L^3 leverages the natural association between what we see and what we hear in videos. By training a neural network to distinguish between corresponding and non-corresponding pairs of video frames and audio snippets, the network can learn to extract meaningful visual and audio features from raw data.

Network Architecture

The proposed network architecture, called L³-Net, is composed of three 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).

3. **Fusion Network:** This module combines the visual and audio features extracted by the previous subnetworks to determine whether the corresponding pairs are genuine or not.

Training Data Sampling

To train L³-Net on unlabelled videos, the authors propose a two-stage sampling process:

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.

Self supervision task: Audio-Visual Correspondence Task (AVC)

The AVC task serves as the 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.

[Paper Link]:
https://openaccess.thecvf.com/content_iccv_2017/html/Arandjelovic_Look_Listen_and_ICCV_2017_paper.html

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.

Network Architecture

Self-labelling employs a two-stage network architecture:

1. **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.

2. **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.

Training Objective

Self-labelling optimizes 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. **Regularization Term:** Prevents overfitting by penalizing overly complex or redundant features.

[Paper Link]: https://openreview.net/forum?id=Hyx-jyBFPr

sHHC: Cross-modal Scalable Hyperbolic Hierarchical Clustering

Core Idea

sHHC embedds data points into a hyperbolic space, where the distance between points captures their similarity in a more meaningful way compared to traditional Euclidean spaces. This embedding allows for the construction of continuous hierarchies, where data points can be grouped into clusters with varying levels of granularity.

Network Architecture

sHHC employs a three-stage network architecture:

1. **Audio-Visual Encoders:** These components extract features from audio and visual data using deep neural networks.

2. **Optimal Transport Assignments**: The extracted features are divided into equal-sized subsets based on optimal transport assignments, ensuring that the resulting clusters are balanced and representative of the data distribution.

3. **Hierarchical Embeddings:** The subsets are then embedded hierarchically to hyperbolic space Dn based on similarities between features. This embedding preserves the underlying relationships between data points while adapting to the curved geometry of hyperbolic space.

Training Objective

sHHC optimizes a combined objective function that balances hierarchical embedding and representation learning:

1. **Embedding Loss:** Encourages the network to learn continuous hierarchical embeddings that accurately represent the relationships between data points.

2. **Representation Learning Loss**: Ensures that the learned representations are close to the hierarchical embeddings, capturing the underlying hierarchical characteristics of the data.

Paper Link: https://openaccess.thecvf.com/content/ICCV2023/html/Long_Cross-modal_Scalable_Hierarchical_Clustering_in_Hyperbolic_space_ICCV_2023_paper.html