# Audio-Visual Generalized Zero-Shot Learning using Pre-Trained Large Multi-Modal Models

Supplementary Material

## A. Additional Details about Textual Feature Extraction

### A.1. CLIP Feature Extraction

To boost zero-shot classification performance, [1] calculate normalized CLIP text embeddings for an ensemble of text prompts to retrieve final textual embeddings. Then the mean is taken and the result is normalized again. Normalizing the individual CLIP text representations is necessary in order to obtain a meaningful averaged vector. The second normalization facilitates the calculation of cosine similarity scores. Note that image embeddings are normalized as well.

For UCF-GZSL$^{cls}$ and ActivityNet-GZSL$^{cls}$, we use an ensemble of 48 different prompt templates for each class. UCF-GZSL$^{cls}$ and ActivityNet-GZSL$^{cls}$ have a similar context since both are action recognition datasets. Hence, we use the same text prompts for these two datasets (see listing 1). These templates are taken from the CLIP repository[1].

```
CLIP_prompt_templates = [
    'a photo of a person {}.',
    'a video of a person {}.',
    'a example of a person {}.',
    'a demonstration of a person {}.',
    'a photo of the person {}.',
    'a video of the person {}.',
    'a example of the person {}.',
    'a demonstration of the person {}.',
    'a photo of a person using {}.',
    'a video of a person using {}.',
    'a example of a person using {}.',
    'a demonstration of a person using {}.',
    'a photo of the person using {}.',
    'a video of the person using {}.',
    'a example of the person using {}.',
    'a demonstration of the person using {}.',
    'a photo of a person doing {}.',
    'a video of a person doing {}.',
    'a example of a person doing {}.',
    'a demonstration of a person doing {}.',
    'a photo of the person doing {}.',
    'a video of the person doing {}.',
    'a example of the person doing {}.',
    'a demonstration of the person doing {}.',
    'a photo of a person during {}.',
    'a video of a person during {}.',
    'a example of a person during {}.',
    'a demonstration of a person during {}.',
    'a photo of the person during {}.',
    'a video of the person during {}.',
    'a example of the person during {}.',
    'a demonstration of the person during {}.',
    'a photo of a person performing {}.',
    'a video of a person performing {}.',
    'a example of a person performing {}.',
    'a demonstration of a person performing {}.',
    'a photo of the person performing {}.',
    'a video of the person performing {}.',
    'a example of the person performing {}.',
    'a demonstration of the person performing {}.',
    'a photo of a person practicing {}.',
    'a video of a person practicing {}.',
    'a example of a person practicing {}.',
    'a demonstration of a person practicing {}.',
    'a photo of the person practicing {}.',
    'a video of the person practicing {}.',
    'a example of the person practicing {}.',
    'a demonstration of the person practicing {}.',
]
```

Listing 1. Text prompt templates that were used to create CLIP label embeddings for UCF-GZSL$^{cls}$ and ActivityNet-GZSL$^{cls}$.

VGGSound-GZSL$^{cls}$ contains videos of a variety of categories and hence more general prompts are required. The prompts that we used to create CLIP text embeddings for VGGSound-GZSL$^{cls}$ can be seen in listing 2.

```
VGGSound_CLIP_prompt_templates = [
    'a photo of {}.',
    'a video of {}.',
    'a example of {}.',
    'a demonstration of {}.',
    'a photo of the person {}.',
    'a video of the {}.',
    'a example of the {}.',
    'a demonstration of the {}.'
]
```

Listing 2. Text prompt templates that were used to create CLIP text embeddings for VGGSound-GZSL$^{cls}$.

### A.2. CLAP Feature Extraction

We use the same procedure as in A.1 to extract textual CLAP embeddings. For UCF-GZSL$^{cls}$ and ActivityNet-GZSL$^{cls}$ we use the prompts as in listing 3. For VGGSound-GZSL$^{cls}$, we use the prompts given in listing 4.

```
CLAP_prompt_templates = [
    'a person {} can be heard.',
```

[1] https://github.com/openai/CLIP/blob/main/data/prompts.md#ucf101

```
    'a example of a person {} can be heard.',
    'a demonstration of a person {} can be heard.',
    'the person {} can be heard.',
    'a example of the person {} can be heard.',
    'a demonstration of the person {} can be heard.',
    'a person using {} can be heard.',
    'a example of a person using {} can be heard.',
    'a demonstration of a person using {} can be heard.',
    'a example of the person using {} can be heard.',
    'a demonstration of the person using {} can be heard.',
    'a person doing {} can be heard.',
    'a example of a person doing {} can be heard.',
    'a demonstration of a person doing {} can be heard.',
    'a example of the person doing {} can be heard.',
    'a demonstration of the person doing {} can be heard.',
    'a example of a person during {} can be heard.',
    'a demonstration of a person during {} can be heard.',
    'a example of the person during {} can be heard.',
    'a demonstration of the person during {} can be heard.',
    'a person performing {} can be heard.',
    'a example of a person performing {} can be heard.',
    'a demonstration of a person performing {} can be heard.',
    'a example of the person performing {} can be heard.',
    'a demonstration of the person performing {} can be heard.',
    'a person practicing {} can be heard.',
    'a example of a person practicing {} can be heard.',
    'a demonstration of a person practicing {} can be heard.',
    'a example of the person practicing {} can be heard.',
    'a demonstration of the person practicing {} can be heard.'
]
```

Listing 3. Text prompt templates that were used to create CLAP label embeddings for UCF-GZSL$^{cls}$ and ActivityNet-GZSL$^{cls}$.

```
VGGSound_CLAP_prompt_templates = [
    'a {} can be heard.',
    'a example of a {} can be heard.',
    'a demonstration of a {} can be heard.',
    'the {} can be heard.',
    'a example of the {} can be heard.',
    'a demonstration of the {} can be heard.',
    '{} can be heard.',
    'a example of {} can be heard.',
    'a demonstration of {} can be heard.'
]
```

Listing 4. Text prompt templates that were used to create CLAP text embeddings for VGGSound-GZSL$^{cls}$.

## References


[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 1