ICLR International Conference On Learning Representations

Revisiting Deep Audio-Text Retrieval through the Lens of Transportation

Manh Luong, Khai Nguyen, Nhat Ho, Dinh Phung, Gholamreza Haffari, Lizhen Qu

MONASH University

Motivation

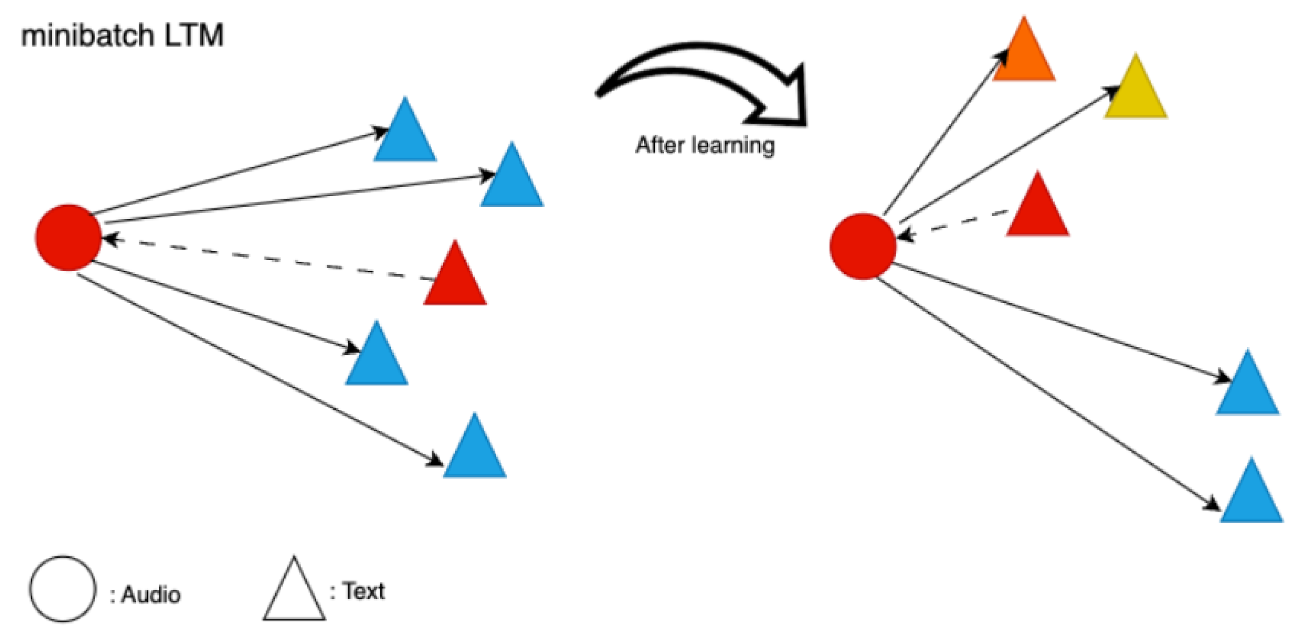


Figure 1: The minibatch Learning-to-Match framework.

- Both contrastive and triplet loss for audio-text retrieval treat all negative samples equally, therefore, they might learn a suboptimal metric space.
- Both contrastive and triplet loss are sensitive to noisy correspondence training data.
- To tackle these aforementioned issues, we propose the minibatch Learning-to-Match(m-LTM) framework to learn the joint embedding space across audio and text through the lens of optimal transport.

Mini-batch Learning-to-Match

Definition 1. *Given two encoder functions $f_\theta : \mathcal{X} \to \mathcal{Z}$ and $g_\phi : \mathcal{Y} \to \mathcal{Z}$, a metric $d : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}^+$, the Mahalanobis enhanced ground metric is defined as:*

$$c_{\theta,\phi,M}(x, y) = \sqrt{(f_\theta(x_i) - g_\phi(y_j))^\top M(f_\theta(x_i) - g_\phi(y_j))}, \tag{1}$$

for $\theta \in \Theta$ and $\phi \in \Phi$ which are spaces of parameters and M is a positive definite matrix.

Mini-batch learning to match with Mahalanobis-Enhanced Ground Metric. By using the family of Mahalanobis-Enhanced ground metrics in Definition 1, the m-LTM objective is defined as follows:

$$\min_{(\theta,\phi,M)\in\Theta\times\Phi\times\mathcal{M}} \mathbb{E}_{(X^b,Y^b)\sim D}[\mathrm{KL}(\hat{\pi}^b || \pi^{X^b,Y^b}_{\epsilon,c_{\theta,\phi,M}})], \tag{2}$$

where $\mathcal{M}$ is the set of all possible positive definite matrices e.g., $x^\top Mx > 0$ for all $x \in \mathcal{Z}$.

Hybrid stochastic gradient descent. the optimization problem in Equation 2 consists of three parameters θ, ϕ, and M. In contrast to θ and ϕ which are unconstrained, M is a constrained parameter. Therefore, we propose to use a hybrid stochastic gradient descent algorithm. In particular, we still update θ, ϕ using the estimated gradients. However, we update M using the projected gradient descent update. We first estimate the stochastic gradient with respect to M:

$$\nabla_M \mathbb{E}_{(X^b,Y^b)\sim D}[\mathrm{KL}(\hat{\pi}^b || \pi^{X^b,Y^b}_{\epsilon,c_{\theta,\phi,M}})] \approx \frac{1}{B}\sum_{i=1}^{B} \nabla_M \mathrm{KL}(\hat{\pi}^b || \pi^{(X^b,Y^b)_i}_{\epsilon,c_{\theta,\phi,M}})]. \tag{3}$$

After that, we update $M = \mathrm{Proj}(F(M, \nabla M))$ where $F(M, \nabla M)$ denotes the one-step update from a chosen optimization scheme

Partial OT for Noisy Correspondence

Setup. Given the training data $D = \{(x_i, y_i)\}_{i=1}^N$ where N is the number of training samples, a proportion of training data N_{cor}, $N_{cor} < N$, is corrupted, for instance, due to the data collection process. We denote a random variable $z \in \{0, 1\}$ which is sampled from a binomial distribution $Binomial(N, \frac{N_{cor}}{N})$, if $z = 1$ indicates the audio-text pair is shuffled. The training data is now $\tilde{D} = \{(z_i, x_i, y_i)\}_{i=1}^N$, where $z_i \sim Binomial(N, \frac{N_{cor}}{N})$

POT for noisy correspondence. we propose to use Partial OT, which relaxes the transportation preservation constraint, to mitigate the harmfulness of noisy empirical matching for approximating the incomplete matching $\bar{\pi}$. The objective function 2 is rewritten as

$$\min_{(\theta,\gamma,M)\in\Theta\times\Phi\times\mathcal{M}} \mathbb{E}_{(\tilde{X}^b,\tilde{Y}^b)\sim\tilde{D}}[\mathrm{KL}(\hat{\pi}^b || \pi^{\tilde{X}^b,\tilde{Y}^b}_{s,\epsilon,c_{\theta,\phi,M}})], , \tag{4}$$

, where $(\tilde{X}^b, \tilde{Y}^b)$ is a minibatch sampled from noisy training data $\tilde{D}$, and $\pi^{\tilde{X}^b,\tilde{Y}^b}_{\epsilon,s}$ is the optimal solution of the equation

$$\pi^{\tilde{X}^b,\tilde{Y}^b}_{s,\epsilon,c_{\theta,\gamma,M}} = \operatorname*{argmin}_{\pi\in\Pi_s(P_{\tilde{X}^b},P_{\tilde{Y}^b})} \sum_{i=1}^{b}\sum_{j=1}^{b} \pi_{ij} c(x_i, y_j) - \epsilon \sum_{i=1}^{b}\sum_{j=1}^{b} \pi_{ij} \log \pi_{ij}, \tag{5}$$

where $\Pi_s(P_{\tilde{X}^b}, P_{\tilde{Y}^b}) = \{\pi \in \mathbb{R}_+^{b\times b} | \pi\mathbb{1} \le P_{\tilde{X}^b}, \pi^\top\mathbb{1} \le P_{\tilde{Y}^b}, \mathbb{1}\pi^\top\mathbb{1} = s\}$.

Quantitative Results

Table 1: The comparison of m-LTM framework with baselines on audio-text retrieval task on two benchmark datasets, AudioCaps and Clotho dataset.

Dataset	Method	Text->Audio			Audio->Text		
		R@1	R@5	R@10	R@1	R@5	R@10
Audiocaps	(Oncescu et al., 2021)	28.1	-	79.0	33.7	-	83.7
	(Mei et al., 2022)	33.9	69.7	82.6	39.4	72	83.9
	(Deshmukh et al., 2022)	33.07	67.30	80.3	39.76	73.72	84.64
	(Wu et al., 2022b)	36.7	70.9	83.2	45.3	78	87.7
	m-LTM(our)	**39.10**	**74.06**	**85.78**	**49.94**	**80.77**	**90.49**
Clotho	(Oncescu et al., 2021)	9.6	-	40.1	10.7	-	40.8
	(Mei et al., 2022)	14.4	36.6	49.9	16.2	37.5	50.2
	(Deshmukh et al., 2022)	15.79	36.78	49.93	17.42	40.57	54.26
	(Wu et al., 2022b)	12.0	31.6	43.9	15.7	36.9	51.3
	m-LTM(our)	**16.65**	**39.78**	**52.84**	**22.1**	**44.4**	**56.74**

Expressiveness and Transferability

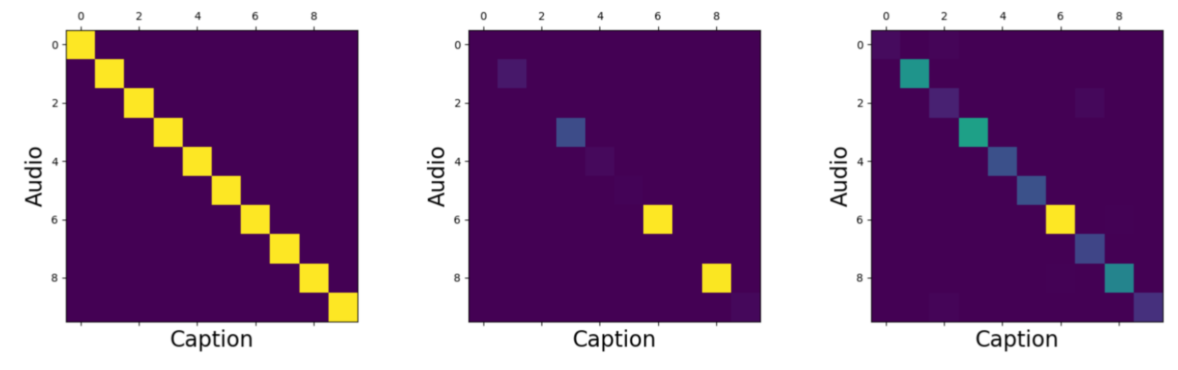
Table 2: The zero-shot sound event detection on the ESC50 test set, the $R@1$ score is equivalent to accuracy.

Loss	Audio->Sound			
	R@1	R@5	R@10	mAP
Triplet	71.25	91.75	95.75	80.09
Contrastive	72.25	93	96.75	80.84
m-LTM	81.0	97.0	99.25	87.57

Table 3: The modality gap between audio and text embedding in the shared embedding space. Lower is better for downstream tasks.

Loss	Modality gap($\|\vec{\Delta}_{gap}\|$)		
	AudioCaps	Clotho	ESC50
Triplet	0.149	0.283	0.937
Contrastive	0.181	0.266	0.922
m-LTM	0.117	0.142	0.224

Qualitative Results



(a) True matching (b) Contrastive loss (c) m-LTM loss

Figure 2: Qualitative results for text-to-audio retrieval task. top-1, top-2, and top-3 retrieved audio results are from left to right in the figure. The ground-truth audio for the caption is marked in red border.

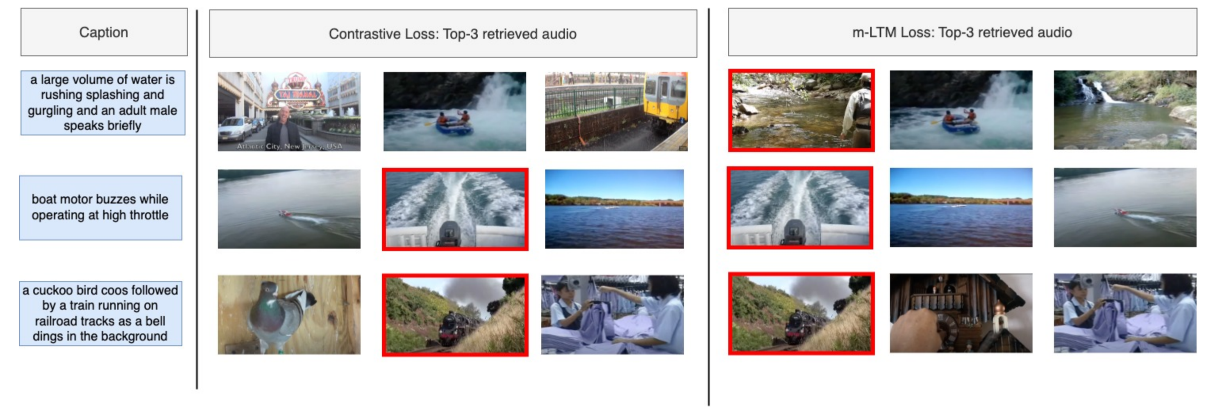


Figure 10: Qualitative results for text-to-audio retrieval task. top-1, top-2, and top-3 retrieved audio results are from left to right in the figure. The ground-truth audio for the caption is marked in red border.

Noisy Correspondence Tolerance

Table 4: The performance of learning-to-match and metric learning methods for audio-text retrieval task under the variant ratio of noisy training data.

Noise	Method	Text->Audio			Audio->Text		
		R@1	R@5	R@10	R@1	R@5	R@10
20%	Triplet loss	23.01	54.98	69.98	28.52	58.09	70.11
	Contrastive loss	31.34	67.73	81.27	40.12	70.84	82.54
	m-LTM	35.51	71.32	84.01	46.64	78.68	87.87
	m-LTM with POT	**35.92**	**72.28**	**84.11**	**47.12**	**79.2**	**88.19**
40%	Triplet loss	0.1	1.19	2.75	1.25	5.43	9.4
	Contrastive loss	26.68	62.98	78.18	34.69	66.66	78.99
	m-LTM	32.58	67.75	80.89	40.31	71.16	84.57
	m-LTM with POT	**33.64**	**69.23**	**82.27**	**42.63**	**73.35**	**86.1**
60%	Triplet loss	0.1	0.52	1.06	0.1	0.52	1.46
	Contrastive loss	20.58	53.96	70.72	27.37	58.72	75.21
	m-LTM	25.26	59.72	75.03	34.08	66.77	79.62
	m-LTM with POT	**27.73**	**62.61**	**76.17**	**35.42**	**68.65**	**80.56**