Supplemental Materials for DMC

1. Approximated Maximization Function

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

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

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

$$\begin{array}{c}\max\left\{x_1,x_2\right\}=\frac{1}{2}\left(\left|x_1+x_2\right|+\left|x_1-x_2\right|\right),\\ s.t.\quad x_1\geq 0,x_2\geq 0,\end{array} \quad (2)$$

Hence the approximation for maximization is turned for the absolute value function $f\left(x\right)=\left|x\right|$. As the derivative function of $f\left(x\right)$ is $f'\left(x\right)=\left\{\begin{array}{l}+1,\ x\geq 0\\ -1,\ x<0\end{array}\right.$, it can be directly replaced by the adaptive tanh function [6], i.e., $f'\left(x\right)=\lim\limits_{z\to+\infty}\frac{e^{zx}-e^{-zx}}{e^{zx}+e^{-zx}}$. Then, we can obtain the approximated absolute value function via integral

$$f\left(x\right)=\lim_{z\to+\infty}\frac{1}{z}\log\left(e^{zx}+e^{-zx}\right). \quad (3)$$

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

$$\begin{array}{l}\max\left\{x_1,x_2\right\}\\ =\lim\limits_{z\to+\infty}\frac{1}{2z}\log\left(e^{2zx_1}+e^{2zx_2}+e^{-2zx_1}+e^{-2zx_2}\right).\end{array} \quad (4)$$

As $z\to+\infty$ and $x_1\geq 0,x_2\geq 0$, Eq. 4 can be approximated into

$$\max\left\{x_1,x_2\right\}\approx\lim_{z\to+\infty}\frac{1}{z}\log\left(e^{zx_1}+e^{zx_2}\right). \quad (5)$$

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 $c=\max\left\{x_1,x_2\right\}$, then

$$\begin{aligned}\max\left\{x_1,x_2,x_3\right\}&=\max\left\{c,x_3\right\}\\ &\approx\lim_{z\to+\infty}\tfrac{1}{z}\log\left(e^{\log\left(e^{zx_1}+e^{zx_2}\right)}+e^{zx_3}\right)\\ &=\lim_{z\to+\infty}\tfrac{1}{z}\log\left(e^{zx_1}+e^{zx_2}+e^{zx_3}\right).\end{aligned} \quad (6)$$

Hence, for multivariable, we can have

$$\max\left\{x_1,x_2,...,x_n\right\}\approx\lim_{z\to+\infty}\frac{1}{z}\log\left(\sum_{i=1}^{p}e^{zx_i}\right). \quad (7)$$

2. Derivation of Eq. 6

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

$$\frac{\partial d_{ij}}{\partial c_j}=-\frac{\partial\left(\frac{u_i^Tc_j}{\left\|c_j\right\|}\right)}{\partial c_j}=-\frac{u_i}{\left\|c_j\right\|}+u_i^Tc_j\cdot\frac{c_j}{\left\|c_j\right\|^3}. \quad (8)$$

Then, by taking Eq. 8 into $\sum\limits_{i=1}^{p}s_{ij}\frac{\partial d_{ij}}{\partial c_j}=0$, we can have

$$\sum_{i=1}^{p}s_{ij}\frac{u_i^Tc_j}{\left\|c_j\right\|}\cdot\frac{c_j}{\left\|c_j\right\|}=\sum_{i=1}^{p}s_{ij}u_i. \quad (9)$$

By taking the modulus of expression in both sides of Eq. 9, we can have

$$\left\|\sum_{i=1}^{p}s_{ij}\frac{u_i^Tc_j}{\left\|c_j\right\|}\right\|\cdot\left\|\frac{c_j}{\left\|c_j\right\|}\right\|=\left\|\sum_{i=1}^{p}s_{ij}u_i\right\|. \quad (10)$$

As $\left\|\frac{c_j}{\left\|c_j\right\|}\right\|$=1, Eq. 10 becomes

$$\left\|\sum_{i=1}^{p}s_{ij}u_i\right\|=\left\|\frac{\sum\limits_{i=1}^{p}s_{ij}u_i^T\cdot c_j}{\left\|c_j\right\|}\right\|=\left\|\sum_{i=1}^{p}s_{ij}u_i\right\|\left|\cos\theta\right| \quad (11)$$

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

$$\frac{c_j}{\left\|c_j\right\|}=\frac{\sum\limits_{i=1}^{p}s_{ij}u_i}{\left\|\sum\limits_{i=1}^{p}s_{ij}u_i\right\|}. \quad (12)$$

3. Connection to Capsule

Capsule net is first proposed in [3] then developed in [9]. 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.

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 [9]. Hence, we employ different training methods in the multimodal cases. And the multimodal clustering module is also the distinct property that aims to capture the correspondence between modalities.

4. Audio Feature Evaluation

For the evaluation of audio features, we provide more experimental results on a different acoustic classification dataset, i.e., DCASE2014 [10]. DCASE2014 focuses on natural scenes sounds. It 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 [2, 1] is employed, where 60 subclips of 5s long are excerpted for each sample. Mean accuracy of the 10 scenes is measured for evaluation. The same extraction and classification strategy as the ESC-50 dataset is employed. And Table 1 shows the results in mAP. Similar with the results on ESC-50, DMC and ‡ DMC enjoy the top-two score among these comparison methods, which confirms the effectiveness of modality clustering and elaborate correspondence learning.

Methods	mAP
RG [7]	0.69
LTT [5]	0.72
RNH [8]	0.77
Ensamble [10]	0.78
SoundNet [2]	0.88
L^3 [1]	0.91
†L^3 [1]	0.93
†AVTS [4]	0.94
DMC	0.94
‡DMC	**0.96**

Table 1. Acoustic Scene Classification on DCASE2014 [10]. 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 [1], similarly for †AVTS. ‡DMC takes supervision from the well-trained vision network for training the audio subnet.

References

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

[2] 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.

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

[4] B. Korbar, D. Tran, and L. Torresani. Co-training of audio and video representations from self-supervised temporal synchronization. *arXiv preprint arXiv:1807.00230*, 2018.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.