VisXP model documentation

Representing video segments in feature space

L^3 leverages the natural association between what we see and what we hear in videos. From training a neural network to distinguish between corresponding and non corresponding pairs of video frames and audio snippets, and subsequently clustering and pseudo-labelling the data, the network has learned to extract meaningful visual and audio features from raw data.

Network Architecture

The network architecture consists of two main components:

1. Visual Subnetwork: This part extracts visual features from video frames using convolutional neural networks (CNNs).
2. Audio Subnetwork: This component extracts audio features from audio snippets using recurrent neural networks (RNNs).

Training Data

VGGSound Data ~200K videos were used for training. This dataset is described in more detail on https://www.robots.ox.ac.uk/~vgg/data/vggsound/ and https://github.com/hche11/VGGSound
The videos were cut into 1-second clips, resulting in around ~2M clips.
Each clip has two modalities,

- visual, represented as the frame in the middle.
- audio, represented as the spectrogram of the 1-second waveform.

Training regime

The model was trained in two stages: pretraining using an audio-visual correspondence task (AVC), and further training with a self-labelling task.

In addition, a more complex version of the model was trained in a different set-up. It performed comparably to the simpler version, so we used the simpler one for inference. However, the complex model yielded some interesting insights and is therefore mentioned here for reference:
https://openaccess.thecvf.com/content/ICCV2023/html/Long_Cross-modal_Scalable_Hierarchical_Clustering_in_Hyperbolic_space_ICCV_2023_paper.html

Pretraining: Self-supervised Audio-Visual Correspondence Task (AVC).

The AVC task serves as the first training objective for L³-Net. Given a video frame and an audio snippet, the network must determine whether they belong to the same underlying video or not. This binary classification task forces the network to learn meaningful visual and audio representations that capture the inherent correspondence between the two modalities.

For this training phase, a **Fusion Network module** was added to the model architecture. This module combines the visual and audio features extracted by the previous subnetworks to determine whether the corresponding pairs are genuine or not.

To train L^3-Net on unlabelled videos, a two-stage sampling process was applied:

1. Positive Sampling: Identify pairs of video frames and audio snippets that correspond to each other, typically taken from the same video.
2. Negative Sampling: Find pairs that do not correspond, ensuring that the audio and video are not from the same video and may even depict unrelated objects or events.

The Paper describing this pre-training phase can be found here: https://openaccess.thecvf.com/content_iccv_2017/html/Arandjelovic_Look_Listen_and_ICCV_2017_paper.html

Training: self-labelling

Self-labelling addresses the limitations of traditional self-supervised learning methods by simultaneously learning representations and pseudo-labels from unlabeled data. This approach leverages the complementary nature of clustering and representation learning to extract meaningful features that are both discriminative and informative.

For this training phase, two components were added to the model architecture:

- **Clustering Network**
  This component utilizes the Visual Subnetwork and Audio Subnetwork fusion to cluster the unlabeled data points based on their learned representations. The network establishes connections between data points based on their feature similarities, and the clustering process partitions the data into distinct groups.
- **Representation Learning Network**
  This subnetwork refines the learned representations by incorporating the clustering information. It takes the representations from the clustering network as input and updates them to better align with the pseudo-labels generated by the clustering process.

Self-labelling optimises a combined objective function that balances clustering and representation learning:

1. Clustering Loss: Encourages the clustering network to partition the data points into meaningful clusters.
2. Representation Loss: Ensures that the learned representations are consistent with the pseudo-labels generated by the clustering network.
3. Regularisation: Prevents overfitting by penalising overly complex or redundant features.

The Paper describing this training phase can be found here: https://openreview.net/forum?id=Hyx-jyBFPr