Supplementary Materials: Hyperbolic Audio-visual Zero-shot Learning

Jie Hong[1,2], Zeeshan Hayder[2], Junlin Han[1,2], Pengfei Fang[3]*,
Mehrtash Harandi[4], Lars Petersson[2]
[1]Australian National University, [2]Data61-CSIRO, [3]Southeast University, [4]Monash University
jie.hong@anu.edu.au, zeeshan.hayder@data61.csiro.au, junlinhcv@gmail.com,
fangpengfei@seu.edu.cn, mehrtash.harandi@monash.edu, lars.petersson@data61.csiro.au

1. Methodology

Prior to the Hyperbolic Alignment Module, we introduce two toy models, *Hyper-embedding* and *Hyper-net*, as shown in Figure 1 (a) and (b), in addition to our Hyperbolic Alignment Module.

1.1. Hyper-embedding.

Hyper-embedding, illustrated in Figure 1 (a), projects all features onto a hyperbolic tangent space before computing the loss $\mathcal{L}_{avca}$. The hyperbolic tangent space $T_z\mathbb{H}_c^n$ is a Euclidean space, so the loss $\mathcal{L}_{avca}$ can be directly applied.

1.2. Hyper-net.

Hyper-net replaces all Euclidean MLP with Hyperbolic MLP, enabling the entire network to be learned under the Poincaré ball space $\mathbb{H}_c^n$ (See Figure 1 (b)). We investigate the potential of applying hyperbolic geometry in audio-visual zero-shot learning by designing and testing two toy models.

2. Experiment

2.1. Ablation Study

Align similarity matrix *vs*. features. We align the similarity matrices between audio and visual features. This can be understood as preserving the latent structure between the two modalities, as captured by the respective similarities. Consequently, it is not necessary for one modality to precisely replicate the representation space of the other modality, as this would impose a strong constraint on the feature space. We conduct an ablation study on UCF-GZSL^{cls} where we align the features directly. The results are shown in Tab. 1.

2.2. Result Analysis

Results. The experimental results of Hyper-embedding and Hyper-net are presented in Table 2 and 3, respectively.

	Similarity alignment		Feature alignment	
Method	HM ↑	ZSL ↑	HM ↑	ZSL ↑
Hyper-alignment	42.52	39.80	29.11	29.74
Hyper-single	44.99	39.86	31.01	30.19
Hyper-multiple	48.30	52.11	37.19	39.23

Table 1. Ablation study: align similarity matrix *vs*. features. Different alignments are tested on dataset UCF-GZSL^{cls}.

AVCA [2] serves as the baseline. The two toy models that incorporate hyperbolic geometry outperform the baseline in certain cases. For example, on VGGSound-GZSL^{cls}, Hyper-embedding achieves $8.74\%/7.19\%$ in HM/ZSL, which is higher than the $8.31\%/6.91\%$ of AVCA. On UCF-GZSL, Hyper-net surpasses AVCA in ZSL by 1.84%. These results suggest hyperbolic geometry may be a promising approach in audio-visual zero-shot learning. In addition to the visualized examples presented in the paper, more visualizations are provided in Figure 2.

References

[1] Pratik Mazumder, Pravendra Singh, Kranti Kumar Parida, and Vinay P Namboodiri. Avgzslnet: Audio-visual generalized zero-shot learning by reconstructing label features from multi-modal embeddings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3090–3099, 2021. 2

[2] Otniel-Bogdan Mercea, Lukas Riesch, A Koepke, and Zeynep Akata. Audio-visual generalised zero-shot learning with cross-modal attention and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10553–10563, 2022. 1, 2

[3] Kranti Parida, Neeraj Matiyali, Tanaya Guha, and Gaurav Sharma. Coordinated joint multimodal embeddings for generalized audio-visual zero-shot classification and retrieval of videos. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3251–3260, 2020. 2

*Corresponding author

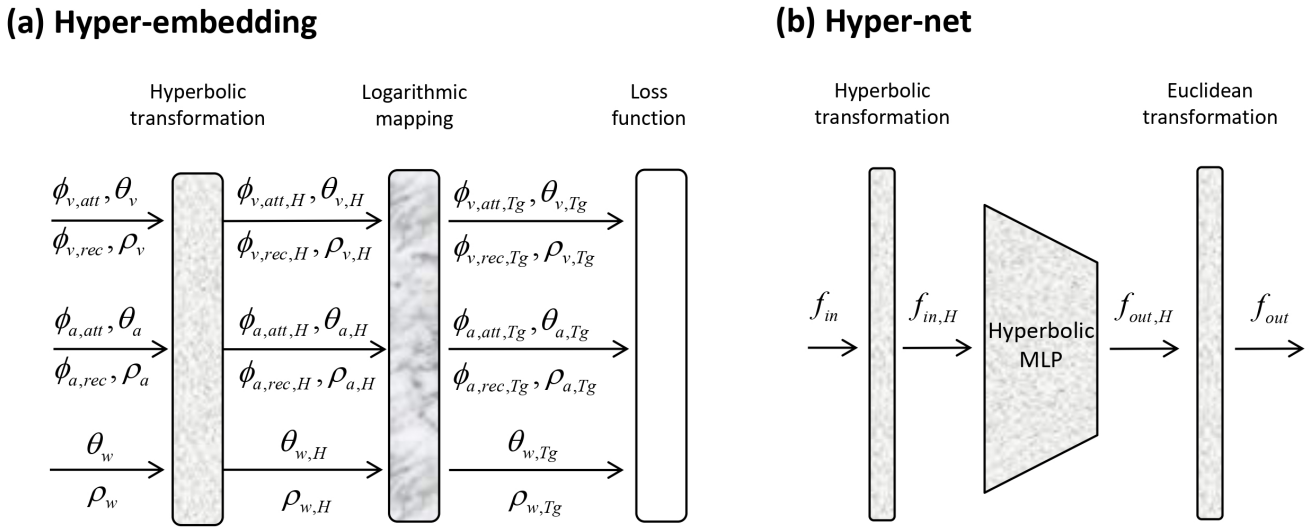


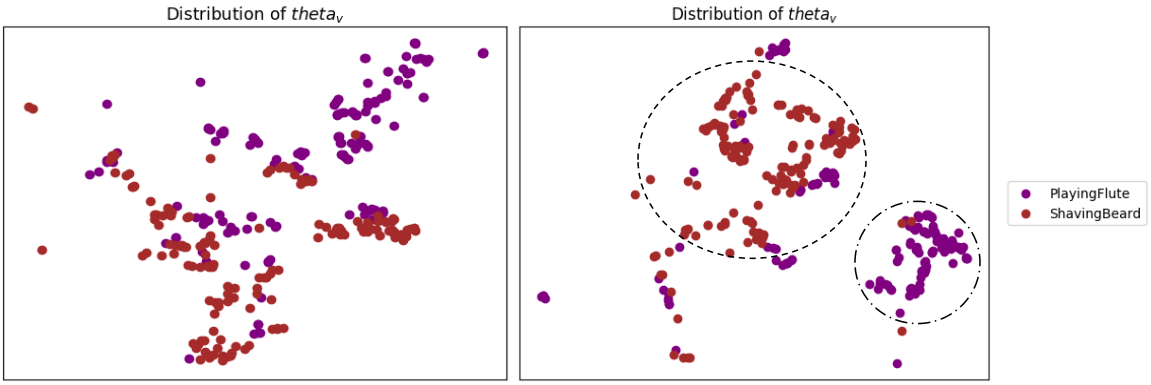
Figure 1. The framework of two toy designs: (a) Hyper-embedding and (b) Hyper-net.

Method	VGGSound-GZSL				UCF-GZSL				ActivityNet-GZSL			
	S ↑	U ↑	HM ↑	ZSL ↑	S ↑	U ↑	HM ↑	ZSL ↑	S ↑	U ↑	HM ↑	ZSL ↑
CJME [3]	8.69	4.78	6.17	5.16	26.04	8.21	12.48	8.29	5.55	4.75	5.12	5.84
AVGZSLNet [1]	18.05	3.48	5.83	5.28	52.52	10.90	18.05	13.65	8.93	5.04	6.44	5.40
AVCA [2]	14.90	4.00	6.31	**6.00**	51.53	18.43	**27.15**	20.01	24.86	8.02	**12.13**	9.13
Hyper-embedding	17.49	5.19	**8.00**	5.68	64.47	16.36	26.10	19.29	31.33	7.21	11.72	**9.90**
Hyper-net	8.34	4.46	5.81	5.80	54.56	17.19	26.14	**21.85**	16.17	8.87	11.46	**9.90**
Hyper-multiple	15.02	6.75	**9.32**	**7.97**	63.08	19.10	**29.32**	**22.24**	23.38	8.67	**12.65**	9.50

Table 2. Experimental results of audio-visual zero-shot learning on three datasets (main feature). AVCA [2] is adopted as the baseline for the proposed toy designs, Hyper-embedding and Hyper-net. The best results in HM and ZSL are in **bold**.

Method	VGGSound-GZSL^{cls}				UCF-GZSL^{cls}				ActivityNet-GZSL^{cls}			
	S ↑	U ↑	HM ↑	ZSL ↑	S ↑	U ↑	HM ↑	ZSL ↑	S ↑	U ↑	HM ↑	ZSL ↑
CJME [3]	10.86	2.22	3.68	3.72	33.89	24.82	28.65	29.01	10.75	5.55	7.32	6.29
AVGZSLNet [1]	15.02	3.19	5.26	4.81	74.79	24.15	36.51	31.51	13.70	5.96	8.30	6.39
AVCA [2]	12.63	6.19	8.31	6.91	63.15	30.72	41.34	37.72	16.77	7.04	9.92	7.58
Hyper-embedding	15.88	6.03	**8.74**	**7.19**	72.47	29.06	**41.48**	36.10	31.10	7.45	**12.02**	**8.37**
Hyper-net	7.56	3.16	4.45	4.91	33.77	37.24	35.42	**37.88**	21.15	7.60	11.18	8.32
Hyper-multiple	15.62	6.00	**8.67**	**7.31**	74.26	35.79	**48.30**	**52.11**	36.98	9.60	**15.25**	**10.39**

Table 3. Experimental results of audio-visual zero-shot learning on three datasets (cls feature). AVCA [2] is adopted as the baseline for the proposed toy designs, Hyper-embedding and Hyper-net. The best results in HM and ZSL are in **bold**.



(a) AVCA (b) Hyper-Multiple

Figure 2. Visualization examples on UCF-GZSL^{cls}. We give t-SNE visualizations of $\boldsymbol{\theta}_v$ from two unseen classes: "PlayingFlute" and "ShavingBeard".