COMPARISON OF SOFT AND HARD TARGET RNN-T DISTILLATION FOR LARGE-SCALE ASR

Dongseong Hwang, Khe Chai Sim, Yu Zhang, Trevor Strohman

Google LLC, USA
{dongseong, khechai, ngyuzh, strohman}@google.com

ABSTRACT

Knowledge distillation is an effective machine learning technique to transfer knowledge from a teacher model to a smaller student model, especially with unlabeled data. In this paper, we focus on knowledge distillation for the RNN-T model, which is widely used in state-of-the-art (SoTA) automatic speech recognition (ASR). Specifically, we compare using soft and hard target distillation to train large-scale RNN-T models on the LibriSpeech/LibriLight public dataset (60k hours) and our in-house data (600k hours). We found that hard targets are more effective when the teacher and student have different architecture, such as large teacher and small streaming student. On the other hand, soft target distillation works better in self-training scenario like iterative large teacher training. For a large model with 0.6B weights, we achieve a new SoTA word error rate (WER) on LibriSpeech (8% relative improvement on dev-other) using Noisy Student Training with soft target distillation. It also allows our production teacher to adapt new data domain continuously.

Index Terms— RNN Transducer, Knowledge Distillation, Noisy Student Training, Semi-supervised learning

1. INTRODUCTION

The success of end-to-end (E2E) speech recognition models [1, 2, 3] are highly dependent on having a large amount of high-quality transcribed speech data. However, it is expensive and difficult to get the high-quality human transcriptions, which restricts the development of automatic speech recognition (ASR).

Knowledge distillation [4] is an effective technique to transfer knowledge from a teacher model to a student model. Noisy Student Training (NST) [5, 6] is a well established method that applies knowledge distillation in an iterative training fashion to progressively improve the model. For each iteration, NST transcribes a large amounts of unlabeled data from teacher model and use it to train the student model with data augmentation. NST has been shown to be effectiveness for ImageNet and the LibriSpeech automatic speech recognition task [6].

In this paper, we explored knowledge distillation for the RNN-T [7] model. RNN-T is widely used in large-scale ASR systems [8, 9] and achieves state-of-the-art results on the LibriSpeech dataset [10, 11, 12]. NST training of RNN-T models was first studied in [6] using hard target distillation [4, 13], where the student model is trained using pseudo labels generated by a teacher model. Hard target distillation was also used in the follow-up works [10, 12] that further improved the SoTA results on LibriSpeech by combining with pre-training. More recently, soft target distillation for RNN-T was explored in [14, 15], where the KL divergence between the teacher and student output label distribution is used as the loss function, similar to those used in [4]. However, it was only used for model compression [14] and streaming ASR models [15].

Motivated by the success of using soft target distillation in image domain [5] and a recent theoretical analysis [16] claiming that soft target distillation is better than hard target, this paper investigates the optimal knowledge distillation method for the large-scale RNN-T architecture.

We demonstrate that soft target distillation brings better word error rate (WER) in self-training whose teacher and student have the same architecture. Otherwise, hard target distillation brings better WER. Teacher self-training with soft target distillation makes new LibriSpeech SoTA WER on a similar setup to W2v-BERT paper [12]; WERs (dev/dev-other/test/test-other) of Conformer 600M model without external language model from 1.3/2.6/1.4/2.7 to 1.3/2.4/1.4/2.6. In addition, we succeeded to train new production teacher without performance degradation by soft distillation, in the case that training data distribution is shifted.

The contributions of this work include the following: (1) A systematic study of soft and hard target distillation on large-scale SoTA RNN-T. (2) A practical solution of soft/hard target distillation given different situation. (3) A more efficient way to do soft distillation and achieve new SoTA on LibriSpeech.

2. RELATED WORK

2.1. RNN-T model

In this section, we briefly summarize RNN Transducer [7] (RNN-T). The RNN-T loss is given by the sum of the negative log posterior probability for all possible sequences through the $U \times T$ lattice, as shown in Fig 1, where $U$ is the length of target token sequence $y$ and $T$ is the number of audio input features. Each node at $(u, t)$ in the lattice represents a log probability made of a pair of acoustic $(a_{mt})$ and label $(l_{nt})$ states. The RNN-T joint network combines these states to output the probabilities of the next label (such as characters or word pieces) token (vertical transition) or a special blank token (horizontal transition). The RNN-T loss can be computed efficiently using the forward-backward algorithm.

2.2. RNN-T distillation

The first knowledge distillation [4] paper introduces both soft target and hard target distillation in classification task. Soft target is a categorical distribution over all classes while hard target is a one-hot vector.

In RNN-T distillation, hard target is a transcript label, which is represented by a sequence of one-hot vectors. The first RNN-T NST [6] utilizes hard target distillation and then hard target distilla
3.1. Multi-stage training using self/semi-supervised learning

We combine W2v-BERT and NST as follows:
1. Prepare the existing strong teacher.
2. Pretrain RNN-T encoder of the student by W2v-BERT.
3. Distillation from the teacher to the pre-trained student.

The existing strong teacher is bi-directional model trained by W2v-BERT and multi generation NST. We call the distillation target as student. In the paper, the student model is both large bi-directional model and small streaming model.

3.2. Distillation methods

We use a linear combination of the RNN-T loss ($L_{RNNT}$) and the KL-divergence loss ($L_{KL}$) as the overall training loss:

$$L = \alpha L_{RNNT} + (1 - \alpha) L_{KL}$$

It can express different training methods by adjusting $\alpha$. With $\alpha = 1$ and human labeled data, it is a conventional RNN-T training. When pseudo-labeled data are used, we can achieve hard target distillation with $\alpha = 1$ and soft target distillation with $\alpha = 0$. Furthermore, we can mix both the hard and soft target distillation by setting $\alpha$ to be between 0 and 1. We can also mix soft target distillation with supervised training by using $L_{RNNT}$ on labeled data and $L_{KL}$ on unlabeled data, which is used in existing distillation work in other domains [4, 20, 21]. In this paper, most of the experiments use $\alpha = 0$ because we found that using $L_{KL}$ alone achieves better WERs, as shown in Section 4.3.

We implement RNN-T soft target distillation efficiently. In Eq 1, the KL divergence for each $(u,t)$ can be computed independently, so it is free to compute all at once or break them into smaller groups to balance time and space trade-off. We found that computing 8 time frames each iteration works best. With this implementation, vanilla soft target distillation and efficient RNN-T distillation [14] consume similar memory.

3.3. Enhanced Conformer

The Conformer [20] is widely used for state-of-the-art ASR systems. On the other hand, Primer [23] is the SoTA Transformer architecture for autoregressive language modeling, as discovered by AutoML. It was found to reduce the training costs by 4x and achieved better performance [23].

Motivated by the use of the Squared ReLU activation function in Primer, we change the activation function for feed forward modules in Conformer from SWISH to Squared ReLU and found it to be beneficial, as shown in the experiment results in Section 4.2.

Primer also introduces depth separable convolution for $Q, K, V$. However, we did not find it useful for our setup. In this paper, we use Conformer with Squared ReLU in all the experiments.

4. EXPERIMENT

4.1. Experiment Setup

4.1.1. Model Architecture

In this study, we evaluate three different types of models: the bi-directional large model, and the streaming large and small model. The bi-directional large model is utilized as the teacher model throughout the paper. In Sec 4.3.1 to 4.3.3, we present the results of our experiments when the student model is either the bi-directional large model, or the streaming large/small model, respectively.

The architecture of the bi-directional large model is similar to Conformer XL model (0.6B) in [10]. The audio encoder has 24 Conformer blocks [22] with model dimension 1024. The convolution kernel size is 5 and the self-attention layer consists of 8 heads with 128 left and right context length. The decoder consists of a 2-layer LSTM label encoder with 2048 units projected down to 640 output units, and a joint network with a single feed-forward layer with 640 units. The total number of weights is 0.6B. The model is trained by minimizing the RNN-T loss [7].

The streaming large model architecture is same to the bi-directional model, but self-attention and convolution layer use a
causal mask. The streaming small model is a smaller by depth and width reduction. The Conformer encoder has 17 blocks with 512 dimension. This model has a total of number 150M weights.

4.1.2. Data

We conduct our experiments on both the public data and in-house data. First, public data consists of LibriSpeech and LibriLight dataset as shown in Tab 1. LL is the main target of RNN-T distillation experiments as LL does not have labels. The teacher model produces pseudo labels for hard target distillation. The teacher model produces RNN-T $U \times T$ log probability lattice for soft distillation.

Table 1. Overview of LibriSpeech datasets.

| Data set       | Label | Hours |
|----------------|-------|-------|
| LibriSpeech (LS) | human | 960   |
| LibriLight (LL)  | unlabel | 60k   |

Second, we use a large multi-domain (MD) English dataset [24] as in-house dataset. The data consists of utterances from multiple domains, such as voice search (VS), medium-form (MF) and YouTube (YT). All the data are anonymized and hand-transcribed following Google AI principles [25]. VS is mostly voice command. MF is mostly natural conversation. The YT labels are generated from YouTube video transcription with confidence filter.

Table 2. Overview of in-house datasets. MD denotes all the supervised data (VS + MF + YT, 575k hrs). MDnew denotes the new data (VSunsup + MFunsup + YT, 615k hrs).

| Data set       | Label | Hours |
|----------------|-------|-------|
| Voice search (VS) | human | 27k   |
| Medium-form (MF)  | human | 26k   |
| Youtube (YT)     | semi  | 440k  |
| Voice search (VSunsup) | unlabel | 150k |
| Medium-form (MFunsup) | unlabel | 25k  |

4.2. Prepare the baseline teacher

We enhance RNN-T conformer model using Squared ReLU activation as proposed in 3.3. It makes clear improvement compared to Conformer XL model of W2v-BERT paper [12]. As shown in Tab 3, Squared ReLU experiment has better WERs than W2v-BERT XL (600M parameters) model [12], which uses SWISH activation. B0 is trained with the same methodology as W2v-BERT XL model [12], such as W2v-BERT pre-train and then NST with hard target pseudo labels. B0 goes through multiple NST generations until WERs are converged, which is 4 generations. The rest of experiments use Conformer with Squared ReLU activation, and B0 as the teacher.

Table 3. LibriSpeech WERs comparison for the enhanced baseline.

| Model                  | dev  | dev-other | test | test-other |
|------------------------|------|-----------|------|------------|
| W2v-BERT XL [12]      | 1.3  | 2.6       | 1.4  | 2.7        |
| B0 (Squared ReLU)      | 1.3  | 2.5       | 1.4  | 2.6        |

4.3. LibriSpeech results

4.3.1. Large bi-directional self-training

We distil the baseline teacher (B0) to the same architecture student model. As shown in Tab 4, soft target distillation (E1) makes better WERs in self-training scenario. Hard target distillation (E0) has slightly worse WERs than the teacher (B0).

All LibriSpeech near-SoTA papers in Tab 4 use hard target distillation. W2v-BERT [12] has LibriSpeech SoTA WERs before. WERs of soft target distillation (E1) is better, especially dev-other and test-other. We have new SoTA WERs for 600M parameters models.

Table 4. LibriSpeech WERs comparison for teacher models, without language model (LM). XL denotes 600M model and XXL denotes 1B model.

| Model                  | dev  | dev-other | test | test-other |
|------------------------|------|-----------|------|------------|
| B0 (teacher XL)        | 1.3  | 2.5       | 1.4  | 2.6        |
| E0 (hard target, XL)   | 1.3  | 2.5       | 1.4  | 2.7        |
| E1 (soft target, XL)   | 1.3  | 2.4       | 1.4  | 2.6        |
| BEST-RQ XL [17]        | 1.5  | 2.8       | 1.6  | 2.9        |
| W2v-BERT XL [12]      | 1.3  | 2.6       | 1.4  | 2.7        |
| wav2vec2 XXL [10]      | 1.3  | 2.7       | 1.5  | 2.7        |
| W2v-BERT XXL [12]     | 1.4  | 2.4       | 1.4  | 2.5        |

4.3.2. Small bi-directional student model

As shown in Tab 5, hard target distillation makes better WERs than others, for 150M parameters student model. Soft target distillation still makes better WERs than RNN-T baseline (B1). Soft target distillation does not work well when the architecture of teacher and student is different. On the other hand, hard target is architecture agnostic as hard target is just pseudo label.

E4 combines hard target and soft target distillation by setting $\alpha = 0.5$ in Eq 2. The results are as bad as soft target distillation (E3). Soft target distillation loss (i.e. the KL term in Eq 2) with mismatching teacher and student logits affects the training badly, although the loss includes RNN-T loss. We assume there is an alignment mismatch between RNN-T loss and KL divergence loss, which affects the training badly.

Table 5. LibriSpeech WERs comparison for 150M student models.

| Model                  | dev  | dev-other | test | test-other |
|------------------------|------|-----------|------|------------|
| B0 (teacher XL)        | 1.3  | 2.5       | 1.4  | 2.6        |
| B1 (student)           | 1.9  | 4.4       | 2.1  | 4.6        |
| E2 (hard target)       | 1.5  | 3.6       | 1.7  | 3.6        |
| E3 (soft target)       | 1.7  | 3.7       | 1.8  | 3.8        |
| E4 (hard+soft target)  | 1.7  | 3.7       | 1.8  | 3.8        |

4.3.3. Small streaming student model

Soft target distillation is even worse with a streaming student because the architecture is more different. As shown in Tab 6, hard
target distillation (E5) makes the best WERs for 150M parameters streaming student model. Soft target distillation (E6) makes even worse WERs than RNN-T baseline (B2). E7 combines hard target and soft target distillation, which has the worse WERs among all the streaming experiments. Soft target distillation loss (i.e. KL divergence in Eq 2) with mismatching teacher and student logits fights against hard target distillation loss (i.e. RNN-T loss in Eq 2), which results in bad WERs.

Table 6. LibriSpeech WERs comparison for streaming 150M student models.

| Model          | Data     | WER | dev | dev-other | test | test-other |
|----------------|----------|-----|-----|-----------|------|------------|
| B0 (teacher XL)| dev      | 1.3 | 2.5 | 1.4       | 2.6  |            |
| B2 (student)  |          | 4.2 | 10.4| 4.8       | 9.5  |            |
| E5 (hard target)|         | 4.0 | 9.4 | 4.4       | 8.6  |            |
| E6 (soft target)|        | 4.4 | 11.1| 4.6       | 11.1 |            |
| E7 (hard+soft target)| | 4.6 | 11.8| 4.7       | 11.7 |            |

4.4. Production data experiments

We compare hard target and soft target distillation with our in-house data as mentioned in Sec 4.1.2. We conduct distillation experiments with the existing strong teacher from the previous work [9]. The existing teacher is B3 in Tab 7 which is trained by W2v-BERT and multi generations NST with MD data as shown in Tab 2. The model architecture is very similar to LibriSpeech Conformer XL teacher. The biggest difference is that the model input feature is 4 contiguous frames of 128-dimension log-mel features [3] sub-sampled by a factor of 3 and a one-hot domain-id vector of size 16.

First, we conduct both hard and soft target distillation on the same architecture student model with the same MD data. As shown in the second group of Tab 7, soft target distillation (E9) maintains the WERs of the teacher (B3), but hard target distillation (E8) degrades the performance. B3 from the previous work [9] uses hard target distillation but has better WERs than E8, because it mixes pseudo labels with 1k hours human labels. Soft target distillation successfully maintains teacher WERs without any human labels.

Second, we conduct the distillation experiments with newer and bigger data (MD_{new}). As shown in the third group of Tab 7, soft target distillation (E11) improves WERs, while hard target distillation (E10) degrades WERs. Soft target distillation is more robust in data domain shift scenario. It enables the teacher to adapt new data without human label. It allows our production teacher to adapt new data domain in continual learning.

5. DISCUSSION

5.1. Soft target vs hard target

As shown in Sec 4.3 and Sec 4.4, soft target distillation works better in self-training, but hard target distillation works better when the teacher and student have different architecture. Hard targets allow the student model to learn the alignment by itself using RNN-T loss and therefore works well when the teacher and student have different architecture such as bi-directional teacher and streaming student. On the other hand, soft distillation computes node-wise KL divergence over the $U \times T$ lattice, so it assumes that both the teacher and student models have similar optimal alignment. This is true when the teacher and student have the same architecture. When the alignment is not an issue, soft targets can transfer more information. We leave alignment aware soft target distillation for a future study.

5.2. Self-sup pretrain for student

So far, we pretrain the student using W2v-BERT before distillation. When the student is randomly initialized, soft target distillation (E13) could not make the optimal WERs, as shown in Tab 8. Self-supervised learning and soft target distillation are complementary methods. We recommend to pretrain audio encoder of the student model by the latest self-supervised learning before teacher student distillation.

Table 7. In-house data WERs comparison for production 600M teacher models. E8 and E9 are trained with the same data to the teacher B3. E10 and E11 are trained with out of domain data. Soft distillation (E11) recovers in-domain VS and MF WERs without in-domain training.

| Model          | Data     | WER (%) | VS | MF | YT |
|----------------|----------|---------|----|----|----|
| B3 (old teacher)| MD       | 4.1     | 4.3| 8.0|
| E8 (hard target)| MD       | 4.2     | 4.4| 8.1|
| E9 (soft target)| MD       | 4.1     | 4.3| 8.0|
| E10 (hard target)| MD_{new}| 4.3     | 4.4| 7.9|
| E11 (soft target)| MD_{new} | 4.1     | 4.3| 7.8|

5.3. Limitation

In this work, we explore optimal distillation method given the existing strong teacher. In the case that we do not have strong teacher, the suggested method may not work without modification. Self-Adaptive Distillation paper [15] reports that RNN-T loss and soft target distillation loss together works well with weak teacher. We leave making weak teacher to strong teacher by knowledge distillation for future study. As the teacher is stronger, the distillation method should converge to the suggested method in this paper.

6. CONCLUSION

In this paper, we conducted extensive empirical studies for RNN-T distillation. Our results show that soft target distillation works better when both the teacher and student models have the same architecture. Otherwise, hard target distillation works better. We further demonstrate that soft target distillation achieved better results for self-training NST, where iterative distillation is applied to the same model. This enhances a 600M Conformer RNN-T model to achieve a new SoTA WERs for LibriSpeech.
7. REFERENCES

[1] Tara N Sainath, Yanzhang He, Bo Li, Arun Narayanan, Ruoming Pang, Antoine Bruguier, Shuo-yin Chang, Wei Li, Raziel Alvarez, Zhiyeng Chen, et al., “A streaming on-device end-to-end model surpassing server-side conventional model quality and latency,” in ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020, pp. 6059–6063.

[2] Bo Li, Anmol Gulati, Jiahui Yu, Tara N Sainath, Chung-Cheng Chiu, Arun Narayanan, Shuo-Yin Chang, Ruoming Pang, Yanzhang He, James Qin, et al., “A better and faster end-to-end model for streaming asr,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 5634–5638.

[3] Arun Narayanan, Rohit Prabhavalkar, Chung-Cheng Chiu, David Rybach, Tara N Sainath, and Trevor Strohman, “Recognizing long-form speech using streaming end-to-end models,” in IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). IEEE, 2019, pp. 920–927.

[4] Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al., “Distilling the knowledge in a neural network,” arXiv preprint arXiv:1503.02531, vol. 2, no. 7, 2015.

[5] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le, “Self-training with noisy student improves imagenet classification,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10687–10698.

[6] Daniel S Park, Yu Zhang, Ye Jia, Wei Han, Chung-Cheng Chiu, Bo Li, Yonghui Wu, and Quoc V Le, “Improved noisy student training for automatic speech recognition,” arXiv preprint arXiv:2005.09629, 2020.

[7] Alex Graves, “Sequence transduction with recurrent neural networks,” in International Conference on Machine Learning: Representation Learning Workshop. 2012.

[8] Yu Zhang, Daniel S Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, et al., “Bigsls: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition,” IEEE Journal of Selected Topics in Signal Processing, 2022.

[9] Dongseong Hwang, Khe Chai Sim, Zhouyuan Huo, and Trevor Strohman, “Pseudo label is better than human label,” in INTERSPEECH 2022, 2022.

[10] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu, “Pushing the limits of semi-supervised learning for automatic speech recognition,” arXiv preprint arXiv:2010.10504, 2020.

[11] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli, “Self-training and pre-training are complementary for speech recognition,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 3030–3034.

[12] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu, “W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training,” arXiv preprint arXiv:2108.06209, 2021.

[13] Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao, “Knowledge distillation: A survey,” International Journal of Computer Vision, vol. 129, no. 6, pp. 1789–1819, 2021.

[14] Sankaran Panchapagesan, Daniel S Park, Chung-Cheng Chiu, Yuan Shangguan, Qiao Liang, and Alexander Grusenstein, “Efficient knowledge distillation for rnn-transducer models,” in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 5639–5643.

[15] Isabel Leal, Neeraj Gaur, Parisa Haghani, Brian Farris, Pedro J Moreno, Manasa Prasad, Bhuvana Ramabhadran, and Yun Zhu, “Self-adaptive distillation for multilingual speech recognition: Leveraging student independence,” in Interspeech, 2021, pp. 2556–2560.

[16] Aditya Krishna Menon, Ankit Singh Rawat, Sashank J Reddi, Seunyeon Kim, and Sanjiv Kumar, “Why distillation helps: a statistical perspective,” arXiv e-prints, pp. arXiv--2005, 2020.

[17] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu, “Self-supervised learning with random-projection quantizer for speech recognition,” arXiv preprint arXiv:2202.01855, 2022.

[18] Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli, “wav2vec: Unsupervised pre-training for speech recognition,” arXiv preprint arXiv:1904.05862, 2019.

[19] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli, “wav2vec 2.0: A framework for self-supervised learning of speech representations,” Advances in Neural Information Processing Systems, vol. 33, pp. 12449–12460, 2020.

[20] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le, “Unsupervised data augmentation for consistency training,” Advances in Neural Information Processing Systems, vol. 33, pp. 6256–6268, 2020.

[21] Tal Ridnik, Hussam Lawen, Emanuel Ben-Baruch, and Asaf Noy, “Solving imagenet: a unified scheme for training any backbone to top results,” arXiv preprint arXiv:2204.03475, 2022.

[22] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang, “Conformer: Convolution-augmented transformer for speech recognition,” in Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020, Helen Meng, Bo Xu, and Thomas Fang Zheng, Eds. 2020, pp. 5036–5040, ISCA.

[23] David So, Wojciech Mafié, Hanxiao Liu, Zihang Dai, Noam Shazeer, and Quoc V Le, “Searching for efficient transformers for language modeling,” Advances in Neural Information Processing Systems, vol. 34, pp. 6010–6022, 2021.

[24] A. Narayanan, A. Misra, K. C. Sim, G. Pandak, A. Tripathi, M. Elfeky, P. Haghani, T. Strohman, and M. Bacchiani, “Towards domain-invariant speech recognition via large scale training,” in 2018 IEEE Spoken Language Technology Workshop (SLT), 2018, pp. 441–447.

[25] “Artificial intelligence at google: Our principles,” https://ai.google/principles/.