Deep Co-Clustering for Unsupervised Audiovisual Learning

Di Hu
Northwestern Polytechnical University
hdui831@mail.nwpu.edu.cn

Feiping Nie
Northwestern Polytechnical University
feipingnie@gmail.com

Xuelong Li
Chinese Academy of Sciences
xuelong_li@opt.ac.cn

Abstract

The seen birds twitter, the running cars accompany with noise, people talks by face-to-face, etc. These naturally audiovisual correspondences provide the possibilities to explore and understand the outside world. However, the mixed multiple objects and sounds make it intractable to perform efficient matching in the unconstrained environment. To settle this problem, we propose to adequately excavate audio and visual components and perform elaborate correspondence learning among them. Concretely, a novel unsupervised audiovisual learning model is proposed, named as Deep Co-Clustering *(DCC), that synchronously performs sets of clustering with multimodal vectors of convolutional maps in different shared spaces for capturing multiple audiovisual correspondences. And such integrated multimodal clustering network can be effectively trained with max-margin loss in the end-to-end fashion. Amounts of experiments in feature evaluation and audiovisual tasks are performed. The results demonstrate that DCC can learn effective unimodal representation, with which the classifier can even outperform human. Further, DCC shows noticeable performance in the task of sound localization, multi-source detection, and audiovisual understanding.*

1. Introduction

When we see a dog, why the sound emerged in our mind is mostly barking instead of miaow or others? It seems easy to answer "we can only meet the barking dog in our daily life". Intuitively, as specific visual appearance and acoustic signal usually occur together, we can realize that they are strongly correlated, which accordingly makes us recognize the sound-maker of visual dog and the distinctive barking sound. Hence, the concurrent audiovisual message provides the possibilities to better explore and understand the outside world [19]. For example, the fused audiovisual messages provide stronger experience than single one [37], the out-of-view sound can anticipate a genuine event, and the speech content is better recognized by watching the speaker's lip-movement in the "cocktail party" environment.

The cognitive researchers have noticed such phenomenon in the last century and named it as multisensory processing [19]. They found that some neural cells in superior temporal sulcus (a brain region in the temporal cortex) can simultaneously response to the visual, auditory, and tactile signal [17]. When the concurrent audiovisual message is perceived by the brain, such neural cells could provide the corresponding mechanism to correlate these different messages, which is further reflected in various tasks, such as lip-reading [7] and sensory substitute [30].

In view of the merits of audiovisual learning in human beings, it is highly expected to make the machine possess similar ability, i.e., exploring and perceiving the world via the concurrent audiovisual message. More importantly, in contrast to the expensive human-annotation, the audiovisual correspondence can also provide valuable supervision, and it is pervasive, reliable, and free [27]. As a result, the audiovisual correspondence learning has been given more and more attention recently. In the beginning, the coherent property of audiovisual signal is supposed to provide cross-modal supervision information, where the knowledge of one modality is transferred to supervise the other primitive one. However, the learning capacity is obviously limited by the transferred knowledge and it is difficult to expand the correspondence into unexplored cases. Instead of this, a natural question emerges out: can the model learn the audiovisual perception just by their correspondence without any prior knowledge? The recent works give definitive answers [3, 26]. They propose to train an audiovisual two-stream network by simply appending a correspondence judgement on the top layer. In other words, the model learns to match the sound with the image that contains correct sound source. Surprisingly, the visual and auditory subnets

have learnt to response to specific object and sound after training the model, which can be then applied for unimodal classification, sound localization, etc.

Actually, the correspondence assumption behind the previous works relies on specific audiovisual scenario where the sound-maker should exist in the captured visual appearance and single sound source condition is expected. However, such rigorous scenario is not entirely suitable for the real-life video. First, the unconstrained visual scene contains multiple objects which could be sound-makers or not, and the corresponding soundscape is a kind of multi-source mixture. Simply performing the global corresponding verification without having an insight into the complex scene components could result in the inefficient and inaccurate matching, which therefore need amounts of audiovisual pairs to achieve acceptable performance [3] but may still generate semantic irrelevant matching [35]. Second, the sound-maker does not always produce distinctive sound, such as honking cars, barking dogs, so that the current video clip does not contain any sound but next one does, which therefore creates inconsistent conditions for the correspondence assumption. Moreover, the sound-maker may be even out of the screen so we cannot see it in the video, e.g., the voiceover of photographer. The above intricate audiovisual conditions make it extremely difficult to analyze and understand the realistic environment, especially to correctly match different sound-makers and their produced sounds. Hence, a kind of elaborate correspondence learning is expected.

As mentioned above, each modality involves multiple concrete components in the unconstrained scene, hence it is difficult to correlate the real audiovisual pairs. To settle this problem, we propose to disentangle each modality into a set of distinct components instead of the conventional indiscriminate fashion. Then, we aim to effectively learn the correspondence between these distributed representations of different modalities. More specifically, we argue that the activation vectors across convolution maps have distinct responses for different input components, which just meets the clustering assumption. Hence, we introduce the Kmeans clustering into the two-stream audiovisual network to distinguish concrete objects or sounds captured by video. To align the sound and its corresponding producer, sets of shared spaces for audiovisual pairs are effectively learnt by minimizing the associated triplet loss. As the clustering module is embedded into the multimodal network, the proposed model is named as *Deep Co-Clustering* (DCC).

To comprehensively evaluate the proposed audiovisual model, extensive experiments are conducted. They show that our model trained by wild audiovisual pairs has strong ability in generating effective unimodal features, which further promotes the image/acoustic classification performance, where the results even exceed the human performance level. Moreover, quantitative and qualitative evaluations in the audiovisual tasks, i.e., single sound localization and multisource Sound Event Detection (SED), verify that our model can learn valuable correspondence from the unconstrained videos. And the ultimate audiovisual understanding seems to have preliminary perception ability in real-life scene.

2. Related Works

Source	Supervis.	Task	Reference
Sound	Vision	Acoustic Classif.	[5, 14, 13, 6]
Vision	Sound	Image Classif.	[27, 6]
Sound & Vision	Match	Classification	[3]
		Sound Localization	[4, 35, 26, 42]
		Source Separation	[8, 26, 12, 42]

Table 1. Audiovisual learning settings and relevant tasks.

Audiovisual correspondence is a kind of natural phenomena, which actually comes from the fact that "Sound is produced by the oscillation of object". The simple phenomena provides the possibilities to discover the audiovisual appearances and build their complex correlations. That's why we can match the barking to the dog appearance from numerous sound candidates (sound separation) and find the dog appearance according to the barking sound from the complex visual scene (sound source localization). As usually, the machine model is also expected to possess similarly ability as human.

In the past few years, there have been several works that focus on audiovisual machine learning. The learning settings and relevant tasks can be categorized into three phases according to source and supervision modality, as shown in Table 1. The early works consider that the audio and visual messages of the same entity should have similar class information. Hence, it is expected to utilize the well-trained model of one modality to supervise the other one without additional annotation. Such "teacher-student" learning fashion has been successfully employed for image classification by sound [27] and acoustic recognition by vision [5]. Concretely, Owens *et al.* [27] cluster the ambient sound to provide supervision for visual network. Aytar *et al.* [5] propose to utilize the predicted class information of the visual network to train the corresponding sound net, similarly for [14, 13]. Further, Aytar *et al.* [6] propose to simultaneously train three student models for sound, vision, and text modality from the predicted class probabilities of the ImageNet model.

Although the above models have shown promised cross-modal learning capacity, they actually rely on stronger supervision signal than human. That is, we are not born with

a well-trained brain that have recognized kinds of objects or sounds. Hence, the recent (almost concurrent) works propose to train a two-stream network just by given the audiovisual correspondence without the teacher supervision, as shown in Table 1. Arandjelović and Zisserman [3] train their audiovisual model to judge whether the image and audio clip are corresponding. Although such model is trained without the supervision of any teacher-model, it has learnt highly effective unimodal representation and cross-modal correlation [3]. Hence, it becomes feasible to execute relevant audiovisual tasks, such as sound localization and source separation. For the first task, Arandjelović and Zisserman [4] revise their previous model [3] to find the visual area with the maximum similarity for the current audio clip. Owens et al. [26] propose to adopt the similar model as [3] but use 3D convolution network for the visual pathway instead, which can capture the motion information for sound localization. However, these works rely on the simple global correspondence. When there exist multiple sound-producers in the shown visual modality, it becomes difficult to exactly locate the correct producer. Recently, Senocak *et al.* [35] introduce the attention mechanism into the audio-visual model, where the relevant area of the visual feature maps learn to attend specific input sound. However, there still exists another problem that the real-life acoustic environment is usually a mixture of multiple sounds. To localize the source of specific sound, efficient sound separation is also required.

In the sound separation task, most works propose to reconstruct specific audio streams from the manually mixed tracks with the help of visual embedding. For example, Zhao *et al.*[42] focus on the musical sound separation, while Casanovas *et al.* [8], Owens *et al.* [26], and Ariel *et al.* [10] perform the separation for mixed speech messages. However, the real-life sound is more complex and general than the specific imitated examples, which even lacks the groundtruth for the separated sound sources. Hence, Gao *et al.* [12] propose to decouple the naturally mixed sound into basis vectors via non-negative matrix factorization, then correlate them with predicted labels of pre-trained visual model. In contrast, our proposed method jointly disentangles the audio and visual components, and establishes elaborate correspondence between them, which naturally covers both of the sound separation and localization task.

3. The Proposed Model

The ultimate goal is to align the audio and visual components in the unconstrained natural scene, then perform the challenging audiovisual tasks, i.e., sound-localization, multisource SED, and audiovisual understanding. As the supervision comes from the audiovisual correspondence, the inputs to our model are the associated image-sound pairs. To achieve the above goal, we first provide exhaustive descriptions for both modalities via a two-stream convolutional network. Then, we treat these descriptions as the feature candidates for multimodal co-clustering that learns to recognize and align audiovisual entities.

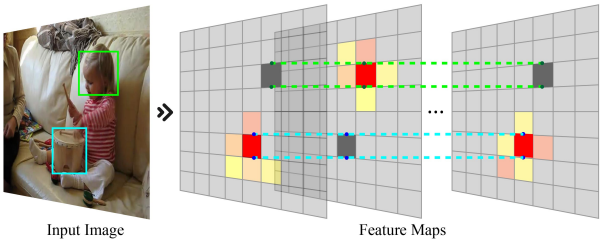

Figure 1. An illustration of activation distributions. It is obvious that different visual components have distinct activation vectors across the feature maps. Such property helps to distinguish different visual components. Best viewed in color.

3.1. Visual and Audio subnet

Visual subnet. The visual pathway directly adopts the off-the-shelf VGG16 architecture but without the fully connected and softmax layers [36]. As the input to the network is resized into 256×256 image, the generated 512 feature maps fall into the size of 8×8. To enable the efficient alignment across modalities, the pixel values are scaled into the range of $[-1, 1]$ that have comparable scale to the log-mel spectrogram of audio signal. As the associated visual components for the audio signal have been encoded into the feature maps, the corresponding entries across all the maps can be viewed as their feature representations, as shown in Fig. 1. In other words, the original feature maps of $8 \times 8 \times 512$ is reshaped into 64×512, where each raw means the representations for specific visual area. Hence, the final visual representations become $\left\{u_1^v, u_2^v, ..., u_p^v | u_i^v \in R^n\right\}$, where $p = 64$ and $n = 512$.

Audio subnet. The audio pathway employs the VGGish model to extract the representations from the input log-mel spectrogram of mono sound, where the VGGish net is a variant of the VGG model [16] and the fully connected layers are also removed. In practice, different from the default configurations in [16], the input audio clip is extended to 496 frames of 10ms each but other parameters about short-time Fourier transform and mel-mapping are kept. Hence, the input to the network becomes 496×64 log-mel spectrogram, and the corresponding output feature maps become $31 \times 4 \times 512$. To prepare the audio representation for the second-stage clustering, we also perform the same operation as the visual ones. That is, the audio feature maps are reshaped into $\left\{u_1^a, u_2^a, ..., u_q^a | u_i^a \in R^n\right\}$, where $q = 124$ and $n = 512$.

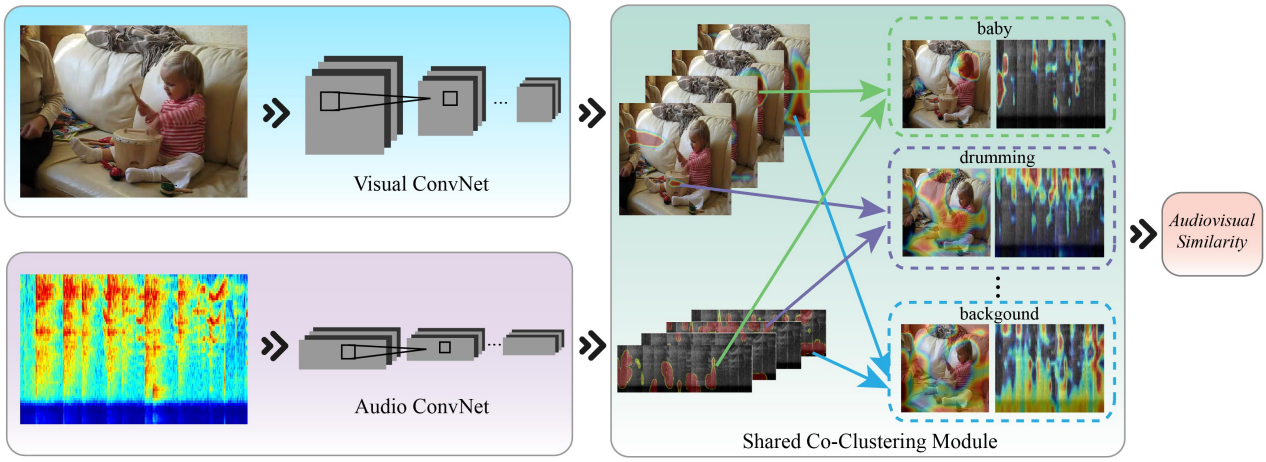

Figure 2. The diagram of the proposed deep co-clustering model. The two modality-specific ConvNets first process the pairwise visual image and audio spectrogram into respective feature maps, then these maps are co-clustered into corresponding components that indicate concrete audiovisual contents, such as baby and its voice, drumming and its sound. Finally, the model takes the similarity across modalities as the supervision for training.

3.2. Multimodal co-clustering module

As the convolutional network shows strong ability in describing the high-level semantics for different modalities [36, 16, 41], we argue that the elements in the feature maps have similar activation probabilities for the same unimodal component, as shown in Fig. 1. It becomes possible to excavate the audiovisual entities by aggregating their similar feature vectors. Hence, we propose to cluster the unimodal feature vectors into object-level representations, and align them in the coordinated audiovisual environment, as shown in Fig. 2. For simplicity, we take $\{u_1, u_2, ..., u_p | u_i \in R^n\}$ for feature representation without regard to the type of modality.

To cluster the unimodal features into k clusters, we propose to perform Kmeans to obtain the centers $C = \{c_1, c_2, ..., c_k | c_j \in R^m\}$, where m is the center dimensionality. Kmeans aims to minimize the within-cluster distance and assign the feature points into k-clusters [20], hence, the objective function can be formulated as,

$$\mathcal{F}(C) = \sum_{i=1}^{p} \min_{j=1}^{k} d\left(u_i, c_j\right), \tag{1}$$

where $\min_{j=1}^{k} d\left(u_i, c_j\right)$ means the distance between current point and its closest center. However, simply introducing Eq.1 into the deep networks will make it difficult to optimize by gradient descent, as the minimization function in Eq.1 is not differentiable.

To solve this intractable problem, the minimization operation is approximated via utilizing the following equation,

$$\max\left\{d_{i1}, d_{i2}, ..., d_{ik}\right\} \approx \frac{1}{z} \log\left(\sum_{j=1}^{k} e^{d_{ij} z}\right), \tag{2}$$

where z is a parameter about magnitude and $d_{ij} = d\left(u_i, c_j\right)$ for simplicity. Eq.2 shows that the maximum value of a given sequence can be approximated by the *log*-summation of corresponding exponential functions. Intuitively, the differences in the original sequence are amplified sharply with the exponential projection, which tends to ignore the tiny ones and remain the largest one. Then, the reversed logarithm projection gives the approximated maximum value. The rigorous proof for Eq.2 can be found in the appendix.

As we aim to find the minimum value of the distance sequence, Eq. 2 is modified into

$$\min\left\{d_{i1}, d_{i2}, ..., d_{ik}\right\} \approx -\frac{1}{z} \log\left(\sum_{j=1}^{k} e^{-d_{ij} z}\right). \tag{3}$$

Then, the objective function of clustering becomes

$$\mathcal{F}\left(C\right) = -\frac{1}{z} \sum_{i=1}^{p} \log\left(\sum_{j=1}^{k} e^{-d_{ij} z}\right). \tag{4}$$

As Eq. 4 is differentiable everywhere, we can directly compute the derivative w.r.t. each cluster center. Concretely, for the center c_j, the derivative is written as,

$$\frac{\partial \mathcal{F}}{\partial c_j} = \sum_{i=1}^{p} \frac{e^{-d_{ij} z}}{\sum_{l=1}^{k} e^{-d_{il} z}} \frac{\partial d_{ij}}{\partial c_j} = \sum_{i=1}^{p} s_{ij} \frac{\partial d_{ij}}{\partial c_j} \tag{5}$$

where $s_{ij} = \frac{e^{-d_{ij}z}}{\sum_{l=1}^{k} e^{-d_{il}z}} = softmax\left(-d_{ij}z\right)$. The softmax coefficient performs like the soft-segmentation over the whole visual area or audio spectrogram for different centers, and we will give more explanation about it in the following sections.

In practice, the distance d_{ij} between each pair of feature point u_i and center c_j can be achieved in different ways, such as Euclidean distance, cosine proximity, etc. In this paper, inspired by the capsule net [34, 39], we choose the inner-product for measuring the agreement, i.e., $d_{ij} = -\left\langle u_i, \frac{c_j}{\|c_j\|} \right\rangle$. By taking it into Eq. 5 and setting the derivative to zero, we can obtain[1]

$$\frac{c_j}{\|c_j\|} = \frac{\sum_{i=1}^{p} s_{ij} u_i}{\left\| \sum_{i=1}^{p} s_{ij} u_i \right\|}, \tag{6}$$

which means the center and the integrated features lie in the same direction. As the coefficients $s_{\cdot j}$ are the softmax values of distances, the corresponding center c_j emerges in the comparable scope as the features and Eq. 6 can be approximatively computed as $c_j = \sum_{i=1}^{n} s_{ij} u_i$ for simplicity. However, there remains another problem that the computation of s_{ij} depends on the current center c_j, which makes it difficult to get the direct update rules for the centers. Instead, we choose to alternatively update the coefficient $s_{ij}^{(r)}$ and center $c_j^{(r+1)}$, i.e.,

$$c_j^{(r+1)} = \sum_{i=1}^{n} s_{ij}^{(r)} u_i. \tag{7}$$

Actually, the updating rule is much similar to the classical Kmeans algorithm, where the first step corresponds to the re-assignment from the features to the centers and the second step is about updating the centers.

The aforementioned clusters indicate a kind of soft assignment (segmentation) over the input image or spectrogram, where each cluster mostly corresponds to certain content (e.g., baby face and drum in image, voice and drumbeat in sound in Fig. 2), hence they can be viewed as the distributed representations of each modality. And we argue that audio and visual messages should have similar distributed representations when they jointly describe the same natural scene. Hence, we propose to perform different center-specific projections $\{W_1, W_2, ..., W_k\}$ over the audio and visual messages to distinguish the representations of different audiovisual entities, then cluster these projected features into the multimodal centers for seeking concrete audiovisual contents. Formally, the distance d_{ij} and center updating become $d_{ij} = -\left\langle W_j u_i, \frac{c_j}{\|c_j\|} \right\rangle$ and $c_j^{(r+1)} = \sum_{i=1}^{n} s_{ij}^{(r)} W_j u_i$, where the projection matrix W_j is shared across modalities and considered as the association with concrete audiovisual entity. Moreover, W_j also performs as the magnitude parameter z when computing the distance d_{ij}. We show the complete co-clustering in Algorithm 1.

We employ the cosine proximity to measure the difference between audiovisual centers, i.e., $s\left(c_i^a, c_i^v\right)$, where c_i^a and c_i^v are the i-center for audio and visual modality, respectively. To efficiently train the two-stream audiovisual network, we employ the max-margin loss to encourage the network to give more confidence to the realistic image-sound pair than mismatched ones,

$$loss = \sum_{i=1, i\neq j}^{k} \max\left(0, s\left(c_j^a, c_i^v\right) - s\left(c_i^a, c_i^v\right) + \Delta\right), \tag{8}$$

where Δ is a margin hyper-parameter and c_j^a means the negative audio sample for the positive audiovisual pair of (c_i^a, c_i^v). In practice, the negative example is randomly sampled from the training set but different from the positive one. The Adam optimizer with the learning rate of 10^{-4} is used. Batch-size of 64 is selected for optimization. And we train the audiovisual net for 25,000 iterations, which took 3 weeks on one K80 GPU card.

Algorithm 1 Audiovisual Co-Clustering Algorithm

Input: The feature vectors for each modality:
$\left\{u_1^a, u_2^a, ..., u_q^a | u_i^a \in R^n\right\}, \left\{u_1^v, u_2^v, ..., u_p^v | u_i^v \in R^n\right\}$
Output: The center vectors for each modality:
$\left\{c_1^a, c_2^a, ..., c_k^a | c_j^a \in R^m\right\}, \left\{c_1^v, c_2^v, ..., c_k^v | c_j^v \in R^m\right\}$
Initialize the distance $d_{ij}^a = d_{ij}^v = 0$
for $t = 1$ to T, x in $\{a, v\}$ **do**
  for $i = 1$ to $q\,(p)$, $j = 1$ to k **do**
    Update weights: $s_{ij}^x = softmax\left(-d_{ij}^x\right)$
    Update centers: $c_j^x = \sum_{i=1}^{n} s_{ij}^x W_j u_i^x$
    Update distances: $d_{ij}^x = -\left\langle W_j u_i^x, \frac{c_j^x}{\|c_j^x\|} \right\rangle$
  end for
end for

3.3. Connection to Capsule

Capsule net is first proposed in [18] then developed in [34]. It aims to represent the various properties of a given entity by the activations of an active capsule, which is similar with our model that describes the audiovisual components via different clusters. Further, we can also view the capsule as another kind of cluster center for the audiovisual description.

[1] Detailed derivation is shown in the appendix.

(a) DCASE2014

Methods	Accuracy
RG [31]	0.69
LTT [24]	0.72
RNH [33]	0.77
Ensamble [38]	0.78
SoundNet [5]	0.88
L^3 [3]	0.91
†L^3 [3]	0.93
DCC	0.94
‡DCC	**0.96**

(b) ESC-50

Methods	Accuracy
Autoencoder [5]	0.399
Rand. Forest [29]	0.443
ConvNet [28]	0.645
SoundNet [5]	0.742
L^3 [3]	0.761
†L^3 [3]	0.793
DCC	0.798
‡DCC	**0.816**
Human Perfor.	0.813

Table 2. Acoustic Scene Classification on DCASE2014 and ESC-50. For fairness, we provide a weakened version of L^3 that is trained with the same audiovisual set as ours, while †L^3 is trained with more data in [3]. ‡DCC takes supervision from the well-trained vision network for training the audio subnet.

However, there exist some key differences between the capsule net and our model. First, our model does not contain the "squashing" function that shrinks the length of capsule vector for representing the possibility. In contrast, the cluster in our model represents a kind of soft assignment over the input features, which is then used for cross-modal comparison. Second, our task is to identify the correspondence between the audio and visual messages in the unconstrained scene instead of the unimodal classification task in [34]. Hence, we employ different training methods in the multimodal cases. And the co-clustering module is also the distinct property that aims to capture the correspondence between modalities.

4. Feature Evaluation

Ideally, the unimodal networks should learn to respond to different objects or sound scenes after training the DCC model. Hence, we propose to evaluate the learned audio and visual representations of the CNN internal layers. For efficiency, the DCC model is trained with 400K unlabeled videos that are randomly sampled from the SoundNet-Flickr dataset [5]. The input audio and visual message are the same as [5], where pairs of 5s sound clip and corresponding image are extracted from each video with no overlap. Note that, the constituted $\sim$1.6M audiovisual pairs are about 17 times less than the ones in L^3 [3] and 5 times less than SoundNet [5].

4.1. Audio Features

The audio representation is evaluated in the typical acoustic classification task. Similar with [5, 3], two benchmark datasets of DCASE2014 [38] and ESC-50 [29] are chosen for evaluation.

DCASE2014. The first task aims to recognize natural scenes of the input sounds. The employed DCASE2014 dataset contains 10 acoustic scenes of 20 samples each and each sample is 30 seconds long. The 20 samples of the same scene are equally partitioned for training and testing. The same experimental setup as [5, 3] is employed, where 60 subclips of 5s long are excerpted for each sample. Mean accuracy of the 10 scenes is measured for evaluation.

ESC-50. The second task focuses on more complex environmental sound. The adopted ESC-50 dataset is a collection of 2000 audio clips of 5s each. They are equally partitioned into 50 categories. Hence, each category contains 40 samples. For fairness, each sample is also partitioned into 1s audio excerpts for data argumentation [5], and these overlapped subclips constitute the audio inputs to the VGGish network. Different from DCASE2014, the mean accuracy is computed over the five leave-one-fold-out evaluations. Note that, the human performance on this dataset is 0.813.

Evaluation. The audio representations of both datasets are extracted by pooling the feature maps[2]. And a multi-class one-vs-all linear SVM is trained with the extracted audio representations. The final accuracy of each clip is the mean value of its subclip scores. To be fair, we also modify the DCC model into the "teacher-student" scheme (‡DCC) where the VGG net is pretrained with ImageNet and kept fixed during training. In Table 2, we show the classification performance on DCASE2014 and ESC-50. It is obvious that the DCC model exceeds all the previous methods, especially the supervised one of SoundNet. Note that, such performance is achieved with less training data (just 400K videos), which confirms that our model can utilize more audiovisual correspondences in the unconstrained videos to effectively train the unimodal network. And the cross-modal supervision version ‡DCC improves the accuracy further, where the most noticeable point is that ‡DCC outperforms human [29] (81.6% vs 81.3%). Such result even goes beyond our expectation, as the audio net of VGGish is just trained with the corresponding visual centers instead of human-annotations. Yet, it exactly verifies that the elaborative alignment between audio and visual contents efficiently works and the audiovisual correspondence indeed helps to learn the unimodal perception.

4.2. Visual Features

The visual representation is evaluated in the object recognition task. The chosen PASCAL VOC 2007 dataset contains 20 object categories that are collected in realistic scenes [11]. We perform global-pooling over the conv5_1 features of VGG16 net to obtain the visual features. A multi-class one-vs-all linear SVM is also employed as the classifier, and the results are evaluated using *Mean Average*

[2]Similar with SoundNet [5], we evaluate the performance of different VGGish layers and select conv4_1 as the extraction layer.

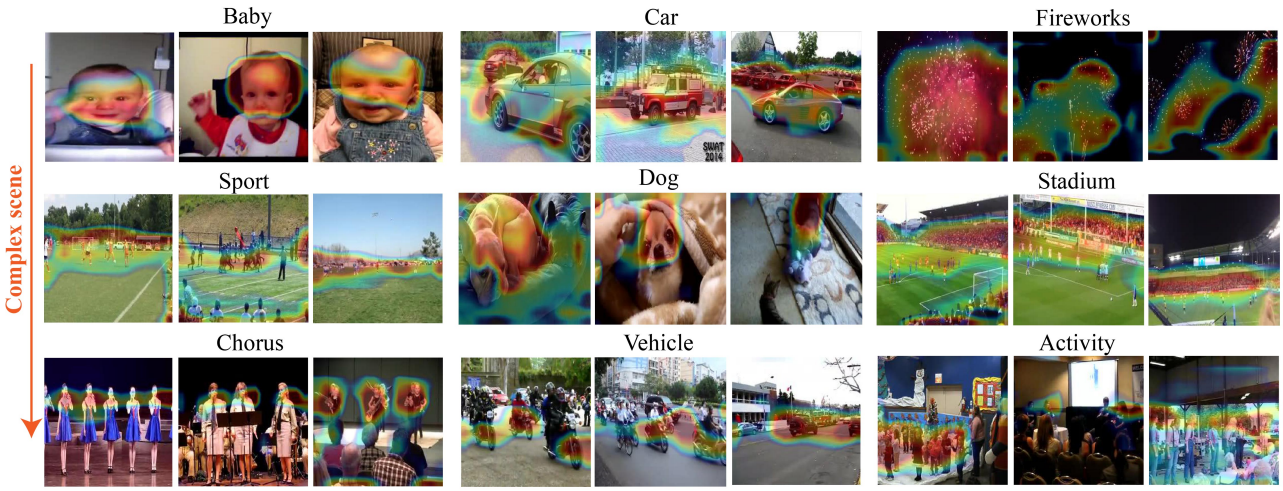

Figure 3. Qualitative examples of sound source localization. After feeding the audio and visual messages into the DCC model, we visualize the soft assignment that belongs to the most related visual cluster to the audio messages. Note that, the visual scene becomes more complex from top to bottom, and the label is just for visualization purpose.

Precision (mAP). As the DCC model does not contain the standard FC layer as previous works, the best conv/pooling features of other methods are chosen for comparison, which have been reported in [27]. To validate the effectiveness of co-clustering in DCC, we choose to compare with the visual model in [27], which treats the separated sound clusters as object indicators for visual supervision. In contrast, DCC model jointly learns the audio and visual representation rather than the above single-flow from sound to vision, hence it is more flexible for learning the audiovisual correspondence. As shown in Table 3, our model indeed shows noticeable improvement over the simple cluster supervision, even its multi-label variation (in binary) [27]. Moreover, we also compare with the pretrained VGG16 net on ImageNet. But, what surprises us is that the DCC model is comparable the human performance in acoustic classification but it has a large gap with the image classification benchmark. Such differences may come from the complexity of visual scene compared with the acoustic ones. Nevertheless, our model still provides meaningful insights in learning effective visual representation via audiovisual correspondence.

Methods	mAP
Kmeans [22]	0.348
Tracking [40]	0.422
Patch Pos. [9]	0.467
Egomotion [2]	0.311
Sound (cluster) [27]	0.458
Sound (binary) [27]	0.467
DCC	**0.514**
ImageNet	0.672

Table 3. Image Classification on Pascal VOC 2007[11]. The shown results are the best ones reported in [27] except the ones with FC features.

5. Audiovisual Evaluation

In this section, we propose to evaluate the DCC model in the practical audiovisual tasks. The difficulty of these tasks is gradually increased according to the complexity of audiovisual scenes, where the simple single sound condition and complex real-life environment are both considered.

5.1. Single Sound Localization

In this task, we aim to localize the sound source in the visual scene as [4, 26], where the simple case of single source is considered. As only one sound appears in the audio track, the generated audio features should share identical center. In practice, we perform average-pooling over the audio centers into c^a, then compare it with all the visual centers via cosine proximity, where the number of visual centers is set to 2 (i.e., sound-make and other components) in this simple case. And the visual center with the highest score is considered as the indicator of corresponding sound source. In order to visualize the sound source further, we resort to the soft assignment of the selected visual center c^v_j, i.e., $s^v_{\cdot j}$. As the assignment $s^v_{ij} \in [0, 1]$, the coefficient vector $s^v_{\cdot j}$ is reshaped back to the size of original feature map and viewed as the heatmap that indicates the cluster property.

In Fig. 3, we show the qualitative examples with respect to the sound-source locations of different videos from the SoundNet-Flickr dataset. It is obvious that the DCC model has learnt to distinguish different visual appearances and correlate the sound with corresponding visual source, al-

Methods	cIoU(0.5)	cIoU(0.7)	AUC
Random	12.0	-	32.3
Unsupervised [35]	66.0	18.8	55.8
Supervised [35]	80.4	25.5	60.3
Sup.+Unsup. [35]	**82.8**	28.8	62.0
DCC	69.3	**62.9**	**63.8**

Table 4. . The evaluation of sound source localization. The cIoUs with threshold 0.5 and 0.7 are shown. The area under the cIoU curve by varying the threshold from 1 to 0 (AUC) is also provided.

though the training phase is entirely performed in the unsupervised fashion. Concretely, in the simple scene, the visual source of baby voice, car noise, and fireworks explosion is easy to localize. When the visual scene becomes more complex, the dcc model can also successfully localize the corresponding source. The dog appearance highly responses to the barking sound while the cat does not. In contrast to the audience and background, only the onstage choruses response to the singing sound. And the moving vehicles are successfully localized regardless of the driver or other visual contents in the complex traffic environment.

Apart from the qualitative analysis, we also provide the quantitative evaluation. We directly adopt the annotated sound-sources dataset [35], which are originally collected from the SoundNet-Flickr dataset. This sub-dataset contains 2,786 audio-image pairs, where the sound-maker of each pair is individually located by three subjects. 250 pairs are randomly sampled to construct the testing set. By setting an arbitrary threshold over the assignment $s^v_{\cdot j}$, we can obtain a binary segmentation over the visual objects that probably indicate the sound locations. Hence, to compare the automatic segmentation with human annotations, we employ the *consensus Intersection over Union* (cIoU) and corresponding AUC area in [35] as the evaluation metric[3]. As shown in Table. 4, the proposed DCC model is compared with the recent sound location model with attention mechanism [35]. It is obvious that DCC shows superior performance over the attention model, especially the supervised ones. Moreover, when the threshold becomes larger (i.e., 0.7), DCC can still enjoy reliable localization while the other models fail. Such phenomenon indicates that the clustering mechanism in DCC can effectively distinguish different modality components and exactly correlate them among different modalities.

[3]In practice, the cIoU metric is the ratio between the overlap area and the annotated bounding boxes. More details can be found in [35].

5.2. Real-Life Sound Event Detection

In this section, in contrast to specific sound separation task[4], we focus on a more general and complicated sound task, i.e., multisource SED. In the realistic environment, multiple sound tracks usually exist at the same time, i.e., the street environment maybe a mixture of people speaking, car noise, walking sound, brakes squeaking, etc. It is expected to detect the existing sounds in every moment, which is much more challenging than the previous single acoustic recognition [15]. In the DCASE2017 acoustic challenges, the third task[5] is exactly the multisource SED. The audio dataset used in this task focuses on the complex street acoustic scenes that consist of different traffic levels and activities. The whole dataset is divided for development and evaluation, and each audio is 3-5 minutes long. Segment-based F-score and error rate are calculated as the evaluation metric.

As our model provides elaborative visual supervision for training the audio subnet, the corresponding representation should provide sufficient description for the multitrack sound. To validate the assumption, we directly replace the input spectrum with our generated audio representation in the baseline model of MLP [15]. As shown in Table. 5, the DCC model is compared with the top five methods in the challenge, the audiovisual net L^3, and the VGGish net. It is obvious that our model takes the first place on F1 metric and is comparable to the best model in error rate. Specifically, there are three points we should pay attention to. First, by utilizing the audio representation of DCC model instead of raw spectrum, we can have a noticeable improvement. Such improvement indicates that the correspondence learning across modalities indeed provides effective supervision in distinguishing different audio contents. Second, as the L^3 net simply performs global matching between audio and visual scene without exploring the concrete content inside, it fails to provide effective audio representation for multisource SED. Third, although the VGGish net is trained on a preliminary version of YouTube-8M (with labels) that is much larger than our training data, our model still outperforms it. This comes from the more efficient audiovisual correspondence learning of DCC model.

5.3. Audiovisual Understanding

As introduced in the Section 1, the real-life audiovisual environment is unconstrained, where each modality consists of multiple instances or components, such as speaking, brakes squeaking, walking sound in the audio modality and building, people, cars, road in the visual modality of

[4]The sound separation task mostly focuses specific task scenarios and need effective supervision from the original sources[8, 26, 42], which goes beyond our conditions.

[5] http://www.cs.tut.fi/sgn/arg/dcase2017/challenge/task-sound-event-detection-in-real-life-audio

Figure 4. Qualitative examples of complex audiovisual understanding. We first feed the messages into the DCC model, then the most representative audiovisual clusters are selected for shown, where the corresponding assignments are binarized into the masks over each modality via a threshold of 0.7. The labels in the figure indicate the learned audiovisual content, which is not used in the training procedure.

Methods	Segment F1	Segment Error
J-NEAT-E [23]	44.9	0.90
SLFFN [23]	43.8	1.01
ASH [1]	41.7	**0.79**
MICNN [21]	40.8	0.81
MLP [15]	42.8	0.94
§MLP [15]	39.1	0.90
L^3 [3]	43.24	0.89
VGGish [16]	50.96	0.86
DCC	**52.14**	0.83

Table 5. Real life sound event detection on the evaluation dataset of DCASE 2017 Challenge. We choose the default STFT parameters of 25ms window size and 10 window hop [16]. The same parameters are also adopted by §MLP, L^3, and VGGish, while other methods adopt the default parameters in [15].

the street environment. Hence, it is difficult to disentangle them within each modality and establish exact correlations between modalities. In this section, we attempt to employ the DCC model to perform the audiovisual understanding in such cases where only the qualitative evaluation is provided due to the absent annotations. To illustrate the results better, we turn the soft assignment of clustering into a binary map via a threshold of 0.7. And Fig. 4 shows the matched audio and visual clustering results of different real-life videos, where the sound is represented in spectrogram. In the "baby drums" video, the drumming sound and corresponding motion are captured and correlated, meanwhile the baby face and people voice are also picked out from the intricate audiovisual content. These two distinct centers jointly describe the audiovisual structures. In more complex indoor and outdoor environments, the DCC model can also capture the people yelling and talking sound from background music and loud environment noise by clustering the audio feature vectors, and correlate them with the corresponding sound-makers (i.e., the visual centers) via the shared projection matrix. However, there still exist some failure cases. Concretely, the out-of-view sound-maker is inaccessible for current visual clustering. Hence, the DCC model improperly correlates the background music with kitchenware in the second video. Similarly, the talking sound comes from the visible woman and out-of-view photographer in the third video, but our model simply extracts all the human voice and assigns them to the visual center of woman. Such failure cases also remind us that the real-life audiovisual understanding is far more difficult than what we have imagined.

Moreover, to perceive the audio centers more naturally, we reconstruct the audio signal from the masked spectrogram information and show them in the released video demo[6].

6. Discussion

In this paper, we aim to explore the elaborate correspondence between audio and visual messages in unconstrained environment by resorting to the deep co-clustering method. In contrast to the previous rough correspondence, our model can efficiently learn more effective audio and visual features, with which we can even exceed the human performance. Further, such elaborate learning contributes to noticeable improvements in the complicated audiovisual tasks, such as sound localization, multisource SED, and audiovisual understanding.

Although the proposed DCC shows considerable superiority over other methods in these tasks, there still remains one problem that the number of clusters k is pre-fixed instead of automatically determined. When there is single sound, it is easy to set $k = 2$ for foreground and background. But when multiple sound-makers emerge, it becomes difficult to pre-determine the value of k. Although we can obtain distinct clusters after setting $k = 10$ in the audiovisual understanding task, more reliable method for determining the number of audiovisual components is still expected [32], which will be focused in the future work. Besides, a kind of elaborate annotation for realistic audiovisual video is also expected, where both the spectrogram of different sounds and the corresponding visual source should be annotated and aligned for quantitative evaluation.

7. Acknowledgement

We gratefully thank Jianlin Su for the constructive discussion in this work.

References

[1] S. Adavanne and T. Virtanen. A report on sound event detection with different binaural features. Technical report, DCASE2017 Challenge, September 2017.

[2] P. Agrawal, J. Carreira, and J. Malik. Learning to see by moving. In *Computer Vision (ICCV), 2015 IEEE International Conference on*, pages 37–45. IEEE, 2015.

[3] R. Arandjelovic and A. Zisserman. Look, listen and learn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 609–617. IEEE, 2017.

[4] R. Arandjelović and A. Zisserman. Objects that sound. *arXiv preprint arXiv:1712.06651*, 2017.

[5] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[6]https://www.youtube.com/watch?v=O4aR76CWnH0

[6] Y. Aytar, C. Vondrick, and A. Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[7] G. A. Calvert, E. T. Bullmore, M. J. Brammer, R. Campbell, S. C. Williams, P. K. McGuire, P. W. Woodruff, S. D. Iversen, and A. S. David. Activation of auditory cortex during silent lipreading. *science*, 276(5312):593–596, 1997.

[8] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval. Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 12(5):358–371, 2010.

[9] C. Doersch, A. Gupta, and A. A. Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1422–1430, 2015.

[10] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[11] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[12] R. Gao, R. Feris, and K. Grauman. Learning to separate object sounds by watching unlabeled video. *arXiv preprint arXiv:1804.01665*, 2018.

[13] D. Harwath and J. R. Glass. Learning word-like units from joint audio-visual analysis. *arXiv preprint arXiv:1701.07481*, 2017.

[14] D. Harwath, A. Torralba, and J. Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2016.

[15] T. Heittola and A. Mesaros. DCASE 2017 challenge setup: Tasks, datasets and baseline system. Technical report, DCASE2017 Challenge, September 2017.

[16] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

[17] K. Hikosaka, E. Iwai, H. Saito, and K. Tanaka. Polysensory properties of neurons in the anterior bank of the caudal superior temporal sulcus of the macaque monkey. *Journal of neurophysiology*, 60(5):1615–1637, 1988.

[18] G. E. Hinton, A. Krizhevsky, and S. D. Wang. Transforming auto-encoders. In *International Conference on Artificial Neural Networks*, pages 44–51. Springer, 2011.

[19] N. P. Holmes and C. Spence. Multisensory integration: space, time and superadditivity. *Current Biology*, 15(18):R762–R764, 2005.

[20] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666, 2010.

[21] I.-Y. Jeong, S. Lee, Y. Han, and K. Lee. Audio event detection using multiple-input convolutional neural network. Technical report, DCASE2017 Challenge, September 2017.

[22] P. Krähenbühl, C. Doersch, J. Donahue, and T. Darrell. Data-dependent initializations of convolutional neural networks. *arXiv preprint arXiv:1511.06856*, 2015.

[23] C. Kroos and M. D. Plumbley. Neuroevolution for sound event detection in real life audio: A pilot study. Technical report, DCASE2017 Challenge, September 2017.

[24] D. Li, J. Tam, and D. Toub. Auditory scene classification using machine learning techniques. *IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events*, 2013.

[25] X. Li, D. Hu, and F. Nie. Deep binary reconstruction for cross-modal hashing. In *Proceedings of the 2017 ACM on Multimedia Conference*, pages 1398–1406. ACM, 2017.

[26] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. *arXiv preprint arXiv:1804.03641*, 2018.

[27] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, pages 801–816. Springer, 2016.

[28] K. J. Piczak. Environmental sound classification with convolutional neural networks. In *Machine Learning for Signal Processing (MLSP), 2015 IEEE 25th International Workshop on*, pages 1–6. IEEE, 2015.

[29] K. J. Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pages 1015–1018. ACM, 2015.

[30] M. J. Proulx, D. J. Brown, A. Pasqualotto, and P. Meijer. Multisensory perceptual learning and sensory substitution. *Neuroscience & Biobehavioral Reviews*, 41:16–25, 2014.

[31] A. Rakotomamonjy and G. Gasso. Histogram of gradients of time-frequency representations for audio scene classification. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(1):142–153, 2015.

[32] S. Ray and R. H. Turi. Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques*, pages 137–143. Calcutta, India, 1999.

[33] G. Roma, W. Nogueira, and P. Herrera. Recurrence quantification analysis features for environmental sound recognition. In *Applications of Signal Processing to Audio and Acoustics (WASPAA), 2013 IEEE Workshop on*, pages 1–4. IEEE, 2013.

[34] S. Sabour, N. Frosst, and G. E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.

[35] A. Senocak, T.-H. Oh, J. Kim, M.-H. Yang, and I. S. Kweon. Learning to localize sound source in visual scenes. *arXiv preprint arXiv:1803.03849*, 2018.

[36] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] B. E. Stein and M. A. Meredith. *The merging of the senses.* The MIT Press, 1993.

[38] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley. Detection and classification of acoustic scenes and events. *IEEE Transactions on Multimedia*, 17(10):1733–1746, 2015.

[39] D. Wang and Q. Liu. An optimization view on dynamic routing between capsules. 2018.

[40] X. Wang and A. Gupta. Unsupervised learning of visual representations using videos. *arXiv preprint arXiv:1505.00687*, 2015.

[41] X. Zhang, J. Zhao, and Y. LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[42] H. Zhao, C. Gan, A. Rouditchenko, C. Vondrick, J. McDermott, and A. Torralba. The sound of pixels. *arXiv preprint arXiv:1804.03160*, 2018.

8. Appendix

8.1. Approximated Maximization Function

Although the maximization function is not differentiable, it can be approximated via the following equation,

$$\max\{d_{i1}, d_{i2}, ..., d_{ik}\} \approx \lim_{z\to+\infty} \frac{1}{z} \log\left(\sum_{j=1}^{k} e^{d_{ij}z}\right), \quad (9)$$

where z is a hype-parameter that controls the precision of approximation. Instead of the above multi-variable case, we first consider the maximization function of two variables $\{x_1, x_2\}$. Actually, it is well known that

$$\begin{aligned} \max\{x_1, x_2\} &= \tfrac{1}{2}\left(|x_1 + x_2| + |x_1 - x_2|\right), \\ s.t. \quad & x_1 \geq 0, x_2 \geq 0, \end{aligned} \quad (10)$$

Hence the approximation for maximization is turned for the absolute value function $f(x) = |x|$. As the derivative function of $f(x)$ is $f'(x) = \begin{cases} +1, & x \geq 0 \\ -1, & x < 0 \end{cases}$, it can be directly replaced by the adaptive tanh function [25], i.e., $f'(x) = \lim_{z\to+\infty} \frac{e^{zx} - e^{-zx}}{e^{zx} + e^{-zx}}$. Then, we can obtain the approximated absolute value function via integral

$$f(x) = \lim_{z\to+\infty} \frac{1}{z} \log\left(e^{zx} + e^{-zx}\right). \quad (11)$$

Hence, the maximization function over two variables can be written as

$$\begin{aligned} &\max\{x_1, x_2\} \\ &= \lim_{z\to+\infty} \tfrac{1}{2z} \log\left(e^{2zx_1} + e^{2zx_2} + e^{-2zx_1} + e^{-2zx_2}\right). \end{aligned} \quad (12)$$

As $z \to +\infty$ and $x_1 \geq 0, x_2 \geq 0$, Eq. 12 can be approximated into

$$\max\{x_1, x_2\} \approx \lim_{z\to+\infty} \frac{1}{z} \log\left(e^{zx_1} + e^{zx_2}\right). \quad (13)$$

At this point, the maximization function has become differentiable for two variables. And it can also be extended to three more variables. Concretely, for three variables $\{x_1, x_2, x_3\}$, let $c = \max\{x_1, x_2\}$, then

$$\begin{aligned} \max\{x_1, x_2, x_3\} &= \max\{c, x_3\} \\ &\approx \lim_{z\to+\infty} \tfrac{1}{z} \log\left(e^{\log(e^{zx_1} + e^{zx_2})} + e^{zx_3}\right) \\ &= \lim_{z\to+\infty} \tfrac{1}{z} \log\left(e^{zx_1} + e^{zx_2} + e^{zx_3}\right). \end{aligned} \quad (14)$$

Hence, for multivariable, we can have

$$\max\{x_1, x_2, ..., x_n\} \approx \lim_{z\to+\infty} \frac{1}{z} \log\left(\sum_{i=1}^{n} e^{zx_i}\right). \quad (15)$$

8.2. Derivation of Eq. 6

To substitute $d_{ij} = -\left\langle u_i, \frac{c_j}{\|c_j\|} \right\rangle$ into $\sum_{i=1}^{n} s_{ij} \frac{\partial d_{ij}}{\partial c_j} = 0$, we first give the derivative of d_{ij} w.r.t. c_j,

$$\frac{\partial d_{ij}}{\partial c_j} = -\frac{\partial\left(\frac{u_i^T c_j}{\|c_j\|}\right)}{\partial c_j} = -\frac{u_i}{\|c_j\|} + u_i^T c_j \cdot \frac{c_j}{\|c_j\|^3}. \quad (16)$$

Then, by taking Eq. 16 into $\sum_{i=1}^{n} s_{ij} \frac{\partial d_{ij}}{\partial c_j} = 0$, we can have

$$\sum_{i=1}^{n} s_{ij} \frac{u_i^T c_j}{\|c_j\|} \cdot \frac{c_j}{\|c_j\|} = \sum_{i=1}^{n} s_{ij} u_i. \quad (17)$$

By taking the modulus of expression in both sides of Eq. 17, we can have

$$\left\|\sum_{i=1}^{n} s_{ij} \frac{u_i^T c_j}{\|c_j\|}\right\| \cdot \left\|\frac{c_j}{\|c_j\|}\right\| = \left\|\sum_{i=1}^{n} s_{ij} u_i\right\|. \quad (18)$$

As $\left\|\frac{c_j}{\|c_j\|}\right\| = 1$, Eq. 18 becomes

$$\left\|\sum_{i=1}^{n} s_{ij} u_i\right\| = \left\|\frac{\sum_{i=1}^{n} s_{ij} u_i^T \cdot c_j}{\|c_j\|}\right\| = \left\|\sum_{i=1}^{n} s_{ij} u_i\right\| |\cos\theta| \quad (19)$$

As $d_{ij} = -\left\langle u_i, \frac{c_j}{\|c_j\|} \right\rangle$, we expect to maximize the cosine proximity between these two vectors, i.e., $\theta = 0$. Hence, $\sum_{i=1}^{n} s_{ij} u_i$ and c_j should lie in the same direction, i.e.,

$$\frac{c_j}{\|c_j\|} = \frac{\sum_{i=1}^{n} s_{ij} u_i}{\left\|\sum_{i=1}^{n} s_{ij} u_i\right\|}. \quad (20)$$