IMPROVING AUDIO-LANGUAGE LEARNING WITH MIXGEN AND MULTI-LEVEL TEST-TIME AUGMENTATION

Eungbeom Kim¹*, Jinhee Kim²*, Yoori Oh², Kyungsu Kim², Minju Park², Jaeheon Sim¹, Jinwoo Lee², Kyogu Lee¹,²,³

¹Interdisciplinary Program in Artificial Intelligence, Seoul National University
²Department of Intelligence and Information, Seoul National University
³Artificial Intelligence Institute, Seoul National University

ABSTRACT

In this paper, we propose two novel augmentation methods 1) audio-language MixGen (AL-MixGen) and 2) multi-level test-time augmentation (Multi-TTA) for audio-language learning. Inspired by MixGen, which is originally applied to vision-language learning, we introduce an augmentation method for the audio-language domain. We also explore the impact of test-time augmentations and present Multi-TTA which generalizes test-time augmentation over multiple layers of a deep learning model. Incorporating AL-MixGen and Multi-TTA into the baseline achieves 47.5 SPIDEr on audio captioning, which is an +18.2% over the baseline and outperforms the state-of-the-art approach with a 5x smaller model. In audio-text retrieval, the proposed methods surpass the baseline performance as well.

Index Terms— Data augmentation, Audio-language learning, Automated audio captioning, Audio-text retrieval

1. INTRODUCTION

Deep learning has been successful along with datasets containing a large number of data samples. However, collecting data samples is a cost-consuming process. Data augmentation would be an alternative way to improve generalization performance and alleviate data scarcity. Multi-modal learning is one of the areas where augmentation is prospective as various augmentation strategies are applicable. We can apply different augmentations on each modality, such as audio augmentations [1, 2] and text augmentations [3, 4]. At the same time, we can also augment multi-modalities jointly. MixGen [5], for instance, is a simple way to generate image-text pairs by applying mixup [6] on images and concatenating each corresponding text.

However, data augmentation on audio-language learning remains poorly understood even though it has the potential to improve training performances under a constrained data environment. AudioCaps [7], the largest public audio-language dataset, only includes about 40K audio clips and 46K captions in opposition to the MS COCO Captions [8] which is a widely known vision-language dataset containing 330K images and 1.5M captions. For this reason, in this study, we delve into data augmentation on audio-language learning.

As a train-time augmentation, we propose audio-language MixGen (AL-MixGen), a simple and effective multi-modal augmentation method originated from MixGen [5]. AL-MixGen leverages a mixup of audio clips and concatenation of their corresponding text captions to generate audio-text pairs as illustrated in Figure 1. We compare various audio mixup strategies for AL-MixGen such as waveform-level and mel spectrogram-level mixup. Furthermore, we suggest a method that randomly selects either waveform-level or mel spectrogram-level mixup with a certain probability for each sample.

Test-time augmentation (TTA) is another data augmentation approach which is shown to increase performances across diverse tasks [9, 10]. However, TTA is disregarded in most audio-language learning studies. In this paper, we investigate the effect of TTA on audio-language learning and present a novel multi-level test-time augmentation (Multi-TTA) methodology, which is a generalization of TTA over multiple layers. Multi-TTA successfully outperforms the baselines and we discover that there is still a room for improvement using augmentations at test-time.

We evaluate the proposed methods across automated audio captioning (AAC) and audio-text retrieval on AudioCaps. In AAC, AL-MixGen with audio augmentations achieves 46.6 SPIDEr, which is a 15.9% over the baseline without modifying its architecture and outperforms the state-of-the-art method with a 5x smaller model. Applying AL-MixGen and Multi-TTA results in an extra improvement and attains 47.5 SPIDEr.

AL-MixGen and Multi-TTA are also effective in audio-text retrieval. Compared to the baseline, recall at rank 10 (R@10) increases from 82.7% to 87.0% for audio to text retrieval, and from 81.3% to 83.3% for text to audio retrieval.

2. AUDIO-LANGUAGE MIXGEN

In this section, we introduce audio-language MixGen (AL-MixGen), a simple but effective augmentation method that generates audio-
text pairs. The original vision-language MixGen mixes randomly selected raw images and concatenates their corresponding texts. We extend this approach to the audio-language domain and present AL-MixGen. We propose different types of audio mixup strategies for AL-MixGen by selectively using waveform-level or mel spectrogram-level mixup. Given randomly sampled \(N\) audio-text pairs \(\{(a_i, t_i)\}_{i=1}^N\) containing audio \(a_i\) and text \(t_i\), Waveform-based AL-MixGen (Wav AL-MixGen) first generates \(\hat{a}\) by mixing two waveforms and then extracts its mel spectrogram \(\hat{m}\). Finally, Wav AL-MixGen generates a new audio-text pair \((\hat{m}, \hat{t})\) by concatenating text inputs to be consistent with original MixGen. Also, we observe that the randomness of WavMel AL-MixGen is helpful for the generalization of models. Since AL-MixGen can be easily applied with any other uni-modal augmentations, it is possible to drastically scale up the size of the dataset with a simple variant.

3. MULTI-LEVEL TEST-TIME AUGMENTATION

Test-time augmentation (TTA) contributes to generalizing models by making multiple predictions from augmented inputs and averaging the predictions. TTA has no additional training cost because TTA is only implemented at test-time as opposed to traditional augmentations which mainly focus on train-time. In this paper, we discuss the effects of conventional TTA for audio-language learning and present the generalized TTA method. As shown in Figure 2, Output-level TTA (Out-TTA) denotes conventional TTA that averages outputs from augmented inputs. We also present middle-level TTA (Mid-TTA) that averages intermediate representations and outputs a single prediction. Finally, we propose a novel multi-level TTA (Multi-TTA) methodology by generalizing Out-TTA and Mid-TTA. Multi-TTA flexibly chooses the number of TTA operations on each layer and sequentially implements a TTA operation on multiple layers.

We define a Multi-TTA strategy \(S = \{\{P_h\}_{h=1}^H, \tau\}\) given a model \(f\) containing \(H\) layers where \(\tau\) is the total number of input augmentations as follows. Let \(P_h\) be a partition of \([1, ..., |P_{h-1}|]\). Each element of \(P_h\) contains indices of the previous layer’s outputs that should be aggregated together. We define \(P_h[i]\) as an \(i\)-th element of \(P_h\) for simplicity. For clarity, \(|P_h|\) decides the number of \(h\)-th layer’s outputs and \(|P_h[i]|\) decides the number of required \(h\)-th layer’s inputs for an \(i\)-th output of \(h\)-th layer. A strategy \(S\) satisfies \(|P_h[i]| = 1\) to output a single prediction and \(\tau = \Pi_{h=1}^H |P_h|\). We define an \(i\)-th output of a \(h\)-th layer \(o_{h,i}\) for \(i = 1, ..., |P_h|\) as follows:

\[
o_{h,i} = \frac{1}{|P_h[i]|} \sum_{j \in P_h[i]} f_h(o_{h-1,j}).
\]

From the notations defined above, Out-TTA becomes a special case of Multi-TTA when \(|P_h| = \tau\) for \(1 \leq h < H - 1\) and \(|P_h| = 1\). Mid-TTA is also a special case of Multi-TTA when \(|P_h| = \tau\) for \(1 \leq h \leq h'\) and \(|P_h| = 1\) for \(h > h'\).

4. EXPERIMENTS

4.1. Dataset

AudioCaps is the largest audio-text paired dataset for audio-language learning. AudioCaps contains approximately 46K audio clips of 10 seconds extracted from AudioSet and their corresponding text descriptions annotated by humans. We investigate the capability of our proposed methods on AudioCaps for automated audio captioning and audio-text retrieval. We reproduce the baselines without unavailable audio clips for a fair comparison with the proposed methods.
4.2. Audio-Language Learning

4.2.1. Automated Audio Captioning

Audio Captioning Transformer (ACT) [12] has an encoder-decoder structure from Transformer [13]. The encoder block of ACT is initialized with DeiT [14] which was trained for image classification tasks and then pre-trained on an audio tagging task using AudioSet. We utilize the pre-trained ACT encoder and randomly initialize the ACT-m decoder for a baseline of AAC. For evaluation, we employ BLEU [15], METEOR [16], and ROUGE-L [17] from machine translation metrics. We also report CIDEr [18], SPICE [19], and SPIDEr [20]. Captions are decoded with beam search up to size 3.

4.2.2. Audio-Text Retrieval

Pre-trained encoders are employed and trained using contrastive loss for audio-text retrieval, following the previous study [21]. We choose PANNs [22] as the audio encoder, which has ResNet38 architecture trained on the audio tagging task using AudioSet. For the text encoder, we initialize our model with pre-trained BERT [23] and extract a text representation using the [CLS] token. Retrieval models are evaluated by Recall at rank k (R@k), mean average precision at rank k (mAP@k), and mean rank (meanR).

4.3. Experimental Setup

We experiment Wav AL-MixGen, Mel AL-MixGen, and WavMel AL-MixGen with different ratios \( K \in \{0.125, 0.25, 0.5, 0.6\} \) with other uni-modal audio augmentations further enhances the result.

Table 1: Experiment on different audio mixing types in automated audio captioning.

| Method       | CIDEr | SPICE | SPIDEr |
|--------------|-------|-------|--------|
| AL-MixGen    | 73.9  | 17.7  | 45.8   |
| Concat AL-MixGen | 72.7  | 17.2  | 45.0   |

where \( K \) is a ratio of \(#(\text{generated samples})\) in a single mini-batch to a size of mini-batch. Also, we compare a fixed mixup ratio \( \lambda = 0.5 \) to \( \lambda \sim \text{Beta}(0.1, 0.1) \). Additionally, we explore another design choice such as audio concatenation instead of audio mixup for AL-MixGen. We set \( N = 2 \), which is the number of audio-text pairs to generate a new pair, for every experiment and additionally load \( N \) audio-text pairs for AL-MixGen to avoid duplication within a minibatch. To validate the compatibility of AL-MixGen with uni-modal augmentations, Gaussian noise and reverberation are employed for audio with a probability of 0.5.

We validate Multi-TTA strategies where TTA operation is implemented at an intermediate layer \( h_E \) and an output layer \( H \). We set \( h_E \) as the last layer of the encoder for AAC and the last layer of the PANN audio encoder for audio-text retrieval. We compare the results of Out-TTA, Mid-TTA, and Multi-TTA with \( \tau \in \{10, 25, 50, 100\} \). For Multi-TTA, we employ the strategy \( S \) satisfying \( \tau, |P_{h_E}[i]|, |P_H[j]| \in \{10, 25, 50, 100\} \) for every \( i, j \). As we already mentioned, \( |P[k]| \) denotes the number of required \( l \)-th layer’s inputs for an single \( k \)-th output of layer \( l \). Note that we set \( |P[k]| \) for \( \forall k \) as identical value for simplicity of experiments and \( S \) satisfies \( \tau = |P_{h_E}[k]| \times |P_H[j]| \). We stabilize a prediction by averaging the output at which no augmentation is applied and the output of TTA. Types of data augmentation on test-time are consistent with train-time. Multi-modal augmentations are not included because generated samples are needless for evaluation. In test-time, we halve the maximum width and height of SpecAugment masking, reverberation degree, and the probability for applying audio augmentations compared to train-time.

All models are trained for 30 epochs using AdamW optimizer of learning rate \( 10^{-4} \) with the weight decay of \( 10^{-4} \). SpecAugment is applied to every model. Other hyperparameters follow [12] for AAC and [21] for audio-text retrieval.

5. RESULTS

Table 2 shows an overview of the results on AAC using AudioCaps. AL-MixGen with the uni-modal audio augmentations achieves 46.6% CIDEr.
Table 2: Evaluation of the model performance in automated audio captioning.

| Method         | Params | BLEU1 | BLEU2 | BLEU3 | BLEU4 | METEOR | ROUGE_L | CIDEr | SPICE | SPIDEr |
|----------------|--------|-------|-------|-------|-------|--------|---------|-------|-------|--------|
| Baseline [21]  | 108M   | 65.4  | 47.5  | 33.6  | 23.4  | 22.4   | 47.1    | 63.5  | 16.8  | 40.2   |
| + AL-MixGen    |        | 69.3  | 52.9  | 38.9  | 28.3  | 24.1   | 49.9    | 75.5  | 17.7  | 46.6   |
| + Multi-TTA    |        | 70.0  | 53.4  | 39.5  | 28.9  | 24.2   | 50.2    | 76.9  | 18.1  | 47.5   |
| [24]           | 494M   | 69.9  | 52.3  | 38.1  | 26.6  | 24.1   | 49.3    | 75.3  | 17.6  | 46.5   |

Table 3: Evaluation of the model performance in audio-text retrieval.

| Method         | Audio to Text | Text to Audio |
|----------------|---------------|---------------|
|                | R@1 | R@5 | R@10 | R@50 | mAP@10 | meanR | R@1  | R@5 | R@10 | R@50 | mAP@10 | meanR |
| Baseline [21]  | 37.9 | 71.0 | 82.7 | 97.4 | 28.7   | 8.8   | 33.0 | 67.9 | 81.3 | 96.5 | 47.7   | 10.0  |
| + AL-MixGen    | 40.0 | 73.1 | 86.8 | 97.3 | 30.5   | 6.6   | 34.4 | 69.6 | 83.1 | 97.5 | 49.5   | 8.2   |
| + Multi-TTA    | 40.0 | 73.4 | 87.0 | 97.7 | 30.7   | 6.4   | 34.7 | 70.0 | 83.3 | 97.5 | 49.7   | 8.0   |
| [25]           | 39.6 | 76.8 | 86.7 | 98.2 | -      | 6.5   | 36.1 | 72.0 | 84.5 | 97.6 | -      | 7.5   |

Fig. 5: Examples of generated captions in automated audio captioning. AL-MixGen correctly captures the overlapping events in a single audio clip unlike the others.

Table 4: Evaluation of the model performance in audio-text retrieval.

| Method         | Audio to Text | Text to Audio |
|----------------|---------------|---------------|
|                | R@1 | R@5 | R@10 | R@50 | mAP@10 | meanR | R@1  | R@5 | R@10 | R@50 | mAP@10 | meanR |
| Baseline       | 37.9 | 71.0 | 82.7 | 97.4 | 28.7   | 8.8   | 33.0 | 67.9 | 81.3 | 96.5 | 47.7   | 10.0  |
| + AL-MixGen    | 40.0 | 73.1 | 86.8 | 97.3 | 30.5   | 6.6   | 34.4 | 69.6 | 83.1 | 97.5 | 49.5   | 8.2   |
| + Multi-TTA    | 40.0 | 73.4 | 87.0 | 97.7 | 30.7   | 6.4   | 34.7 | 70.0 | 83.3 | 97.5 | 49.7   | 8.0   |
| [25]           | 39.6 | 76.8 | 86.7 | 98.2 | -      | 6.5   | 36.1 | 72.0 | 84.5 | 97.6 | -      | 7.5   |

5.2. Multi-Level Test-Time Augmentation

We study the effects of the number of augmentations $\tau$ on Out-TTA, Mid-TTA, and Multi-TTA. Out-TTA and Mid-TTA are comparable to Multi-TTA when $\tau = 10$, but Multi-TTA outperforms Out-TTA and Mid-TTA when $\tau \geq 25$ as shown in Figure 4. Finally, Multi-TTA with $\tau = 100$, $|P_{h_{[i]}[i]}| = 5$, and $|P_{i_{[j]}[j]}| = 20$ successfully improves every evaluation metric (Table 3).

Applying Multi-TTA to the audio-text retrieval model leads to an additional gain in performance. With $\tau = 25$, $|P_{h_{[i]}[i]}| = 5$, and $|P_{i_{[j]}[j]}| = 5$, R@10 of the model with Multi-TTA shows 87.0% for audio to text retrieval and 83.3% for text to audio retrieval. Unlike AAC where we can easily select intermediate layers to apply TTA as it has an encoder-decoder structure, audio-text retrieval models have more freedom to select any layers inside the encoder. Therefore, future research on automated layer search could further improve current results.

6. CONCLUSIONS

We demonstrate that appropriate augmentations improve audio-language learning performance. Specifically, experimental results show that AL-MixGen simply generates new audio-text pairs effectively. Also, we generalize TTA and propose Multi-TTA which enhances the efficiency of augmentations. Our proposed methods surpass the state-of-the-art method in AAC and shows competitive results in audio-text retrieval. Since AL-MixGen and Multi-TTA are flexible, they can be simply incorporated into other architectures.
7. REFERENCES

[1] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, “Spec augment: A simple data augmentation method for automatic speech recognition,” arXiv preprint arXiv:1904.08779, 2019.

[2] Helin Wang, Yuxian Zou, and Wenwu Wang, “Specaugment++: A hidden space data augmentation method for acoustic scene classification,” arXiv preprint arXiv:2103.16858, 2021.

[3] Jason Wei and Kai Zou, “Eda: Easy data augmentation techniques for boosting performance on text classification tasks,” arXiv preprint arXiv:1901.11196, 2019.

[4] Sergey Edunov, Myle Ott, Michael Auli, and David Grangier, “Understanding back-translation at scale,” arXiv preprint arXiv:1808.09381, 2018.

[5] Xiaoshuai Hao, Yi Zhu, Srikar Appalaraju, Aston Zhang, Wanqian Zhang, Bo Li, and Mu Li, “Mixgen: A new multi-modal data augmentation,” arXiv preprint arXiv:2206.08358, 2022.

[6] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz, “mixup: Beyond empirical risk minimization,” arXiv preprint arXiv:1710.09412, 2017.

[7] 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.

[8] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick, “Microsoft coco captions: Data collection and evaluation server,” arXiv preprint arXiv:1504.00325, 2015.

[9] Nikita Moshkov, Botond Mathe, Attila Kertesz-Farkas, Reka Hollandi, and Peter Horvath, “Test-time augmentation for deep learning-based cell segmentation on microscopy images,” Scientific reports, vol. 10, no. 1, pp. 1–7, 2020.

[10] Dmitry Molchanov, Alexander Lyzhov, Yuliya Molchanova, Arseni Ashukha, and Dmitry Vetrov, “Greedy policy search: A simple baseline for learnable test-time augmentation,” arXiv preprint arXiv:2002.09103, 2020.

[11] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, “Audio set: An ontology and human-labeled dataset for audio events,” in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2017, pp. 776–780.

[12] Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang, “Audio captioning transformer,” arXiv preprint arXiv:2107.09817, 2021.

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

[14] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou, “Training data-efficient image transformers & distillation through attention,” in International Conference on Machine Learning. PMLR, 2021, pp. 10347–10357.

[15] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, “Bleu: a method for automatic evaluation of machine translation,” in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.

[16] Satanjeev Banerjee and Alon Lavie, “Meteor: An automatic metric for mt evaluation with improved correlation with human judgments,” in Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.

[17] Chin-Yew Lin, “Rouge: A package for automatic evaluation of summaries,” in Text summarization branches out, 2004, pp. 74–81.

[18] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, “Cider: Consensus-based image description evaluation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.

[19] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, “Spice: Semantic propositional image caption evaluation,” in European conference on computer vision. Springer, 2016, pp. 382–398.

[20] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy, “Improved image captioning via policy gradient optimization of spider,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 873–881.

[21] 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.

[22] Qujiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, “Panns: Large-scale pre-trained audio neural networks for audio pattern recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2880–2894, 2020.

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” arXiv preprint arXiv:1810.04805, 2018.

[24] Félix Gontier, Romain Serizel, and Christophe Cerisara, “Automated audio captioning by fine-tuning bart with audioset tags,” in Detection and Classification of Acoustic Scenes and Events-DCASE 2021, 2021.

[25] A Sophia Koepe, Andrea-Maria Oncescu, Joao Henriques, Zeynep Akata, and Samuel Albanie, “Audio retrieval with natural language queries: A benchmark study,” IEEE Transactions on Multimedia, 2022.