A KNOWLEDGE DISTILLATION APPROACH TO IMPROVING LANGUAGE-BASED AUDIO RETRIEVAL MODELS

Technical Report

Paul Primus[1], *Gerhard Widmer*[1,2]

[1]Institute of Computational Perception (CP-JKU)
[2]LIT Artificial Intelligence Lab
Johannes Kepler University, Austria

ABSTRACT

This technical report describes the CP-JKU team's submissions to the language-based audio retrieval task of the 2024 DCASE Challenge (Task 8). All our submitted systems are based on the dual encoder architecture that projects recordings and textual descriptions into a shared audio-caption space in which related examples from the two modalities are similar. We utilized pretrained audio and text embedding models and trained them on audio-caption datasets (WavCaps, AudioCaps, and ClothoV2) via contrastive learning. We further fine-tuned the resulting models on ClothoV2 via knowledge distillation from a large ensemble of audio retrieval models. Our best single system submission based on PaSST and RoBERTa achieves a mAP@10 of 39.77 on the ClothoV2 test split, outperforming last year's best single system submission by around 1pp. without utilizing metadata and synthetic captions. An ensemble of three distilled models achieves 41.91 mAP@10 on the ClothoV2 test split.

1. INTRODUCTION

Task 8 of the 2024 DCASE challenge [1] invited participants to train systems that can retrieve audio recordings from a database based on textual descriptions. Such systems are of practical interest because they allow users to intuitively specify arbitrary acoustic concepts of interest (such as acoustic events, qualities of sound, and temporal relationships) without relying on a predefined set of tags or categories. However, language-based retrieval is difficult from a technical perspective because it requires connecting raw audio recordings with textual descriptions. Typical audio retrieval systems [2–6] achieve this via a dual-encoder architecture that projects the textual query and the candidate audio recordings into a shared multimodal metric space where the audios are then ranked based on their distance to the textual query (for a different approach, see [7]).

Previous studies have explored multiple directions to improve language-based audio retrieval systems, such as using better pretrained embedding models [8], augmentation techniques for both audio and text [9], artificial captions generated with large language models [8, 10, 11], and hybrid content and metadata based retrieval systems [12]. The systems described in this report build on top of our submission to the DCASE Challenge 2023 [13]. This year, we explored a new direction to improving language-based audio retrieval systems, namely via knowledge distillation from an ensemble of pre-trained retrieval models. Our objective was to estimate audio-caption correspondences for all audios in the training set and use those to train better retrieval models. To this end, we propose a two-step training procedure that is illustrated in Figure 1. In the following sections, we motivate and describe the proposed two-stage training strategy; we then detail the experimental setup, present results on the ClothoV2 benchmark [14], and summarize the four submitted systems.

2. MOTIVATION

Our submitted retrieval systems consists of two modality encoder networks, one for audio and one for caption embedding, denoted as $\phi_a(\cdot)$ and $\phi_c(\cdot)$, respectively. These encoders have learned to embed audio recordings and descriptions into a shared D-dimensional embedding space such that representations of matching audios and captions are similar. The agreement between audio a_i and description c_j at training and inference time is estimated via the normalized dot product in the shared embedding space:

$$C_{ij} = \frac{\phi_{\text{a}}(a_i)^T \cdot \phi_{\text{c}}(c_j)}{\|\phi_{\text{a}}(a_i)\|^2 \, \|\phi_{\text{t}}(c_j)\|^2}$$

For training, we relied on the normalized temperature-scaled cross-entropy loss [15], which converts those similarities into conditional probability distributions over audios and captions via a temperature-scaled softmax operation, where

$$q_a(a_i \mid c_j) = \frac{e^{C_{ij}/\tau}}{\sum_{i=1}^{N} e^{C_{ij}/\tau}}$$

gives the estimated probability that audio a_i corresponds to a given caption c_j, and

$$q_c(c_j \mid a_i) = \frac{e^{C_{ij}/\tau}}{\sum_{j=1}^{N} e^{C_{ij}/\tau}}$$

gives the estimated probability that caption c_j corresponds to a given audio a_i. The training objective is then to minimize the cross-entropy (denoted as H) between the estimated and the actual correspondence probabilities, q and p, respectively.

$$\mathcal{L}_{\text{sup}} = H(p_a, q_a) + H(p_c, q_c)$$

However, the correspondence probabilities p for audio a_i and caption c_j with $i \neq j$ are not generally available because audio retrieval data sets typically only provide a set of N matching audio

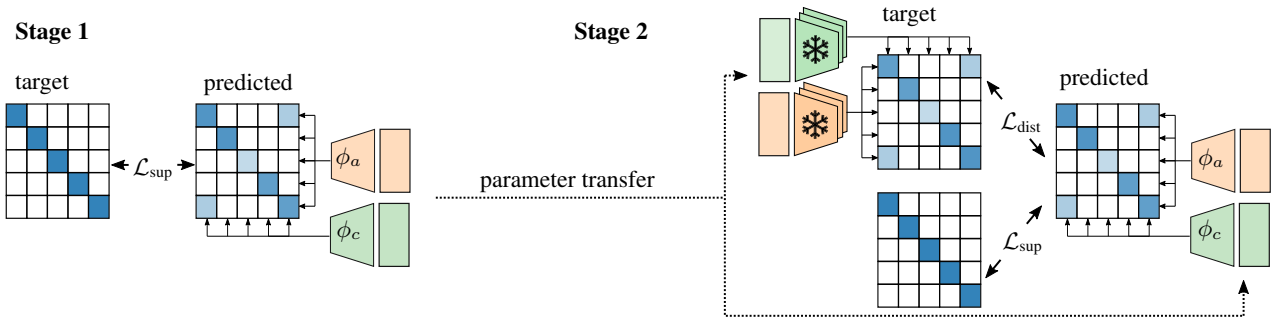

Figure 1: Overview of the two-step training procedure: Audio and descriptions are transformed into the shared audio-caption embedding space via the audio and description embedding models ϕ_{a} and ϕ_{c}, respectively. In stage one, we assume that audio a_i and caption c_j do not match if $i \neq j$ an train the model on $\mathcal{L}_{\text{sup}}$. Stage two uses the ensembled predictions to estimate the correspondence between a_i and c_j; the corresponding loss is denoted as $\mathcal{L}_{\text{dist}}$.

and caption pairs $\{(a_i, c_i)\}_{i=1}^{N}$, but no correspondence annotations for $i \neq j$. Previous studies thus assumed that c_j does not describe a_i if $i \neq j$, which is reasonable if the data set is large and holds a large variety of audio recordings with specific descriptions. The target probability distribution p for audios and captions based on this assumption can then be defined as follows:

$$p_a(a_i \mid c_j) := \mathbb{1}_{i=j} \text{ and } p_c(c_j \mid a_i) := \mathbb{1}_{i=j}$$

We argue that relying on this assumption is not ideal, mainly for two reasons:

1. It is only valid if each caption in the dataset describes *exactly one* audio recording, which is not the case in ClothoV2, AudioCaps, and WavCaps, as illustrated in Table 1.
2. Binary correspondences are limited to exact matches between audio recordings and captions. However, a caption can partially match an audio recording; previous research has shown that soft annotations can provide useful information during learning [16].

Some efforts have been made to obtain pairwise correspondence scores of audios and captions [16], but these valuable annotations are scarce due to the high cost associated with annotating N^2 audio-caption pairs.

3. PROPOSED METHOD

In order to remove the previous assumption without relying on human annotators, we estimated the correspondences in a two-step training procedure: We first pre-trained M models (as described before) and used them to estimate the pairwise correspondences between all audios and captions in the training set. To this end, we averaged the predicted pairwise similarities as follows:

$$\hat{C}_{ij} = \frac{1}{M} \sum_{m=1}^{M} C_{ij}^{m}$$

Then, we fine-tuned the previous models with a knowledge distillation like procedure on the estimated correspondences. We applied the softmax activation over audio recordings

$$\hat{p}_a(a_i \mid c_j) := \frac{e^{C_{ij}/\tau}}{\sum_{i=1}^{N} e^{C_{ij}/\tau}}$$

and captions

$$\hat{p}_c(c_j \mid a_i) := \frac{e^{C_{ij}/\tau}}{\sum_{j=1}^{N} e^{C_{ij}/\tau}}$$

to obtain estimated correspondence probabilities $\hat{p}_a$ and $\hat{p}_c$, and used them as prediction targets instead of p_a and p_c, respectively. The corresponding distillation loss is:

$$\mathcal{L}_{\text{dist}} = H(\hat{p}_a, q_a) + H(\hat{p}_c, q_c)$$

We traded off both losses with hyperparameter λ to fine-tune the models from the previous stage:

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda \mathcal{L}_{\text{dist}}$$

4. EXPERIMENTAL SETUP

The following section explains the implementation details of the submitted systems.

4.1. Datasets

We trained our models in two steps on multiple datasets. First, we performed pretraining on ClothoV [14], AudioCaps [17], and WavCaps [10]. The resulting models were then finetuned with the previously introduced knowledge distillation procedure using only the ClothonV2 data set.

4.1.1. ClothoV2

ClothoV2 [14] contains 10-30 second-long audio recordings and captions that are between 8 and 20 words long. The development set's training, validation, and test split suggested by the organizers contains 3840, 1045, and 1045 recordings, respectively, and each recording is associated with five human-generated captions. The leaderboard evaluation split used for the final system ranking contains 1000 audio recordings and 1000 captions. We used the validation split to monitor the generalization performance and report the performance on the test spit in this report.

data set	caption	audio-IDs
ClothoV2 [14]	A person is walking along a gravel path.	Gravel road walk, BodySC
AudioCaps [17]	A baby cries and a woman speaks	KuYnSQdcsss, a6x-RPqK3L
WavCaps [10]	Music is playing	YNuks8XFdGMk, Y-0Gj8-vB1q4

Table 1: Examples of audio recordings that are associated with the same caption. Each data set contains duplicate captions, but we note that ClothoV2 has a higher diversity and more specific captions.

4.1.2. AudioCaps

AudioCaps [17] contains 51.308 audio recordings taken from AudioSet [18] and one human-written caption for each of them. Each audio recording has a duration of 10 seconds, and the captions are, on average, 9.8 words long. We concatenated the training, validation, and testing split of AudioCaps into one large dataset and utilized it for pretraining.

4.1.3. WavCaps

WavCaps [10] is a weakly-labeled audio-caption dataset that contains 403.050 audio recordings of varying lengths collected from FreeSound, BBC Sound Effects, SoundBible, and the strongly supervised AudioSet subset. Each audio file is associated with a synthetic audio caption that was created by instructing the GPT3.5-turbo model to extract relevant sound events from metadata and output a single-sentence description. The generated captions are, on average, 7.8 words long. The dataset is available on HuggingFace[1]. In order to comply with the updated rules this year, we excluded the overlapping recordings between WavCaps and the evaluation subsets of Clotho.

4.2. Audio Embedding Models

We experimented with three audio embedding models and describe them briefly below.

4.2.1. PaSST

The Patchout faSt Spectrogram Transformer (PaSST) [19] uses pretrained parameters taken from a vision transformer [20,21] and finetunes them on AudioSet for general-purpose audio tagging. PaSST achieves a relatively low computational and memory footprint by dropping patches from the input sequence. The model holds a positional encoding for inputs of up to 10 seconds, so we cut the up to 30-second long inputs into non-overlapping 10-second snippets and averaged their embeddings. We used the version of PaSST without patch overlap and applied structured patchout of 2 and 15 over frequency and time dimensions, respectively. PaSST has approximately 86.2 million trainable weights and achieves a mAP of 46.8 on the AudioSet test set. We used the checkpoint denoted as `passt_s_p16_s16_128_ap468` in our experiments, which is available via GitHub[2].

4.2.2. ATST

ATST-Frame [22] (denoted only as ATST in the following) is also based on the vision transformer architecture [20]; in contrast to PaSST, however, it takes spectrogram frames instead of patches as input and is not initialized with ViT parameters. Instead, ATST is pre-trained in a self-supervised manner on the audio recordings in AudioSet. The model also has a positional encoding limited to 10 seconds, so we again cut the up to 30-second long audio recordings into non-overlapping 10-second snippets and averaged their embeddings. During training, we used frequency warping [22] where at most 10% of the higher frequency bins were dropped. We utilized a publicly available checkpoint of ATST (called `atst_as2M.ckpt`) that was further fine-tuned on the weak labels of AudioSet[3]. ATST has approximately 85.4 million trainable weights and achieves a mAP of 48.0 on the AudioSet test set.

4.2.3. MobileNetV3

We further experimented with a pretrained CNN model [23] that is based on the MobileNetV3 [24] architecture (referred to as MN in the following). The model was pre-trained on AudioSet using knowledge distillation from an ensemble of audio spectrogram transformers [19]. This architecture is particularly well suited for experiments with ClothoV2 because it can handle audio recordings of arbitrary length as input. Pre-trained checkpoints for models of varying sizes are available on GitHub. We used the model with ID `mn40_as_ext` in our experiments [4]. The selected MN has approximately 68.4 million trainable weights, and the selected checkpoint achieves a mAP of 48.7 on the AudioSet test set.

4.3. Sentence Embedding Models

We further conducted our experiments with RoBERTa large [25] as a sentence embedding model because it gave the best performance in the last year's challenge. RoBERTa is a bi-directional self-attention-based sentence encoder that underwent self-supervised pretraining on the BookCorpus [26], and WikiText datasets [27]. We selected the output vector that corresponds to the class token as sentence embedding. The RoBERTa large has around 354 million parameters. Pretrained models were taken from HuggingFace[5].

4.4. Preprocessing

To allow batched processing of recordings of varying lengths, we extracted random 30-second snippets from those audio recordings that are longer than 30 seconds and zero-padded shorter recordings to the maximum duration in the current batch. Depending on the audio embedding model, we used the methods described in the original papers to extract spectrograms from waveforms [19,22,23]. The input sentences were pre-processed by transforming all characters to lowercase and removing punctuation. The resulting strings were

[1] `https://huggingface.co/datasets/cvssp/WavCaps`
[2] `https://github.com/kkoutini/PaSST`
[3] `https://github.com/Audio-WestlakeU/ATST-SED`
[4] `https://github.com/fschmid56/EfficientAT`
[5] `https://huggingface.co/`

audio embedding	text embedding	fine tuning	λ	mAP@10	R@1	R@5	R@10	SID
PaSST	roberta-large	✗	0	35.7	23.83	51.79	64.96	
ATST	roberta-large	✗	0	32.31	21.45	47.01	59.39	
MobileNetV3	roberta-large	✗	0	34.06	22.49	49.19	63.31	
PaSST	roberta-large	✓	1	39.77	27.12	57.13	69.86	2
ATST	roberta-large	✓	1	38.96	26.81	54.87	68.82	4
MobileNetV3	roberta-large	✓	1	37.73	25.21	54.66	68.04	3
ensembe of [2,3,4]				41.90	29.33	59.311	71.923	1

Table 2: Retrieval performance of the three models after pre-training (first section) and after fine tung (second section).

tokenized with the WordPiece tokenizer [28], padded to the maximum sequence length in the current batch, and truncated if they were longer than 32 tokens.

4.5. Training

We pre-trained all models on AudioCaps, WavCaps, and ClothoV2. Both modality encoder models were jointly optimized using gradient descent with a batch size of 64 for PaSST and ATST and 32 for MN. We used the Adam update rule [29] for 20 epochs, with one warmup epoch. Thereafter, the learning rate was decayed from 2×10^{-5} to 10^{-7} using a cosine schedule. The hyperparameters of the optimizer were set to PyTorch's [30] defaults.

Finetuning was done on ClothoV2 only but with the same configuration as for pre-training. We ensembled the similarity scores of all three models as described in Section 3 to obtain audio-caption correspondence estimates. We used $\tau = 0.05$ and $\lambda = 1$ in all our experiments. Our main evaluation criterion for hyperparameter selection was the mean Average Precision among the top-10 results (mAP@10) on the validation set, which is the metric used for ranking submitted systems. In the results section, we additionally report the recall among the top-1, top-5, and top-10 retrieved results.

5. RESULTS

Table 2 summarizes the results for pre-training and fine-tuning in the first and second sections, respectively. We note a considerable increase in mAP@10 when fine-tuning the pretrained models with knowledge distillation, namely 4.07, 6.86, and 3.67 pp. for PaSST, ATST, and MN, respectively. Our best-performing model outperforms last year's best single system submission (Submission 2 of [13]) by around 1pp without utilizing metadata and synthetic captions. An ensemble of three distilled models achieved a mAP@10 of 41.91 on the ClothoV2 test split, which is competitive with the previous year's best ensemble (Submission 1 of [13]) that achieved a score of 41.4 with more than twice as many models.

6. SUBMISSION

Since the training procedure was fairly stable and in order to remain competitive, we retrained all previously discussed models and utilized the whole ClothoV2 development set (i.e., train, validation, and test splits) instead of the ClothoV2 training split only. We submitted predictions generated with four different systems to the challenge; the numbers in the following list correspond to the numbers in the SID column in Table 2:

1. an ensemble of three models (ensembled models are indicated in Table 2)
2. PaSST and roberta-large, finetuned with knowledge distillation on ClothoV2
3. MobileNetV3 and roberta-large, finetuned with knowledge distillation on ClothoV2
4. ATST and roberta-large, fine-tuned with knowledge distillation on ClothoV2

7. ACKNOWLEDGMENT

The LIT AI Lab is financed by the Federal State of Upper Austria. The computational results presented in this work have been partially achieved using the Vienna Scientific Cluster (VSC).

8. REFERENCES

[1] "Language-based audio retrieval, Task description website," https://dcase.community/challenge2024/task-language-based-audio-retrieval, accessed: 2024-06-20.

[2] A. S. Koepke, A. Oncescu, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Trans. Multim.*, vol. 25, pp. 2675–2685, 2023.

[3] Y. Xin, D. Yang, and Y. Zou, "Improving text-audio retrieval by text-aware attention pooling and prior matrix revised loss," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP, Rhodes Island, Greece*, 2023.

[4] S. Lou, X. Xu, M. Wu, and K. Yu, "Audio-text retrieval in context," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing, ICASSP, Singapore*, 2022.

[5] X. Mei, X. Liu, J. Sun, M. D. Plumbley, and W. Wang, "On metric learning for audio-text cross-modal retrieval," in *23rd Annual Conf. of the Int. Speech Communication Association, Interspeech, Incheon, Korea*, 2022.

[6] H. Xie, O. Räsänen, K. Drossos, and T. Virtanen, "Unsupervised audio-caption aligning learns correspondences between individual sound events and textual phrases," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process., ICASSP*, 2022.

[7] E. Labbé, T. Pellegrini, and J. Pinquier, "Killing two birds with one stone: Can an audio captioning system also be used for audio-text retrieval?" in *Proc. of the 8th Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE 2022, Helsinki, Finland*, 2022.

[8] P. Primus, K. Koutini, and G. Widmer, "Advancing natural-language based audio retrieval with passt and large audio-caption data sets," in *Proc. of the 8th Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE, Helsinki, Finland*, 2023.

[9] P. Primus and G. Widmer, "Improving natural-language-based audio retrieval with transfer learning and audio & text augmentations," in *Proc. of the 7th Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE, Nancy, France*, 2022.

[10] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A ChatGPT-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," *CoRR*, vol. abs/2303.17395, 2023.

[11] G. Zhu and Z. Duan, "Cacophony: An improved contrastive audio-text model," *CoRR*, vol. abs/2402.06986, 2024.

[12] P. Primus and G. Widmer, "Fusing audio and metadata embeddings improves language-based audio retrieval," *CoRR*, vol. abs/, 2024.

[13] P. Primus, K. Koutini, and G. Widmer, "Cp-jku's submission to task 6b of the dcase2023 challenge: Audio retrieval with passt and gpt-augmented captions," DCASE2023 Challenge, Tech. Rep., June 2023.

[14] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: an Audio Captioning Dataset," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process., ICASSP*, 2020.

[15] T. Chen, S. Kornblith, M. Norouzi, and G. E. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. of the 37nd Int. Conf. on Machine Learning, ICML*, 2020.

[16] H. Xie, K. Khorrami, O. Räsänen, and T. Virtanen, "Crowdsourcing and evaluating text-based audio retrieval relevances," in *Proc. of the 8th Workshop on Detection and Classification of Acoustic Scenes and Events, DCASE, Helsinki, Finland*, 2023.

[17] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in *Proc. of the North American Ch. of the Ass. for Computational Linguistics: Human Language Technologies, NAACL-HLT*, 2019.

[18] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio Set: An ontology and human-labeled dataset for audio events," in *Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process., ICASSP*, 2017.

[19] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *23rd Annual Conf. of the Int. Speech Communication Association, Interspeech*, 2022.

[20] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *CoRR*, vol. abs/2010.11929, 2020.

[21] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. of the 38th Int. Conf. on Machine Learning, ICML*, 2021.

[22] X. Li, N. Shao, and X. Li, "Self-supervised audio teacher-student transformer for both clip-level and frame-level tasks," *IEEE ACM Trans. Audio Speech Lang. Process.*, 2024.

[23] F. Schmid, K. Koutini, and G. Widmer, "Efficient large-scale audio tagging via transformer-to-CNN knowledge distillation," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing ICASSP, Rhodes Island, Greece*, 2023.

[24] A. Howard, R. Pang, H. Adam, Q. V. Le, M. Sandler, B. Chen, W. Wang, L. Chen, M. Tan, G. Chu, V. Vasudevan, and Y. Zhu, "Searching for MobileNetV3," in *IEEE/CVF Int. Conf. on Computer Vision, ICCV 2019, Seoul, Korea (South)*, 2019.

[25] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019.

[26] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *IEEE Int. Conf. on Computer Vision, ICCV*, 2015.

[27] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," in *5th Int. Conf. on Learning Representations, ICLR*, 2017.

[28] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's neural machine translation system: Bridging the gap between human and machine translation," *CoRR*, vol. abs/1609.08144, 2016.

[29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd Int. Conf. on Learning Representations, ICLR*, 2015.

[30] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Köpf, E. Z. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "PyTorch: An imperative style, high-performance deep learning library," in *Annual Conf. on Neural Information Processing Systems, NEURIPS*, 2019.