{prob}$ and layer normalisation g^{LN} [11], such that

$$x_{1,att} = f_{1,Att}(x_{l,in}) = g^{DL}(g_h(f_{1,att}^1(g^{LN}(x_{1,in})), \cdots, f_{1,att}^H(g^{LN}(x_{1,in})))), \tag{9}$$

with the attention f_{att}^h for the attention head h. We can write the attention for the head h as

$$f_{att}^h(x_{1,in}) = softmax\left(\frac{\mathbf{A}}{\sqrt{d_{head}}}\right)\mathbf{v}, \tag{10}$$

where $\mathbf{A}$ can be split into its cross-attention and self-attention components:

$$\mathbf{A}_c = \mathbf{q}_c\,\mathbf{k}^T + \mathbf{k}\,\mathbf{q}_c^T, \qquad \mathbf{A}_x = \mathbf{q}_a\,\mathbf{k}_v^T + \mathbf{q}_v\,\mathbf{k}_a^T, \tag{11}$$
$$\mathbf{A}_{self} = \mathbf{q}_a\,\mathbf{k}_a^T + \mathbf{q}_v\,\mathbf{k}_v^T.$$

We then get

$$\mathbf{A} = \mathbf{A}_c + \mathbf{A}_x + \mathbf{A}_{self} = \begin{pmatrix} A_{cc} & A_{ca} & A_{cv} \\ A_{ac} & \ddots & \vdots \\ A_{vc} & \dots & 0 \end{pmatrix} + \begin{pmatrix} 0 & \dots & 0 \\ \vdots & \ddots & A_{av} \\ 0 & A_{va} & 0 \end{pmatrix} + \begin{pmatrix} 0 & \dots & 0 \\ \vdots & A_{aa} & \vdots \\ 0 & \dots & A_{vv} \end{pmatrix}, \tag{12}$$

where the A_{mn} with $m, n \in \{c, a, v\}$ describe the attention contributions from the classification token, the audio and the visual modalities respectively.

Our TCAF uses the cross-attention $\mathbf{A}_c + \mathbf{A}_x$ to put special emphasis on the attention across modalities. Results for different model variants that use only the within-modality self-attention ($\mathbf{A}_c + \mathbf{A}_{self}$) or the full attention which combines self-attention and cross-attention are presented in Section 4.3.

Feed forward function. The feed forward function $g_{l,FF} : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{dim}}$ is applied to the output of the attention function

$$x_{l,ff} = g_{l,FF}(x_{l,att}) = g^{DL}(g_{l,F2}(g^{DL}(g^{GD}(g_{l,F1}(g^{LN}(x_{l,att})))))) \tag{13}$$

where $g_{l,F1} : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{ff}}$ and $g_{l,F2} : \mathbb{R}^{d_{ff}} \rightarrow \mathbb{R}^{d_{dim}}$ are linear mappings, g^{GD} is a GELU layer [26] and a dropout layer with dropout probability $drop_{prob}$, g^{DL} is dropout with $drop_{prob}$ and g^{LN} is layer normalisation.

Output prediction. To determine the final class prediction, the audio-visual embedding is projected to the same embedding space as the textual class label representations. We project the output classification token c_o of the temporal cross-attention to $\theta_o = O_{proj}(c_o)$ where $\theta_o \in \mathbb{R}^{d_{out}}$. The projection block is composed of a sequence of two linear layers f_3 and f_4, where $f_3 : \mathbb{R}^{d_{dim}} \rightarrow \mathbb{R}^{d_{fhidd}}$ and $f_4 : \mathbb{R}^{d_{fhidd}} \rightarrow \mathbb{R}^{d_{out}}$. f_3, f_4 are each followed by batch normalisation, a ReLU, and dropout with rate $drop_{proj_o}$. We project the word2vec class label embedding w^j for class j using the projection block $W_{proj}(w^j) = \theta_w^j$, where $\theta_w^j \in \mathbb{R}^{d_{out}}$. W_{proj} consists of a linear projection followed by batch normalisation, ReLU, and dropout with dropout rate $drop_{proj_w}$. The class prediction c is obtained by determining the projected word2vec embedding which is closest to the output embedding:

$$c = \underset{j}{\operatorname{argmin}}(\|\theta_w^j - \theta_o\|_2). \tag{14}$$

3.3 Loss functions

Our training objective l combines a cross-entropy loss l_{ce}, a reconstruction loss l_{rec}, and a regression loss l_{reg}:

$$l = l_{ce} + l_{rec} + l_{reg}. \tag{15}$$

Cross-entropy loss. For the ground-truth label y_i with corresponding class index $k_{gt} \in \mathbb{R}^{K_{seen}}$, the output of our temporal cross-attention θ_{o_i}, and a matrix containing the textual label embeddings for the K_{seen} seen classes $\theta_{w_{seen}}$, we define the cross-entropy loss for n training samples as

$$l_{ce} = -\frac{1}{n}\sum_{i}^{n} y_i \log\left(\frac{\exp\left(\theta_{w_{seen},k_{gt}}\theta_{o_i}\right)}{\sum_{k_j}^{K_{seen}}\exp\left(\theta_{w_{seen},k_j}\theta_{o_i}\right)}\right). \tag{16}$$

Regression loss. While the cross-entropy loss updates the probabilities for both the correct and incorrect classes, our regression loss directly focuses on reducing the distance between the output embedding for a sample and the corresponding projected word2vec embedding. The regression loss is based on the mean squared error metric with the following formulation:

$$l_{reg} = \frac{1}{n}\sum_{i=1}^{n}(\theta o_i - \theta_{w_i})^2, \tag{17}$$

where θ_{o_i} is the audio-visual embedding, and θ_{w_i} is the projection of the word2vec embedding corresponding to the i-th sample.

Reconstruction loss. The goal of the reconstruction loss is to ensure that the embeddings θ_o and θ_w contain semantic information from the word2vec embedding w. We use $D_u : \mathbb{R}^{d_{out}} \mapsto \mathbb{R}^{d_{dim}}$ with $\rho_u = D_u(\theta_u)$ for $u \in \{o, w\}$. D_w is a sequence of one linear layer, batch normalisation, a ReLU, and dropout with rate $drop_{proj_w}$. D_o is composed of a sequence of two linear layers each followed by batch normalisation, a ReLU, and dropout with dropout rate $drop_{proj_o}$. Our reconstruction loss encourages the reconstruction of the output embedding, ρ_{o_i}, and the reconstruction of the word2vec projection, ρ_{w_i}, to be close to the original word2vec embedding w_i:

$$l_{rec} = \frac{1}{n}\sum_{i=1}^{n}(\rho_{o_i} - w_i)^2 + \frac{1}{n}\sum_{i=1}^{n}(\rho_{w_i} - w_i)^2. \tag{18}$$

4 Experiments

In this section, we detail our experimental setup (Section 4.1), and compare to state-of-the-art methods for audio-visual GZSL (Section 4.2). Furthermore, we present an ablation study in Section 4.3 which shows the benefits of using our proposed attention scheme and training objective. Finally, we present t-SNE visualisations of our learnt audio-visual embeddings in Section 4.4.

4.1 Experimental setup

Here, we describe the datasets used, the evaluation metrics, and the implementation details for all models.

Datasets. We use the UCF-GZSL^{cls}, VGGSound-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets [47] for audio-visual (G)ZSL for training and testing all models. [47] introduced benchmarks for two sets of features, the first uses a model pre-trained using self-supervision on the VGGSound dataset from [9], the second takes features extracted from pre-trained VGGish [27] and C3D [66] audio and video classification networks. Since the VGGSound dataset is also used for the zero-shot learning task (VGGSound-GZSL), we selected the second option (using VGGish and C3D) and use the corresponding dataset splits proposed in [47].

In particular, the audio features are extracted using VGGish [27] to obtain one 128-dimensional feature vector for each 0.96 s snippet. The visual features are obtained using C3D [66] pre-trained on Sports-1M [32]. For this, all videos are resampled to 25 fps. A 4096-dimensional feature vector is then extracted for 16 consecutive video frames.

Evaluation metrics. We follow [71,47] and use the mean class accuracy to evaluate all models. The ZSL performance is obtained by considering only the subset of test samples from the unseen test classes. For the GZSL performance, the models are evaluated on the full test set which includes seen and unseen classes. We then report the performance on the subsets of seen (S) and unseen (U) classes, and also report their harmonic mean (HM).

Implementation details. For TCAF, we use $d_{in_a} = 128$, $d_{in_v} = 4096$, $d_{fhidd} = 512$, $d_{dim} = 300$ and $d_{out} = 64$. Furthermore, TCAF has $L = 6$ transformer layers layers for UCF-GZSL^{cls} and ActivityNet-GZSL^{cls}, and $L = 8$ for VGGSound-GZSL^{cls}. We set $d_{pos} = 64$, $d_{ff} = 128$. For ActivityNet-GZSL^{cls}/ UCF-GZSL^{cls}/ VGGSound-GZSL^{cls} we use dropout rates $drop_{enc} = 0.1/0.3/0.2$, $drop_{prob,pos} = 0.2/0.2/0.1$, $drop_{prob} = 0.4/0.3/0.5$, $drop_{proj_w} = 0.1/0.1/0.1$, and $drop_{proj_o} = 0.1/0.1/0.2$. All attention blocks use $H = 8$ heads with a dimension of $d_{head} = 64$ per head. We train all models using the Adam optimizer [34] with running average coefficients $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay 0.00001. We use a batch size of 64 for all datasets. In order to efficiently train on ActivityNet-GZSL^{cls}, we randomly trim the features to a maximum sequence length of 60 during training, and we evaluate on features that have a maximum sequence length of 300 and which are centered in the middle of the video. We note, that TCAF can be efficiently trained on a single Nvidia 2080-Ti GPU. All models are trained for 50 epochs. We use a base learning rate of 0.00007 for UCF-GZSL^{cls} and ActivityNet-GZSL^{cls}, and 0.00006 for VGGSound-GZSL^{cls}. For UCF-GZSL^{cls} and ActivityNet-GZSL^{cls} we use a scheduler that reduces the learning rate by a factor of 0.1 when the HM on the validation set has not improved for 3 epochs. To eliminate the bias that the ZSL methods have towards seen classes, we used calibrated stacking [14] on the search space composed of the interval $[0, 3]$ with a step size of 0.2.

We train all models with a two-stage training protocol [47]. In the first stage, we determine the calibrated stacking [14] and the epoch with the best HM performance on the validation set. In the second stage, using the hyperparameters from the first stage, we re-train the models on the union of the training and validation sets. We evaluate the final models on the test set.

4.2 Quantitative results

We compare our proposed TCaF to state-of-the-art audio-visual ZSL frameworks and to audio-visual frameworks that we adapted to the ZSL task.

Audio-visual ZSL baselines. We compare our TCaF to three audio-visual ZSL frameworks. **CJME** [55] consists of a relatively simple architecture which maps both input modalities to a shared embedding space. The modality-specific embeddings in the shared embedding space are input to an attention predictor module that determines the dominant modality which is used for the output prediction. **AVGZSLNet** [46] builds on CJME by adding a shared decoder and introducing additional loss functions to improve the performance. AVGZSLNet removes the attention predictor network and replaces it with a simple average between the output from the head of each modality. **AVCA** [47] is a recent state-of-the-art method for audio-visual G(ZSL). It uses a simple cross-attention mechanism on the temporally averaged audio and visual input features to combine the information from the two modalities. Our proposed TCaF improves upon the closely related AVCA framework by additionally ingesting temporal information in the audio and visual inputs with an enhanced cross-attention mechanism that gathers information across time and modalities.

Audio-visual baselines adapted to ZSL. We adapt two attention-based audio-visual frameworks to the ZSL setting. **Attention Fusion** [20] is a method for audio-visual classification which is trained to classify unimodal information. It then fuses the unimodal predictions with learnt attention weights. The **Perceiver** [30] is a scalable multi-modal transformer framework for flexible learning with arbitrary modality information. It uses a latent bottleneck to encode input information by repeatedly attending to the input with transformer-style attention. The Perceiver allows for a comparison to another transformer-based architecture with focus on multi-modality. We adapt the Perceiver to use the same positional encodings and model capacity as TCaF. We use 64 latent tokens and the same number of layers and dimensions as TCaF. Both Attention Fusion and Perceiver use the same input features, input embedding functions A_{enc} and V_{enc}, learning rate and loss functions as TCaF. For Attention Fusion, we temporally average the input features after A_{enc} and V_{enc} to deal with non-synchronous modality sequences due to different feature extraction rates.

All baselines, except for the Perceiver, operate on temporally averaged audio and visual features. This decreases the amount of information contained in the inputs, in particular regarding the dynamics in a video. In contrast to methods that use temporally averaged inputs, TCaF exploits the temporal dimension which boosts the (G)ZSL performance.

Results. We compare the results obtained with our TCaF to state-of-the-art baselines for audio-visual (G)ZSL and for audio-visual learning in Table 1. TCaF outperforms all previous methods on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets for both, GZSL performance (HM) and ZSL performance. For ActivityNet-GZSL^{cls}, our proposed model is significantly better than its strongest competitor AVCA, with a HM of 12.20% compared to 9.92% and a ZSL performance of 7.96% compared to 7.58%. The CJME and

Model	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
Attention Fusion	14.13	3.00	4.95	3.37	39.34	18.29	24.97	20.21	11.15	3.37	5.18	4.88
Perceiver	13.25	3.03	4.93	3.44	46.85	26.82	34.11	28.12	18.25	4.27	6.92	4.47
CJME	10.86	2.22	3.68	3.72	33.89	24.82	28.65	29.01	10.75	5.55	7.32	6.29
AVGZSLNet	**15.02**	3.19	5.26	4.81	**74.79**	24.15	36.51	31.51	13.70	5.96	8.30	6.39
AVCA	12.63	6.19	8.31	6.91	63.15	30.72	41.34	37.72	16.77	7.04	9.92	7.58
TCAF	12.63	**6.72**	**8.77**	**7.41**	67.14	**40.83**	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table 1. Performance of our TCAF and of state-of-the-art methods for audio-visual (G)ZSL on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets. The mean class accuracy for GZSL is reported on the seen (S) and unseen (U) test classes, and their harmonic mean (HM). For the ZSL performance, only the test subset of unseen classes is considered.

AVGZSLNet frameworks are weaker than the AVCA model. Similar patterns are exhibited for the VGGSound-GZSL^{cls} and UCF-GZSL^{cls} datasets. Interestingly, the GZSL performance for TCAF is improved by a more significant margin than the ZSL performance compared to AVCA across all three datasets. This shows that using temporal information and allowing our model to attend across time and modalities is especially beneficial for the GZSL task.

Furthermore, we observe that the audio-visual Attention Fusion framework and the Perceiver give worse results than AVGZSLNet and AVCA on all three datasets. In particular, our TCAF yields stronger ZSL and GZSL performances than the Perceiver which also takes temporal audio and visual features as inputs, with a HM of 8.77% on VGGSound-GZSL^{cls} for TCAF compared to 4.93% for the Perceiver. Attention Fusion and the Perceiver architecture were not designed for the (G)ZSL setting that uses text as side information. Our proposed training objective, used to also train the Perceiver, aims to regress textual embeddings which might be challenging for the Perceiver given its tight latent bottlenecks.

4.3 Ablation study on the training loss and attention variants

Here, we analyse different components of our proposed TCAF. We first compare the performance of our model when trained using different loss functions. We then investigate the influence of the attention mechanisms used in the model architecture on the (G)ZSL performance. Finally, we show that using multi-modal inputs is beneficial and results in outperforming unimodal baselines.

Comparing different training losses. We show the contributions of the different components in our training loss function to the (G)ZSL performance in Table 2. Using only the regression loss l_{reg} to train our model results in the weakest performance across all datasets, with HM/ZSL performances of 16.25%/30.17% on UCF-GZSL^{cls} compared to 50.78%/44.64% for our full TCAF. Interestingly, the seen performance (S) when using only l_{reg} is relatively weak, likely caused by the calibrated stacking. Similarly, on ActivityNet-GZSL^{cls}, using only l_{reg} yields a low test performance of 0.43% HM. Jointly training with the regression and cross-entropy loss functions ($l_{reg}+l_{ce}$) improves the GZSL and ZSL performance

Loss	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
l_{reg}	0.10	2.41	0.19	2.50	14.30	18.82	16.25	30.17	1.09	0.27	0.43	2.11
$l_{reg} + l_{ce}$	**13.67**	4.06	6.26	4.31	**75.31**	37.15	49.76	41.75	11.36	5.28	7.21	5.31
$l = l_{reg} + l_{ce} + l_{rec}$	12.63	**6.72**	**8.77**	**7.41**	67.14	**40.83**	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table 2. Influence of using different components of our proposed training objective for training TCaF on the (G)ZSL performance on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets.

significantly, giving a ZSL performance of 4.31% compared to 2.50% for l_{reg} on VGGSound-GZSL^{cls}. The best results are obtained when training with our full training objective l which includes a reconstruction loss term, giving the best performance on all three datasets.

Comparing different attention variants. We study the use of different attention patterns in Table 3. In particular, we analyse the effect of using within-modality ($\mathbf{A}_{self}$) and cross-modal ($\mathbf{A}_x$) attention (cf. Eq. (11)), on the GZSL and ZSL performance. Additionally, we investigate models that use a classification token x^c with corresponding output token c_o (*with class. token*) and models for which we simply average the output of the transformer layers which is then used as input to O_{proj} (*w/o class. token*).

Interestingly, we observe that with no global token, using the full attention $\mathbf{A}_{self} + \mathbf{A}_x$ gives better results than using only cross-attention on UCF-GZSL^{cls} and ActivityNet-GZSL^{cls} for ZSL and GZSL, but is slightly worse on VGGSound-GZSL^{cls}. This suggests that the bottleneck introduced by limiting the information flow in the attention when using only cross-attention is beneficial for (G)ZSL on VGGSound-GZSL^{cls}. When not using the classification token and only self-attention $\mathbf{A}_{self}$, representations inside the transformer are created solely within their respective modalities.

Using a classification token (*with class. token*) and the cross-attention variant ($\mathbf{A}_c + \mathbf{A}_x$) yields the strongest ZSL and GZSL results across all three datasets. The most drastic improvements over full attention can be observed on the UCF-GZSL^{cls} dataset, with a HM of 50.78% for the cross-attention with classification token ($\mathbf{A}_c + \mathbf{A}_x$) compared to 39.18% for the full attention ($\mathbf{A}_c + \mathbf{A}_{self} + \mathbf{A}_x$). Furthermore, when using x_c, cross-attention $\mathbf{A}_x$ instead of self-attention $\mathbf{A}_{self}$ leads to a better performance on all three datasets. For $\mathbf{A}_x$ and x_c, we obtain HM scores of 8.77% and 50.78 % on VGGSound-GZSL^{cls} and UCF-GZSL^{cls} compared to 6.71% and 37.37% with $\mathbf{A}_{self}$ and x_c. This shows that using information from both modalities is important for creating strong and transferable video representations for (G)ZSL. Using the global token relaxes the pure cross-attention setting to a certain extent, since $\mathbf{A}_c$ allows for attention between all tokens from both modalities and the global token. The results in Table 3 have demonstrated the clear benefits of our cross-attention variant used in TCaF.

The influence of multi-modality. We compare using only a single input modality for training TCaF to using multiple input modalities in Table 4. For the unimodal baselines TCaF- audio and TCaF- visual, we train TCaF only

Model	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
w/o class. token												
$\mathbf{A}_{self} + \mathbf{A}_x$	**18.40**	3.78	6.27	4.25	31.70	32.57	32.13	33.26	11.87	3.80	5.75	3.90
$\mathbf{A}_{self}$	16.08	3.56	5.83	4.00	42.59	24.04	30.73	27.49	9.51	4.33	5.95	4.39
$\mathbf{A}_x$	14.62	4.22	6.55	4.59	19.52	29.80	23.62	31.35	1.85	3.50	2.42	3.50
with class. token												
$\mathbf{A}_c + \mathbf{A}_{self} + \mathbf{A}_x$	11.36	5.50	7.41	5.97	36.73	**41.99**	39.18	42.56	17.75	6.79	9.83	6.89
$\mathbf{A}_c + \mathbf{A}_{self}$	12.23	4.63	6.71	5.25	40.14	34.95	37.37	35.74	4.24	3.23	3.67	3.25
$\mathbf{A}_c + \mathbf{A}_x$ (TCAF)	12.63	**6.72**	**8.77**	**7.41**	**67.14**	40.83	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table 3. Ablation of different attention variants with and without a classification token on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets.

Model	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S	U	HM	ZSL	S	U	HM	ZSL	S	U	HM	ZSL
TCAF- audio	5.11	4.06	4.53	4.28	35.51	19.75	25.38	24.24	9.28	4.26	5.84	4.65
TCAF- visual	3.97	3.12	3.50	3.19	38.10	26.84	31.49	27.25	2.75	3.11	2.92	3.11
TCAF	**12.63**	**6.72**	**8.77**	**7.41**	**67.14**	**40.83**	**50.78**	**44.64**	**30.12**	**7.65**	**12.20**	**7.96**

Table 4. Influence of using multiple modalities for training and evaluating our proposed model on the (G)ZSL performance on the VGGSound-GZSL^{cls}, UCF-GZSL^{cls}, and ActivityNet-GZSL^{cls} datasets.

with the corresponding input modality. Using only audio inputs gives stronger GZSL and ZSL results than using only visual inputs on VGGSound-GZSL^{cls} and ActivityNet-GZSL^{cls}. We obtain a HM of 5.84% for audio compared to 2.92% for visual inputs on ActivityNet-GZSL^{cls}. Interestingly this pattern is reversed for the UCF-GZSL^{cls} dataset where using visual inputs only results in a slightly higher performance than using the audio inputs with HM scores of 31.49% compared to 25.38%, and ZSL scores of 27.25% and 24.24%. However, using both modalities (TCAF) increases the HM to 50.78% and ZSL to 44.64% on UCF-GZSL^{cls}. Similar trends can be observed for VGGSound-GZSL^{cls} and ActivityNet-GZSL^{cls} which highlights the importance of the tight multi-modal coupling in our TCAF.

4.4 Qualitative results

We present a qualitative analysis of the learnt audio-visual embeddings in Fig. 3. For this, we show t-SNE [45] visualisations for the audio and visual input features and for the learnt multi-modal embeddings from 7 classes in the UCF-GZSL^{cls} test set. We averaged the input features for both modalities across time. We observe that the audio and visual input features are poorly clustered. In contrast, the audio-visual embeddings (θ_o) are clearly clustered for both, seen and unseen classes. This suggests that our network is actually learning useful representations for unseen classes, too. Furthermore, the word2vec class label embeddings (θ_w^j) lie inside the corresponding audio-visual clusters. This confirms that the learnt

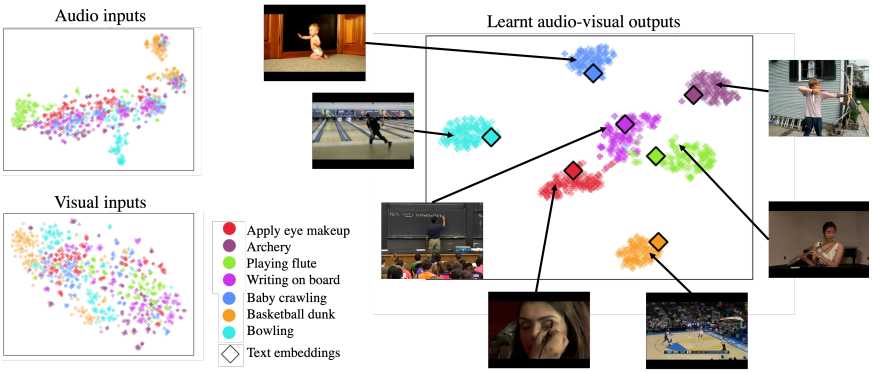


Fig. 3. t-SNE visualisation for five seen (*apply eye makeup, archery, baby crawling, basketball dunk, bowling*) and two unseen (*playing flute, writing on board*) test classes from the UCF-GZSL^{cls} dataset, showing audio and visual input embeddings extracted with C3D and VGGish, and audio-visual output embeddings learned with TCaF. Textual class label embeddings are visualised with a square.

audio-visual embeddings are mapped to locations that are close to the corresponding word2vec embeddings, showing that our embeddings capture semantic information from the word2vec representations.

5 Conclusion

We presented a cross-attention transformer framework that addresses (G)ZSL for video classification using audio-visual input data with temporal information. Our proposed model achieves state-of-the-art performance on the three audio-visual (G)ZSL datasets UCF-GZSL^{cls}, VGGSound-GZSL^{cls}, and ActivityNet-GZSL^{cls}. The use of pre-extracted audio and visual features as inputs results in a computationally efficient framework compared to using raw data. We demonstrated that using cross-modal attention on temporal audio and visual input features and suppressing the contributions from the within-modality self-attention is beneficial for obtaining strong audio-visual embeddings that can transfer information from classes seen during training to novel, unseen classes at test time.

Acknowledgements: This work was supported by BMBF FKZ: 01IS18039A, DFG: SFB 1233 TP 17 - project number 276693517, by the ERC (853489 - DEXIM), and by EXC number 2064/1 – project number 390727645. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting O.-B. Mercea and T. Hummel. The authors would like to thank M. Mancini for valuable feedback.

References

1. Afouras, T., Asano, Y.M., Fagan, F., Vedaldi, A., Metze, F.: Self-supervised object detection from audio-visual correspondence. In: CVPR (2022)
2. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE TPAMI (2018)
3. Afouras, T., Chung, J.S., Zisserman, A.: Asr is all you need: Cross-modal distillation for lip reading. In: ICASSP (2020)
4. Afouras, T., Owens, A., Chung, J.S., Zisserman, A.: Self-supervised learning of audio-visual objects from video. In: ECCV (2020)
5. Akata, Z., Perronnin, F., Harchaoui, Z., Schmid, C.: Label-embedding for image classification. IEEE TPAMI (2015)
6. Akata, Z., Reed, S., Walter, D., Lee, H., Schiele, B.: Evaluation of output embeddings for fine-grained image classification. In: CVPR (2015)
7. Alwassel, H., Mahajan, D., Torresani, L., Ghanem, B., Tran, D.: Self-supervised learning by cross-modal audio-video clustering. In: NeurIPS (2020)
8. Arandjelovic, R., Zisserman, A.: Objects that sound. In: ECCV (2018)
9. Asano, Y., Patrick, M., Rupprecht, C., Vedaldi, A.: Labelling unlabelled videos from scratch with multi-modal self-supervision. NeurIPS (2020)
10. Aytar, Y., Vondrick, C., Torralba, A.: Soundnet: Learning sound representations from unlabeled video. In: NeurIPS (2016)
11. Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
12. Boes, W., Van hamme, H.: Audiovisual transformer architectures for large-scale classification and synchronization of weakly labeled audio events. In: ACM MM (2019)
13. Brattoli, B., Tighe, J., Zhdanov, F., Perona, P., Chalupka, K.: Rethinking zero-shot video classification: End-to-end training for realistic applications. In: CVPR (2020)
14. Chao, W.L., Changpinyo, S., Gong, B., Sha, F.: An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In: ECCV (2016)
15. Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., Zisserman, A.: Localizing visual sounds the hard way. In: CVPR (2021)
16. Chen, Y., Xian, Y., Koepke, A.S., Shan, Y., Akata, Z.: Distilling audio-visual knowledge by compositional contrastive learning. In: CVPR (2021)
17. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: ACL (2019)
18. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
19. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)
20. Fayek, H.M., Kumar, A.: Large scale audiovisual learning of sounds with weakly labeled data. In: IJCAI (2020)
21. Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., Mikolov, T.: Devise: A deep visual-semantic embedding model. In: NeurIPS (2013)
22. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: ECCV (2020)
23. Gan, C., Huang, D., Chen, P., Tenenbaum, J.B., Torralba, A.: Foley music: Learning to generate music from videos. In: ECCV (2020)

24. Gao, R., Grauman, K.: Co-separating sounds of visual objects. In: ICCV (2019)
25. Goldstein, S., Moses, Y.: Guitar music transcription from silent video. In: BMVC (2018)
26. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
27. Hershey, S., Chaudhuri, S., Ellis, D.P., Gemmeke, J.F., Jansen, A., Moore, R.C., Plakal, M., Platt, D., Saurous, R.A., Seybold, B., et al.: Cnn architectures for large-scale audio classification. In: ICASSP (2017)
28. Iashin, V., Rahtu, E.: A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In: BMVC (2020)
29. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
30. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., Carreira, J.: Perceiver: General perception with iterative attention. In: ICML (2021)
31. Jamaludin, A., Chung, J.S., Zisserman, A.: You said that?: Synthesising talking faces from audio. IJCV (2019)
32. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: CVPR (2014)
33. Kerrigan, A., Duarte, K., Rawat, Y., Shah, M.: Reformulating zero-shot action recognition for multi-label actions. NeurIPS (2021)
34. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
35. Koepke, A.S., Wiles, O., Moses, Y., Zisserman, A.: Sight to sound: An end-to-end approach for visual piano transcription. In: ICASSP (2020)
36. Koepke, A.S., Wiles, O., Zisserman, A.: Visual pitch estimation. In: SMC (2019)
37. Korbar, B., Tran, D., Torresani, L.: Cooperative learning of audio and video models from self-supervised synchronization. In: NeurIPS (2018)
38. Li, G., Duan, N., Fang, Y., Gong, M., Jiang, D.: Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In: AAAI (2020)
39. Li, L.H., Yatskar, M., Yin, D., Hsieh, C.J., Chang, K.W.: Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557 (2019)
40. Lin, C.C., Lin, K., Wang, L., Liu, Z., Li, L.: Cross-modal representation learning for zero-shot action recognition. In: CVPR (2022)
41. Lin, Y.B., Wang, Y.C.F.: Audiovisual transformer with instance attention for audio-visual event localization. In: ACCV (2020)
42. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., Wang, Z.: Hit: Hierarchical transformer with momentum contrast for video-text retrieval. In: ICCV (2021)
43. Liu, Y., Guo, J., Cai, D., He, X.: Attribute attention for semantic disambiguation in zero-shot learning. In: CVPR (2019)
44. Lu, J., Batra, D., Parikh, D., Lee, S.: Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In: NeurIPS (2019)
45. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. JMLR (2008)
46. Mazumder, P., Singh, P., Parida, K.K., Namboodiri, V.P.: Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In: WACV (2021)
47. Mercea, O.B., Riesch, L., Koepke, A.S., Akata, Z.: Audio-visual generalised zero-shot learning with cross-modal attention and language. In: CVPR (2022)
48. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: ICLR (2013)

49. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: ICML (2010)
50. Narasimhan, M., Ginosar, S., Owens, A., Efros, A.A., Darrell, T.: Strumming to the beat: Audio-conditioned contrastive video textures. arXiv preprint arXiv:2104.02687 (2021)
51. Narayan, S., Gupta, A., Khan, F.S., Snoek, C.G., Shao, L.: Latent embedding feedback and discriminative features for zero-shot classification. In: ECCV (2020)
52. Owens, A., Efros, A.A.: Audio-visual scene analysis with self-supervised multisensory features. In: ECCV (2018)
53. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Ambient sound provides supervision for visual learning. In: ECCV (2016)
54. Owens, A., Wu, J., McDermott, J.H., Freeman, W.T., Torralba, A.: Learning sight from sound: Ambient sound provides supervision for visual learning. IJCV (2018)
55. Parida, K., Matiyali, N., Guha, T., Sharma, G.: Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In: WACV (2020)
56. Patrick, M., Asano, Y.M., Fong, R., Henriques, J.F., Zweig, G., Vedaldi, A.: Multi-modal self-supervision from generalized data transformations. In: NeurIPS (2020)
57. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog (2019)
58. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. JMLR (2014)
59. Su, K., Liu, X., Shlizerman, E.: Multi-instrumentalist net: Unsupervised generation of music from body movements. arXiv preprint arXiv:2012.03478 (2020)
60. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. arXiv preprint arXiv:1908.08530 (2019)
61. Sun, C., Baradel, F., Murphy, K., Schmid, C.: Learning video representations using contrastive bidirectional transformer. arXiv preprint arXiv:1906.05743 (2019)
62. Sun, C., Myers, A., Vondrick, C., Murphy, K., Schmid, C.: Videobert: A joint model for video and language representation learning. In: ICCV (2019)
63. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. In: EMNLP (2019)
64. Tancik, M., Srinivasan, P., Mildenhall, B., Fridovich-Keil, S., Raghavan, N., Singhal, U., Ramamoorthi, R., Barron, J., Ng, R.: Fourier features let networks learn high frequency functions in low dimensional domains. NeurIPS (2020)
65. Tian, Y., Shi, J., Li, B., Duan, Z., Xu, C.: Audio-visual event localization in unconstrained videos. In: ECCV (2018)
66. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: ICCV (2015)
67. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
68. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The caltech-ucsd birds-200-2011 dataset (2011)
69. Wang, X., Zhu, L., Yang, Y.: T2vlad: global-local sequence alignment for text-video retrieval. In: CVPR (2021)
70. Wiles, O., Koepke, A.S., Zisserman, A.: X2face: A network for controlling face generation using images, audio, and pose codes. In: ECCV (2018)
71. Xian, Y., Lampert, C.H., Schiele, B., Akata, Z.: Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. IEEE TPAMI (2018)
72. Xian, Y., Lorenz, T., Schiele, B., Akata, Z.: Feature generating networks for zero-shot learning. In: CVPR (2018)

73. Xian, Y., Sharma, S., Schiele, B., Akata, Z.: f-vaegan-d2: A feature generating framework for any-shot learning. In: CVPR (2019)
74. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)
75. Xie, G.S., Liu, L., Jin, X., Zhu, F., Zhang, Z., Qin, J., Yao, Y., Shao, L.: Attentive region embedding network for zero-shot learning. In: CVPR (2019)
76. Xu, W., Xian, Y., Wang, J., Schiele, B., Akata, Z.: Attribute prototype network for zero-shot learning. In: NeurIPS (2020)
77. Zhou, H., Liu, Z., Xu, X., Luo, P., Wang, X.: Vision-infused deep audio inpainting. In: ICCV (2019)
78. Zhu, Y., Xie, J., Liu, B., Elgammal, A.: Learning feature-to-feature translator by alternating back-propagation for generative zero-shot learning. In: ICCV (2019)