Deep Multimodal Clustering for Unsupervised Audiovisual Learning

Di Hu, Feiping Nie, Xuelong Li*

School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an 710072, P. R. China

`hdui831@mail.nwpu.edu.cn, feipingnie@gmail.com, xuelong_li@ieee.org`

Abstract

The seen birds twitter, the running cars accompany with noise, 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 Multimodal Clustering *(DMC), 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 DMC can learn effective unimodal representation, with which the classifier can even outperform human performance. Further, DMC shows noticeable performance in sound localization, multisource detection, and audiovisual understanding.*

1. Introduction

When seeing 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 catch the barking dog in our daily life". 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].

The cognitive community has noticed such phenomenon

---

*Corresponding author

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 visual, auditory, and tactile signal [18]. When the concurrent audiovisual message is perceived by the brain, such neural cells could provide corresponding mechanism to correlate these different messages, which is further reflected in various tasks, such as lip-reading [8] and sensory substitute [30].

In view of the merits of audiovisual learning in human beings, it is highly expected to make 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.

The correspondence assumption behind previous works [27, 3, 6] rely 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 multisource 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 [33]. 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. So, a kind of elaborate correspondence learning is expected.

As each modality involves multiple concrete components in the unconstrained scene, 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 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 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 Multimodal Clustering* (DMC). Extensive experiments conducted on wild audiovisual pairs show superiority of our model on unimodal features generation, image/acoustic classification and some audiovisual tasks, such as single sound localization and multisource Sound Event Detection (SED). And the ultimate audiovisual understanding seems to have preliminary perception ability in real-life scene.

2. Related Works

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 audiovisual appearances and build their complex correlations. That's why we can match the barking sound to the dog appear-

Source	Supervis.	Task	Reference
Sound	Vision	Acoustic Classif.	[5, 15, 14, 6]
Vision	Sound	Image Classif.	[27, 6]
Sound & Vision	Match	Classification	[3]
		Sound Localization	[4, 33, 26, 38]
		Source Separation	[9, 26, 13, 38]

Table 1. Audiovisual learning settings and relevant tasks.

ance from numerous audio 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].

Although the above models have shown promised crossmodal 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, recent (almost concurrent) works propose to train a two-stream network just by given the audiovisual correspondence, 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 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.* [33] introduce the attention mechanism into the audiovisual model, where the relevant area of visual feature maps learn to attend specific input sound. However, there still exists another problem
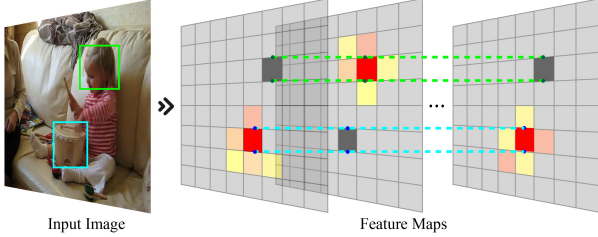
Figure 1. An illustration of activation distributions. It is obvious that different visual components have distinct activation vectors across the feature maps. Such property helps to distinguish different visual components. Best viewed in color.

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 manually mixed tracks with the help of visual embedding. For example, Zhao *et al.*[38] focus on the musical sound separation, while Casanovas *et al.* [9], Owens *et al.* [26], and Ariel *et al.* [11] 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, 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

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 [34]. 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 row means the representations for specific visual area. Hence, the final visual representations become $\{u_1^v, u_2^v, ..., u_p^v | u_i^v \in R^n\}$, 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 [17]. In practice, different from the default configurations in [17], 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 $\{u_1^a, u_2^a, ..., u_q^a | u_i^a \in R^n\}$, where $q = 124$ and $n = 512$.

3.2. Multimodal clustering module

As the convolutional network shows strong ability in describing the high-level semantics for different modalities [34, 17, 37], 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,

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

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 a hard assignment of data points for clusters and not differentiable. To solve this intractable problem, one way is to make a soft assignment for each point. Particularly, *Expectation-Maximization* (EM) algorithm for *Gaussian Mixture Models* (GMMs) makes a soft assignment based on the posterior probabilities and converges to a local optimum [25].

In this paper, we propose another perspective to transform the hard assignment in Eq.(1) to a soft assignment problem and be a differentiable one. The minimization operation in Eq.(1) is approximated via utilizing the following equation,

$$\max \{d_{i1}, d_{i2}, ..., d_{ik}\} \approx \frac{1}{z} \log \left(\sum_{j=1}^{k} e^{d_{ij} z}\right), \qquad (2)$$
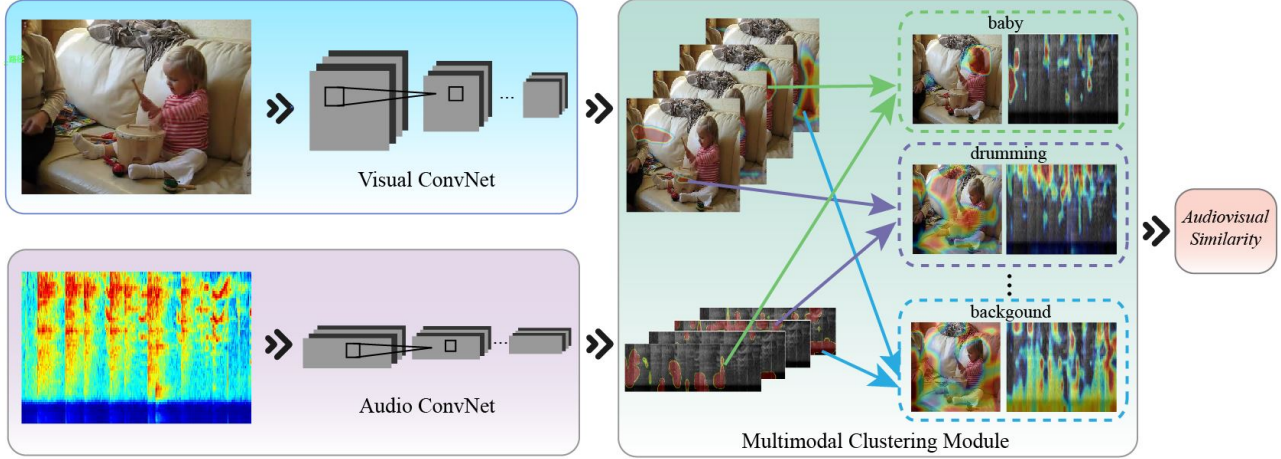
Figure 2. The diagram of the proposed deep multimodal clustering model. The two modality-specific ConvNets first process the pairwise visual image and audio spectrogram into respective feature maps, then these maps are co-clustered into corresponding components that indicate concrete audiovisual contents, such as baby and its voice, drumming and its sound. Finally, the model takes the similarity across modalities as the supervision for training.

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 materials.

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

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

Then, the objective function of clustering becomes

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

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

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

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[1] [32, 35], 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[2]

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

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 is approximatively computed as $c_j = \sum_{i=1}^{p} s_{ij} u_i$ 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.,

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

Actually, the updating rule is much similar to the EM algorithm that maximizes posterior probabilities in GMMs [7].

---

[1]A discussion about capsule and DMC is provided in the materials.
[2]Detailed derivation is shown in the materials.

Specifically, the first step is the *expectation* step or E step, which uses the current parameters to evaluate the posterior probabilities, i.e., re-assigns data points to the centers. The second step is the *maximization* step or M step that aims to re-estimate the means, covariance, and mixing coefficients, i.e., update the centers in Eq.(7).

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 $c_j^{(r+1)} = \sum_{i=1}^{p} s_{ij}^{(r)} W_j u_i$, 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 multimodal clustering in Algorithm 1.

We employ the cosine proximity to measure the difference between audiovisual centers, i.e., $s(c_i^a, c_i^v)$, 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,

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

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.

4. Feature Evaluation

Ideally, the unimodal networks should have learnt to respond to different objects or sound scenes after training the DMC model. Hence, we propose to evaluate the learned audio and visual representations of the CNN internal layers.

Algorithm 1 Deep Multimodal Clustering Algorithm

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

For efficiency, the DMC 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 ~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 complex environmental sound classification task. The adopted ESC-50 dataset [29] 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. The mean accuracy is computed over the five leave-one-fold-out evaluations. Note that, the human performance on this dataset is 0.813.

The audio representations are extracted by pooling the feature maps[3]. 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 DMC model into the "teacher-student" scheme (‡DMC) where the VGG net is pretrained with ImageNet and kept fixed during training. In Table 2 (a), it is obvious that the DMC model exceeds all the previous methods except *Audio-Visual Temporal Synchronization* (AVTS) [22]. 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. We also note that AVTS is trained with the whole

---

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

(a) 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
†AVTS [22]	0.823
DMC	0.798
‡DMC	**0.826**
Human Perfor.	0.813

(b) Pascal VOC 2007	
Methods	Accuracy
Taxton. [27]	0.375
Kmeans [23]	0.348
Tracking [36]	0.422
Patch. [10]	0.467
Egomotion [2]	0.311
Sound(spe.) [27]	0.440
Sound(clu.) [27]	0.458
Sound(bia.) [27]	0.467
DMC	**0.514**
ImageNet	0.672

Table 2. Acoustic Scene Classification on ESC-50 [29] and Image Classification on Pascal VOC 2007[12]. (a) 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]. †AVTS is trained with the whole SoundNet-Flickr dataset [5]. ‡DMC takes supervision from the well-trained vision network for training the audio subnet. (b) The shown results are the best ones reported in [27] except the ones with FC features.

2M+ videos in [5], which is 5 times more than DMC. Even so, DMCs still outperform AVTS on the DCASE2014 benchmark dataset (More details can be found in the materials). And the cross-modal supervision version ‡DMC improves the accuracy further, where the most noticeable point is that ‡DMC outperforms human [29] (82.6% vs 81.3%). Hence, it verifies that the elaborative alignment 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 [12]. 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 Precision* (mAP). As the DMC 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 multimodal clustering in DMC, we choose to compare with the visual model in [27], which treats the separated sound clusters as object indicators for visual supervision. In contrast, DMC 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 2 (b), our model indeed shows noticeable improvement over the simple cluster
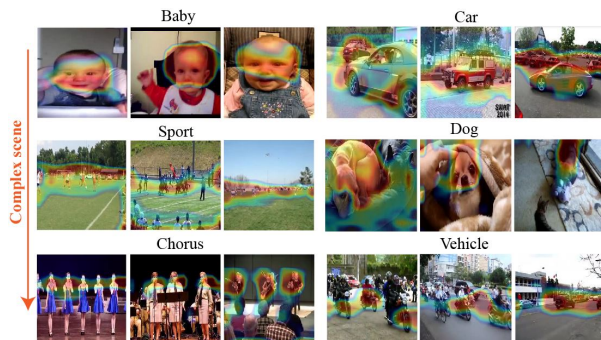

Figure 3. Qualitative examples of sound source localization. After feeding the audio and visual messages into the DMC model, we visualize the soft assignment that belongs to the most related visual cluster to the audio messages. Note that, the visual scene becomes more complex from top to bottom, and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

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 DMC 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.

5. Audiovisual Evaluation

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-maker 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_j^v, i.e., $s_{\cdot j}^v$. As the assignment $s_{ij}^v \in [0, 1]$, the coefficient vector $s_{\cdot j}^v$ 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 DMC model has learnt to distinguish different visual appearances and correlate the sound with corresponding visual source, although the training phase is entirely performed in the un-

supervised fashion. Concretely, in the simple scene, the visual source of baby voice and car noise are easy to localize. When the visual scene becomes more complex, the DMC 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 [33], 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 (with single sound). By setting an arbitrary threshold over the assignment $s_{\cdot j}^v$, 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 [33] as the evaluation metric. As shown in Table. 3, the proposed DMC model is compared with the recent sound location model with attention mechanism [33]. First, it is obvious that DMC shows superior performance over the unsupervised attention model. Particularly, when the cIoU threshold becomes larger (i.e., 0.7), DMC even outperforms the supervised ones. Second, apart from the most related visual centers to the sound track, the unrelated one is also evaluated. The large decline of the unrelated center indicates that the clustering mechanism in DMC can effectively distinguish different modality components and exactly correlate them among different modalities.

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 may be 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 [16]. Hence, it becomes more valuable to evaluate the ability of DMC in learning the effective representation of multi-track sound. In the DCASE2017 acoustic challenges, the third task[5] is exactly the multisource SED.

---

[4]The sound separation task mostly focuses specific task scenarios and need effective supervision from the original sources[9, 26, 38], which goes beyond our conditions.

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

Methods	cIoU(0.5)	cIoU(0.7)	AUC
Random	12.0	-	32.3
Unsupervised† [33]	52.4	-	51.2
Unsupervised [33]	66.0	∼18.8	55.8
Supervised [33]	80.4	∼25.5	60.3
Sup.+Unsup. [33]	82.8	∼28.8	62.0
DMC (unrelated)	10.4	5.2	21.1
DMC (related)	67.1	26.2	56.8

Table 3. 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. The unsupervised† method in [33] employs a modified attention mechanism.

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 audio representation should provide sufficient description for the multi-track sound. To validate the assumption, we directly replace the input spectrum with our generated audio representation in the baseline model of MLP [16]. As shown in Table. 4, the DMC model is compared with the top five methods in the challenge, the audiovisual net L^3 [3], and the VGGish net [17]. 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 DMC 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 DMC 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

Methods	Segment F1	Segment Error
J-NEAT-E [24]	44.9	0.90
SLFFN [24]	43.8	1.01
ASH [1]	41.7	**0.79**
MICNN [21]	40.8	0.81
MLP [16]	42.8	0.94
§MLP [16]	39.1	0.90
L^3 [3]	43.24	0.89
VGGish [17]	50.96	0.86
DMC	**52.14**	0.83

Table 4. 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 [17]. The same parameters are also adopted by §MLP, L^3, and VGGish, while other methods adopt the default parameters in [16].
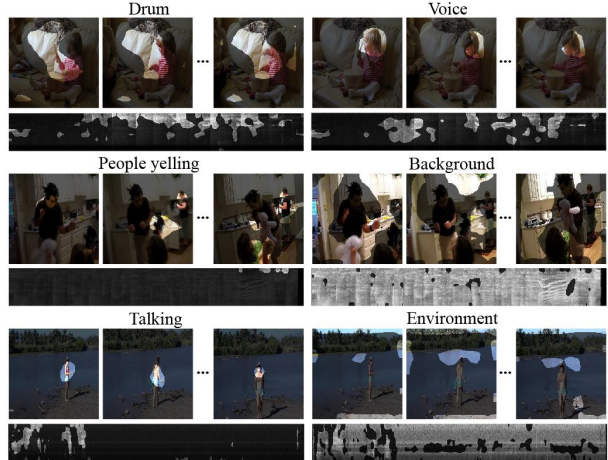

Figure 4. Qualitative examples of complex audiovisual understanding. We first feed the audiovisual messages into the DMC model, then the corresponding audiovisual clusters are captured and shown, where the assignments are binarized into the masks over each modality via a threshold of 0.7. The labels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

the street environment. Hence, it is difficult to disentangle them within each modality and establish exact correlations between modalities, i.e., audiovisual understanding. In this section, we attempt to employ the DMC 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 DMC 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. In contrast, the DMC 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. Discussion

In this paper, we aim to explore the elaborate correspondence between audio and visual messages in unconstrained environment by resorting to the proposed deep multimodal clustering method. In contrast to the previous rough correspondence, our model can efficiently learn more effective audio and visual features, which 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 DMC 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 [31], which will be focused in the future work.

7. Acknowledgement

References

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

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

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

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

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

[6] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[7] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[8] Gemma A Calvert, Edward T Bullmore, Michael J Brammer, Ruth Campbell, Steven CR Williams, Philip K McGuire, Peter WR Woodruff, Susan D Iversen, and Anthony S David. Activation of auditory cortex during silent lipreading. *science*, 276(5312):593–596, 1997.

[9] Anna Llagostera Casanovas, Gianluca Monaci, Pierre Vandergheynst, and Rémi Gribonval. Blind audiovisual source separation based on sparse redundant representations. *IEEE Transactions on Multimedia*, 12(5):358–371, 2010.

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

[11] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[12] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010.

[13] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. *arXiv preprint arXiv:1804.01665*, 2018.

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

[15] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In *Advances in Neural Information Processing Systems*, pages 1858–1866, 2016.

[16] Toni Heittola and Annamaria Mesaros. DCASE 2017 challenge setup: Tasks, datasets and baseline system. Technical report, DCASE2017 Challenge, September 2017.

[17] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan 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.

[18] KAZUO Hikosaka, EIICHI Iwai, H Saito, and KEIJI 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.

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

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

[21] Il-Young Jeong, Subin Lee, Yoonchang Han, and Kyogu Lee. Audio event detection using multiple-input convolutional neural network. Technical report, DCASE2017 Challenge, September 2017.

[22] Bruno Korbar, Du Tran, and Lorenzo Torresani. Co-training of audio and video representations from self-supervised temporal synchronization. *arXiv preprint arXiv:1807.00230*, 2018.

[23] Philipp Krähenbühl, Carl Doersch, Jeff Donahue, and Trevor Darrell. Data-dependent initializations of convolutional neural networks. *arXiv preprint arXiv:1511.06856*, 2015.

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

[25] Brian Kulis and Michael I Jordan. Revisiting k-means: New algorithms via bayesian nonparametrics. *arXiv preprint arXiv:1111.0352*, 2011.

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

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

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

[30] Michael J Proulx, David J Brown, Achille Pasqualotto, and Peter Meijer. Multisensory perceptual learning and sensory substitution. *Neuroscience & Biobehavioral Reviews*, 41:16–25, 2014.

[31] Siddheswar Ray and Rose 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.

[32] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.

[33] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. *arXiv preprint arXiv:1803.03849*, 2018.

[34] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[35] Dilin Wang and Qiang Liu. An optimization view on dynamic routing between capsules. 2018.

[36] Xiaolong Wang and Abhinav Gupta. Unsupervised learning of visual representations using videos. *arXiv preprint arXiv:1505.00687*, 2015.

[37] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657, 2015.

[38] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. *arXiv preprint arXiv:1804.03160*, 2018.