ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 978-1-7281-6327-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICASSP49357.2023.10096972

IMPROVING TEXT-AUDIO RETRIEVAL BY TEXT-AWARE ATTENTION POOLING AND PRIOR MATRIX REVISED LOSS

*Yifei Xin, Dongchao Yang, Yuexian Zou**

School of ECE, Peking University, Shenzhen, China

ABSTRACT

In text-audio retrieval (TAR) tasks, due to the heterogeneity of contents between text and audio, the semantic information contained in the text is only similar to certain frames within the audio. Yet, existing works aggregate the entire audio without considering the text, such as mean-pooling over the frames, which is likely to encode misleading audio information not described in the given text. In this paper, we present a text-aware attention pooling (TAP) module for TAR, which is essentially a scaled dot product attention for a text to attend to its most semantically similar frames. Furthermore, previous methods only conduct the softmax for every single-side retrieval, ignoring the potential cross-retrieval information. By exploring the intrinsic prior of each text-audio pair, we introduce a prior matrix revised (PMR) loss to filter the hard case with high (or low) text-to-audio but low (or high) audio-to-text similarity scores, thus achieving the dual optimal match. Experiments show that our TAP significantly outperforms various text-agnostic pooling functions. Moreover, our PMR loss also shows stable performance gains on multiple datasets.

***Index Terms*—** Text-audio retrieval, text-aware attention pooling, similarity matrix, dual optimal match

1. INTRODUCTION

Given a caption or an audio clip as a query, the text-audio retrieval (TAR) task aims at retrieving a paired item from a set of candidates in another modality. To compute the similarity between the two modalities, a common technique is to embed a text and an audio clip into a shared latent space and then adopt a distance metric like the cosine similarity to measure the relevance of the text and audio.

However, there is a significant disparity between both modalities that makes such a direct interaction challenging, that is, the heterogeneity of contents across different modalities [1–3]. Specifically, the semantic information contained in the text is typically similar to sub-segments of an audio clip. In this case, common text-agnostic aggregation schemes that pool entire audio frames, such as mean-pooling, might encode redundant or even distracting acoustic information that is not described in the given text. Moreover, depending on the input text, the most semantically similar frames would vary, so there could be multiple equally valid texts that match a specific audio clip. Therefore, we would expect the same audio to be retrieved for any of these queries and a retrieval model to prioritize the audio sub-segments that are most pertinent to the provided text.

Besides, previous TAR methods only perform the softmax operation along a single dimension for each retrieval pair [4, 5], which ignores the potential cross-retrieval information and harms the retrieval performance. To solve this, we introduce the dual optimal match hypothesis based on the discovered phenomenon from previous extensive experiments [6–8] that when a text-to-audio or audio-to-text pair reaches the optimal match (single-side match), the symmetric audio-to-text or text-to-audio scores should be the highest. With this hypothesis, a prior matrix revised (PMR) loss is introduced to revise the similarity matrix between the text and audio. Specifically, we first introduce a prior probability matrix calculated in the cross direction to adjust the original similarity score. Then, by conducting the dot product of the prior probability matrix and original scaling similarity matrix, we can filter the case with a high text-to-audio (or audio-to-text) similarity score but a low audio-to-text (or text-to-audio) similarity score, thus achieving the dual optimal match and leading to a more convincing result.

The major contributions of this paper are summarized as follows:

- We present a text-aware attention pooling (TAP) module that allows a model to reason about the most relevant audio frames to a provided text while suppressing frames not described in the given text.
- We introduce a prior matrix revised (PMR) loss to revise the similarity matrix between the text and audio, which imposes a direct constraint to filter those single-side match pairs and highlights more convincing results with the dual optimal match.
- Experiments show that our TAP significantly outper-

This paper was partially supported by NSFC (No: 62176008) and Shenzhen Science & Technology Research Program (No:GXWD20201231165807007-20200814115301001).

* Yuexian Zou is the corresponding author.

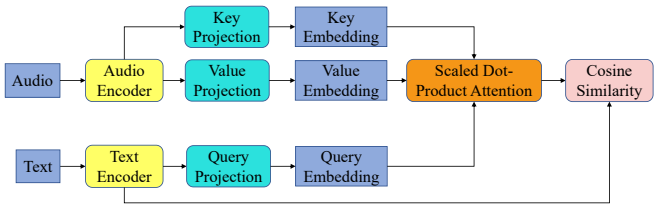

Fig. 1. Overview of our text-aware attention pooling module.

forms text-agnostic audio pooling functions. Furthermore, our PMR loss also shows stable performance gains on multiple datasets.

2. PROBLEM FORMULATION

We define two text-audio retrieval (TAR) tasks, where the text-to-audio retrieval is denoted as $t2a$ and the audio-to-text retrieval is denoted as $a2t$. In $t2a$, we are provided with a query text and an audio set. The target is to rank all audio clips according to their similarities with the query text. Similarly, in $a2t$, we are provided with a query audio clip and a text set. The aim is to retrieve matched texts based on their relevance with the query audio.

The TAR models usually consist of a text encoder (e.g., BERT-styled models that achieve superior performance on various NLP tasks [9]) and an audio encoder (e.g., pretrained audio tagging networks [10–12]), which project the text and audio into a shared embedding space, respectively. Specifically, given a text t and an audio clip a as input, the text encoder outputs the text embedding $c_t \in \mathbb{R}^D$, while the audio encoder is employed to generate the audio embedding $c_a \in \mathbb{R}^{T\times D}$, where D is the size of the model's channel dimension and T is the number of audio frames. In order to embed our given text and audio into a shared space to compute the similarity score, an aggregation function $p(\cdot)$ (e.g., mean-pooling) is utilized to pool the frame-level feature c_a into the clip-level latent embedding $z_a \in \mathbb{R}^D$:

$$z_a = p(c_a), \quad z_t = c_t. \tag{1}$$

Therefore, the similarity of the text and audio can be measured by the cosine similarity of their embeddings:

$$s(t,a) = \frac{z_t \cdot z_a}{\|z_t\| \, \|z_a\|}. \tag{2}$$

Currently, the NT-Xent loss [13, 14] based on symmetrical cross-entropy is widely employed, which has been shown to consistently outperform the previous triplet-based losses [4, 15] on both $t2a$ and $a2t$ tasks. Therefore, we adopt it as the baseline loss function for our work. The NT-Xent loss is formulated as below:

$$L = -\frac{1}{B}\left(\sum_i^B \log\frac{\exp(s(t_i,a_i)/\tau)}{\sum_j^B \exp(s(t_i,a_j)/\tau)} + \sum_i^B \log\frac{\exp(s(t_i,a_i)/\tau)}{\sum_j^B \exp(s(t_j,a_i)/\tau)}\right), \tag{3}$$

where B is the batch size, i and j denote the sample index in a batch, and τ is a temperature hyper-parameter. The training objective is to maximize the similarity of the positive pair relative to all negative pairs within a mini-batch, and the ultimate loss is calculated in both directions.

3. PROPOSED METHODS

3.1. Text-aware Attention Pooling

In existing TAR works, the aggregation function p does not directly consider the input text and is merely a function of the audio frames such as max-pooling, mean-pooling schemes [6]. However, a text is most semantically related to the subsegments of an audio clip. What's more, there could be multiple texts matching a specific audio clip, but the frames that are most semantically similar would vary. As such, text-agnostic aggregation functions would capture superfluous and distracting information not stated in the text, which impairs the TAR performance. Therefore, it is important to match a given text with its most semantically similar audio frames.

To that end, we present a learnable text-aware attention pooling (TAP) module ψ for TAR to perform the cross-modal reasoning on the audio frames that are most semantically related to a given text. The core mechanism is a scaled dot product attention [3, 16] between the text t and the frames of an audio clip a. By conditioning ψ on t, we can generate the audio aggregated embedding that learns to capture the most semantically similar audio frames as described in t. The resulting aggregated audio embedding is denoted as $z_{a|t}$, and our similarity function $s(t,a)$ is defined as:

$$z_{a|t} = \psi(c_a|t), \quad s(t,a) = \frac{z_t \cdot z_{a|t}}{\|z_t\| \, \|z_{a|t}\|}. \tag{4}$$

To elaborate, as shown in Fig. 1, we first project a text embedding $c_t \in \mathbb{R}^D$ output by the text encoder into a query $Q_t \in \mathbb{R}^{1\times D_p}$ and an audio embedding $c_a \in \mathbb{R}^{T\times D}$ generated by the audio encoder into key $K_a \in \mathbb{R}^{T\times D_p}$ and value $V_a \in \mathbb{R}^{T\times D_p}$ matrices, where D_p is the size of the projection dimension. The projections are defined as:

$$Q_t = \mathrm{LN}(c_t^T)W_Q, \tag{5}$$

$$K_a = \mathrm{LN}(c_a)W_K, \tag{6}$$

$$V_a = \mathrm{LN}(c_a)W_V, \tag{7}$$

where LN represents a layer normalization layer [17] and W_Q, W_K and W_V are projection matrices in $\mathbb{R}^{D\times D_p}$. In order to flexibly learn the relevance between the given text and the audio frames, we then adapt the scaled dot product attention from the query-projected text embedding to the key-projected frame embedding. The dot product attention provides relevance weights from a text to each audio frame, which we adopt to aggregate the value-projected frame embedding:

$$\text{Attention}(Q_t, K_a, V_a) = \text{softmax}(\frac{Q_t K_a^T}{\sqrt{D_p}})V_a. \quad (8)$$

Specifically, the query-projected text embedding is utilized to search frames with high relevance from the key-projected frame embedding. The value-projected embedding represents the audio's context, from which we aggregate frames conditioned on the given text. To embed an audio clip into a shared space with a text, we project the aggregated audio feature from the attention module back into $\mathbb{R}^D$ by leveraging a weight $W_O \in \mathbb{R}^{D_p\times D}$ to obtain:

$$z_{a|t} = \text{LN}(\text{Attention}(Q_t, K_a, V_a)W_O), \quad (9)$$

where the resulting output $z_{a|t}$ is an aggregated audio embedding depending on the text t. By introducing the text-aware attention pooling, the model can concentrate on the most pertinent audio frames as described in a given text, thus effectively mitigating the negative impacts of the heterogeneity of contents between text and audio. Next, we will introduce a prior matrix revised loss to further revise the results for both $t2a$ and $a2t$ tasks.

3.2. Prior Matrix Revised Loss

Previous TAR loss functions [13, 18, 19] (e.g., the NTXent loss) only conduct the softmax for every single-side retrieval, which is just inferred with the similarity score for each row in the original similarity matrix, thus ignoring the potential cross-retrieval information.

Based on the introduced dual optimal match hypothesis from previous extensive TAR experiments [6, 7, 13] that when a $t2a$ (or $a2t$) pair reaches the single-side match, the symmetric $a2t$ (or $t2a$) score should also be the highest, a prior probability matrix is introduced to be calculated in the cross direction for $t2a$ and $a2t$ to fully exploit the cross-retrieval information. Specifically, the prior matrix is obtained by calculating the softmax score along each column of the original similarity matrix and then we multiply the prior matrix with the original similarity matrix to revise the similarity score, that is, for the $t2a$ (or $a2t$) task, we first calculate the softmax score of $a2t$ (or $t2a$), and then incorporate it into the loss calculation of $t2a$ (or $a2t$). Lastly, we conduct the softmax operation along each row of the revised similarity matrix to get the final probability result. Our prior matrix revised (PMR) loss is formulated as below:

Table 1. Performance comparison of our TAP and previous text-agnostic pooling methods.

Methods	Text-to-Audio		Audio-to-Text	
	R@1	**R@10**	**R@1**	**R@10**
AudioCaps				
Mean	32.6±0.5	81.0±0.4	37.9±0.8	82.4±0.8
MeanMax	33.9±0.4	82.6±0.3	39.4±1.0	83.9±0.6
NetRVLAD	34.5±0.4	83.4±0.7	40.1±1.0	84.3±0.6
TAP	**36.1±0.2**	**85.2±0.4**	**41.3±0.5**	**86.1±0.3**
CNN14+Mean	29.8±0.5	78.5±0.4	40.3±0.7	81.3±0.3
CNN14+TAP	**33.1±0.4**	**81.3±0.8**	**42.8±0.4**	**84.1±0.2**
Clotho				
Mean	12.6±0.3	45.2±0.6	13.1±0.6	46.6±0.8
MeanMax	14.4±0.4	49.9±0.2	16.2±0.7	50.2±0.7
NetRVLAD	15.1±0.5	50.1±1.2	16.8±0.2	50.5±0.5
TAP	**16.2±0.6**	**50.8±0.3**	**17.6±0.5**	**51.4±0.6**
CNN14+Mean	12.2±0.8	46.1±0.5	12.4±0.6	47.1±0.4
CNN14+TAP	**15.1±0.4**	**49.3±0.6**	**15.3±0.3**	**51.1±0.3**

$$L_{t2a} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(s(t_i,a_i)\cdot Pr_{i,i}^{t2a}/\tau)}{\sum_j^B \exp(\boldsymbol{s(t_i,a_j)}\cdot Pr_{i,j}^{t2a}/\tau)}, \quad (10)$$

$$L_{a2t} = -\frac{1}{B}\sum_{i}^{B}\log\frac{\exp(s(t_i,a_i)\cdot Pr_{i,i}^{a2t}/\tau)}{\sum_j^B \exp(\boldsymbol{s(t_j,a_i)}\cdot Pr_{j,i}^{a2t}/\tau)}, \quad (11)$$

$$L = L_{t2a} + L_{a2t}, \quad (12)$$

where Pr^{t2a}, Pr^{a2t} denote the prior matrix for text-to-audio and audio-to-text tasks, respectively:

$$Pr_{i,j}^{t2a} = \frac{\exp(\omega\cdot s(t_i,a_i))}{\sum_j^B \exp(\omega\cdot \boldsymbol{s(t_j,a_i)})}, \quad (13)$$

$$Pr_{j,i}^{a2t} = \frac{\exp(\omega\cdot s(t_i,a_i))}{\sum_j^B \exp(\omega\cdot \boldsymbol{s(t_i,a_j)})}, \quad (14)$$

where ω represents a logit scaling parameter to smooth the gradients. In this way, $t2a$ and $a2t$ can coordinately revise each other's similarity scores, which provides prior knowledge of each other to filter the outliers and sharpen the more convincing points, thus achieving the dual optimal result.

4. EXPERIMENTS

4.1. Dataset

We evaluate our methods on two publicly available datasets: AudioCaps [20] and Clotho [21] datasets. AudioCaps contains about 50K audio samples, which are all 10-second long. The training set consists of 49274 audio clips, each with one corresponding human-annotated caption. The validation and test sets contain 494 and 957 audio clips, each with five

Table 2. Results of our PMR loss with previous methods.

Methods	Text-to-Audio		Audio-to-Text	
	R@1	**R@5**	**R@1**	**R@5**
AudioCaps				
MeanMax+NTXent	33.9	69.7	39.4	72.0
MeanMax+PMR	**34.1**	**70.2**	**39.6**	**72.8**
TAP+NTXent	36.1	72.0	41.3	75.5
TAP+PMR	**36.8**	**72.7**	**41.7**	**76.2**
CNN14+TAP+NTXent	33.1	68.6	42.8	72.7
CNN14+TAP+PMR	**33.4**	**68.8**	**43.1**	**73.3**
Clotho				
MeanMax+NTXent	14.4	36.6	16.2	37.5
MeanMax+PMR	**14.9**	**37.1**	**16.6**	**37.8**
TAP+NTXent	16.2	39.2	17.6	39.6
TAP+PMR	**17.1**	**39.6**	**18.2**	**39.9**
CNN14+TAP+NTXent	15.1	36.7	15.3	36.5
CNN14+TAP+PMR	**15.6**	**37.2**	**15.9**	**36.8**

human-annotated captions. The Clotho v2 dataset contains 6974 audio samples between 15 and 30 seconds in length. Each audio sample is annotated with 5 sentences. The numbers of training, validation, and test samples are 3839, 1045, and 1045, respectively.

4.2. Training Details and Evaluation metrics

In our work, we follow the same pipeline in [13] to train our network. We adopt BERT [9] as the text encoder, while employing the ResNet-38 in Pre-trained audio neural networks (PANNs) [10] as the audio encoder if not otherwise specified. We conduct experiments by fine-tuning the pre-trained models. The query, key and value projection dimension size is set as $D_p = 512$. Recall at rank k (R@k) is utilized as the evaluation metric, which is a popular cross-modal retrieval evaluation protocol. R@k measures the proportion of targets retrieved within the top-k ranked results, so a higher score means better performance. The results of R@1, R@5, and R@10 are reported.

4.3. Experimental Results

In this section, we first compare the performance of our TAP with various text-agnostic pooling functions on the AudioCaps and Clotho datasets. To demonstrate the superiority of our TAP, we compare it with previously popular aggregation strategies that achieve SOTA results, where Mean denotes the average pooling function, MeanMax [13] denotes we both use an average and max pooling layer to aggregate the frame-level feature, NetRVLAD [6, 22] is a descriptor-based pooling method that enables back-propagation by adopting soft assignment to clusters, and TAP represents our text-aware attention pooling method. The experiments are repeated three times with different training seeds.

As shown in Table 1, our TAP outperforms all other works that use text-agnostic pooling on all datasets and across all metrics, thereby highlighting the importance of our text-aware aggregation scheme that can learn to match a text with its most relevant frames while suppressing distracting information from irrelevant audio frames. In addition to adopting ResNet-38 as the backbone of our audio encoder, we also provide the results of using CNN14 [10] as our audio encoding model. It can be seen that our method also achieves performance boosts by a large margin, which strongly demonstrates the effectiveness and generalization of our TAP module. Notably, although the NetRVLAD aggregation method achieves relatively good results, it needs to manually select the number of clusters for different datasets and its tuning of the hyper-parameter is task and data specific. In contrast, our TAP can adaptively learn the optimal amount of information to extract for each text-audio pair, which thus removes the need to manually specify the hyper-parameter and can be more robust to different tasks and instances.

To evaluate our PMR loss, we compare it with the previous SOTA loss for TAR: NTXent, which has be shown in [13] to outperform the popular triplet-based losses. As can be seen in Table 2, our PMR loss shows stable performance boosts on both AudioCaps and Clotho datasets with different aggregation schemes. Besides, our PMR also achieves consistent performance gains on different baseline models, which further demonstrates the effectiveness of our PMR loss.

5. CONCLUSIONS

In this work, we first highlight the drawbacks of text-agnostic audio pooling functions and then propose a text-aware attention pooling (TAP) module for text-audio retrieval. Our TAP can learn to attend to the most relevant frames to a given text while suppressing frames not described in the text, thereby enabling the model to flexibly extract the most semantically relevant information of the audio frames. Furthermore, we introduce a prior matrix revised (PMR) loss to revise the similarity matrix between the text and audio. By introducing a prior probability matrix calculated in the cross direction, the hard case with only single-side match can be filtered, thus producing more convincing dual optimal results. Experiments show that our TAP performs better than various text-agnostic pooling functions. Moreover, our PMR loss also shows stable performance gains on publicly available AudioCaps and Clotho datasets.

References

[1] Mingyang Li, Shao-Lun Huang, and Lin Zhang, "Otcmr: Bridging heterogeneity gap with optimal transport for cross-modal retrieval," in *Proceedings of the*

30th ACM International Conference on Information & Knowledge Management, 2021, pp. 3216–3220.

[2] Chengyuan Zhang, Jiayu Song, Xiaofeng Zhu, Lei Zhu, and Shichao Zhang, "Hcmsl: Hybrid cross-modal similarity learning for cross-modal retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 17, no. 1s, pp. 1–22, 2021.

[3] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu, "X-pool: Cross-modal language-video attention for text-video retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5006–5015.

[4] Huang Xie, Samuel Lipping, and Tuomas Virtanen, "Dcase 2022 challenge task 6b: Language-based audio retrieval," *arXiv e-prints*, pp. arXiv–2206, 2022.

[5] Yongquan Lai, Jinsong Pan, and Buxian Chen, "A resnet-based clip text-to-audio retrieval system for dcase challenge 2022 task 6b," 2022.

[6] Siyu Lou, Xuenan Xu, Mengyue Wu, and Kai Yu, "Audio-text retrieval in context," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 4793–4797.

[7] A Sophia Koepke, Andreea-Maria Oncescu, Joao Henriques, Zeynep Akata, and Samuel Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, 2022.

[8] Xing Cheng, Hezheng Lin, Xiangyu Wu, Fan Yang, and Dong Shen, "Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss," *arXiv preprint arXiv:2109.04290*, 2021.

[9] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, pp. 4171–4186.

[10] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[11] Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, "Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 646–650.

[12] Yifei Xin, Dongchao Yang, and Yuexian Zou, "Audio pyramid transformer with domain adaption for weakly supervised sound event detection and audio classification," *Proc. Interspeech 2022*, pp. 1546–1550, 2022.

[13] Xinhao Mei, Xubo Liu, Jianyuan Sun, Mark D Plumbley, and Wenwu Wang, "On metric learning for audio-text cross-modal retrieval," *arXiv preprint arXiv:2203.15537*, 2022.

[14] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[15] T Lamort de Gail and D Kicinski, "Take it easy: Relaxing contrastive ranking loss with cider," Tech. Rep., DCASE2022 Challenge, Tech. Rep, 2022.

[16] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[17] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[18] Xinhao Mei, Xubo Liu, Haohe Liu, Jianyuan Sun, Mark D Plumbley, and Wenwu Wang, "Language-based audio retrieval with pre-trained models," Tech. Rep., Tech. rep., DCASE2022 Challenge, 2022.

[19] Andrew Koh and Eng Siong Chng, "Language-based audio retrieval with converging tied layers and contrastive loss," *arXiv preprint arXiv:2206.14659*, 2022.

[20] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, "Audiocaps: Generating captions for audios in the wild," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019, pp. 119–132.

[21] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[22] Antoine Miech, Ivan Laptev, and Josef Sivic, "Learnable pooling with context gating for video classification," *arXiv preprint arXiv:1706.06905*, 2017.