INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Scaling Effect of Self-Supervised Speech Models

Jie Pu[1,2], *Yuguang Yang*[2,*], *Ruirui Li*[2,*], *Oguz Elibol*[2], *Jasha Droppo*[2]

[1]Department of Engineering, University of Cambridge, UK
[2]Amazon Alexa, California, USA

jp936@cam.ac.uk, {yuguay, ruirul, oelibol, drojasha}@amazon.com

Abstract

The success of modern deep learning systems is built on two cornerstones, massive amount of annotated training data and advanced computational infrastructure to support large-scale computation. In recent years, the model size of state-of-the-art deep learning systems has rapidly increased and sometimes reached to billions of parameters. Herein we take a close look into this phenomenon and present an empirical study on the scaling effect of model size for self-supervised speech models. In particular, we investigate the quantitative relationship between the model size and the loss/accuracy performance on speech tasks. First, the power-law scaling property between the number of parameters and the L_1 self-supervised loss is verified for speech models. Then the advantage of large speech models in learning effective speech representations is demonstrated in two downstream tasks: i) speaker recognition and ii) phoneme classification. Moreover, it has been shown that the model size of self-supervised speech networks is able to compensate the lack of annotation when there is insufficient training data.

Index Terms: model size, self-supervised learning, power-law scaling, speaker recognition

1. Introduction

With the advance of deep learning, the performance of state-of-the-art models is stably improving in recent years, along with a rapid increase of their model sizes. Such a trend has been observed in a wide range of machine learning tasks, e.g., computer vision [1, 2], audio analysis [3] and natural language processing [4, 5]. In particular, the GPT-3 model [6] contains 175 billions parameters in total and costs OpenAI more than four-million dollars to train it. Given such a huge model size, GPT-3 is shown to achieve strong performance on many NLP tasks without any fine-tuning on annotated data. Moreover, people tend to train even larger neural network models to push the performance limit with the support of more powerful hardware. A natural question arisen here is, what are the indispensable benefits of these large neural networks that are worthwhile for those invested resources?

Intuitively, it is easy to notice a positive correlation between the model size and its performance, however there are important questions at a deeper level. For instance, does there exist a quantitative relationship, instead of qualitative one, between the size of neural networks and its performance? What would be the high ceiling and low floor for performance when the model size gets scaled up/down? Will the scaling property of different model sizes vary when applied to different downstream tasks? In [7], Jared Kaplan et al. answered some of these questions and proposed that there is an empirical quantitative power-law scaling property between the model size and its cross-entropy loss. This scaling law provides a definitive guide on scaling up models to meet performance requirement under resource constraints. Nevertheless, the discovery is limited in the sense that it only applies to language models and the cross-entropy loss. In this paper, we would like to build upon Jared's work, and investigate the scaling property of speech models, as well as the model performance, i.e. accuracies, in downstream tasks.

To this end, we proposed to use self-supervised speech models with different model sizes, i.e., number of parameters, and examine their performance on both the self-supervised loss and accuracy in speech tasks. The reason to use self-supervised models is twofold: i) it has a self-supervised pre-train loss that resembles to the cross-entropy loss in [7], which allows us to verify the power-law scaling property, ii) unlike supervised speech models trained on one particular classification task [8] [9], the self-supervised speech model is able to generate representations that can be used on several downstream tasks. This allows us to investigate its scaling property across different speech tasks.

Among recent self-supervised speech studies, Mockingjay [10], TERA [11], APC [12] and DeCoAR [13] learn speaker embedding representations based on self-supervised spectrogram reconstructions or predictions. In particular, we choose Mockingjay [10] to serve as the base model in this work. Different from other self-supervised representation learning methods in speech [14, 15, 16], which learn speech representations by predicting future frames based on past frames, Mockingjay alleviates the uni-directionality constraint and is designed to predict the current frame through jointly conditioning on both past and future contexts. By doing so, Mockingjay is shown to provide the state-of-the-art performance for self-supervised speech representation learning. We will scale up/down the model size of the Mockingjay Transformer, and then record its corresponding performance on the L_1 loss and the accuracy in downstream tasks, where the L_1 loss here measures the discrepancies between the predicted frames and original frames. Herein we have studied two downstream tasks, i.e., speaker recognition and phoneme classification.

In our experiments, the quantitative relationship between the model size of self-supervised speech models and their performance is empirically investigated, where the number of parameters cover over five orders of magnitude in scale. The key findings are summarized as following:

1. Similar to language models, power-law scaling property between the self-supervised L_1 loss and the model size largely holds for speech models disregarding variations in model architecture, i.e., width vs. depth.

2. As the model size varies, there exists positive correlation between increasing model size and improving accuracy in downstream tasks.

3. When annotated data is scarce, larger pretrained speech models require fewer training data to fine-tune for downstream tasks.

* Equal contribution.

http://dx.doi.org/10.21437/Interspeech.2021-1935

2. Proposed Methodology

In this section, the power-law scaling property that we aim to verify is first introduced. Then the self-supervised speech model is presented along with a scaling scheme to increase or decrease its model size.

2.1. Power-law scaling property

In [7], Jared Kaplan et al. introduced the power-law relationship between the number of parameters in autoregressive language models and their test loss. Mathematically, this power-law relationship can be defined as follows:

$$L(N) = (N_c/N)^{(\alpha_N)} \tag{1}$$

where L is the test loss and N is the number of parameters. α_N and N_c are constants for language modelling, where $\alpha_N \sim 0.076$ and $N_c \sim 8.8 \times 10^{13}$. The power law α_N specifies the degree of performance improvement expected as we scale up N. Besides, the precise numerical values of α_N and N_c depend on the language training dataset and hence do not have a fundamental meaning.

The power-law scaling property is important as it provides a theoretical framework about the dependence of model performance on N. We can use this framework to conduct predictions, and obtain insights on controlling the compute scaling, magnitude of over-fitting, early stopping step, and data requirements when training deep learning models. In the ideal case, one might interpret the relation as analogues of laws of physics, e.g., the ideal gas law that relates the macroscopic properties of a gas in a universal and deterministic way. Therefore, as a critical step to understand the relationship between the model size and its performance, we would like to verify the power-law scaling property in the speech domain.

2.2. Self-supervised learning: Mockingjay

In this paper, we use the bidirectional Transformer as a basic model, whose model size will be scaled up/down in later experiments. To train the transformer, we use a self-supervised representation learning approach, Mockingjay [10].

Mockingjay aims to achieve self-supervised pre-training for speech representations, by learning to reconstruct masked audio frames. Specifically, at the training time 15% of the input audio frames are randomly masked to either zero or another frame, then the model learns to reconstruct and predict the original frames, based on its left and right context. The L_1 loss is used to minimize reconstruction error between prediction and ground-truth frames on both the selected 15% and untouched 85% frames. After self-supervised training, the output of the last layer in the Transformer encoder serves as speech representations, which will be used for downstream tasks.

On the whole, there are two training stages in order to perform speech tasks. First, the Transformer is pre-trained in the self-supervised manner (Mockingjay) on a large amount of unlabeled speech data. The metric to evaluate this self-supervised pre-training is the L_1 loss between reconstructed and original audio frames. At the second stage, the output from the last layer of the Transformer is used as the feature representation of speech, and then feed into a simple one or two-layer classifier to perform downstream tasks. The weights of Transformer network will not be updated at this stage, while the classifier network will be trained and tested with the accuracy on downstream tasks as its metric. Specifically, we have two downstream tasks, i.e., speaker recognition and phoneme classification. These two tasks strive to learn different speech signals, which are complementary to each other in the sense that speaker recognition focuses on extracting speaker biometrics, while the phoneme classification focuses on identifying speech content.

2.3. Scaling scheme for Transformer

To investigate the relationship between the model size and its performance, a scaling scheme for Mockingjay Transformer is needed. In general, we want to vary the hyper-parameters that control the model size of Transformer. Let L_{num} denote the number of layers, H_{dim} the size of hidden dimension and F_{dim} the size of feed-forward layer. The total number of parameters N can be approximated as follows [7]:

$$\begin{aligned} N &\approx 2H_{dim}L_{num}(2H_{dim} + F_{dim}) \\ &= 12L_{num}H_{dim}^2 \;\; when \;\; F_{dim} = 4H_{dim} \end{aligned} \tag{2}$$

It is easy to see that we can control the model size by varying the values of L_{num}, H_{dim} and fixing the relationship between H_{dim} and F_{dim}. In particular, all possible values of L_{num} in our experiments are defined in the set $L_{num} = \{1, 2, 3, 4, 5, 6\}$. Similarly, for the size of hidden dimension $H_{dim} = \{12, 24, 48, 96, 192, 384, 768\}$. Then the model size of the Transformer is defined by the different combination of L_{num} and H_{dim}.

The minimal model size in our experiments would be $1728 \approx 1.7 \times 10^3$ when $L_{num} = 1$ and $H_{dim} = 12$, while the maximal size $42.47M \approx 4.2 \times 10^7$ when $L_{num} = 6$ and $H_{dim} = 768$. Therefore, the number of parameters in the Transformer covers over five orders of magnitude in scale, which is reasonably adequate to verify the power-law scaling property in speech tasks. It is worth noting that the magnitude of 10^7 is still relatively small compared to top-performing deep learning systems, e.g, wav2vec 2.0 [3] with 3.17×10^8 parameters trained on 128 GPUs and the Conformer [17] has 10^9 parameters and trained on TPUs [18].

3. Experimental Evaluation

This section provides a thorough experimental evaluation of speech models with different sizes. Three sets of experiments are conducted, which are summarized as follows:

- **Power law on L_1 loss**. The power-law scaling property between the model size and its loss, i.e, the L_1 loss in Mockingjay, is first assessed. It is important to verify whether this quantitative relationship holds in speech models and why it would or not be the case. A log-log plot between the model size and the L_1 loss will be generated for a visualized evaluation, where the power-law property is equivalent to a straight line in the plot.
- **Downstream tasks**. In order to evaluate the quality of representation learned by self-supervised speech models for practical applications, we conduct two downstream tasks for the models with different sizes. In particular, speaker recognition and phoneme classification tasks are conducted, where the focus is the relationship between the accuracy performance and the model size.
- **Few annotated data**. The scaling property of the model size would be examined when annotated training data is few. With only 3.6 hours (1%) of transcribed speech data available, the quantitative relationship between model accuracy and its size at the phoneme classification task has been revisited.

The self-supervised speech models are first pre-trained on the LibriSpeech corpus [19] subset, train-clean-360. The Adam optimizer [20] is used where learning rate is warmed up over the

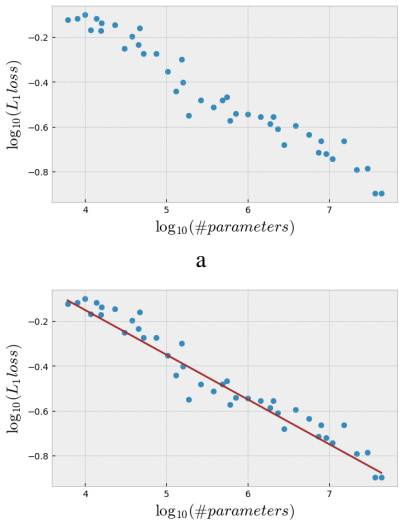

Figure 1: *Scaling results of the L_1 loss. The x-axis is for $\log_{10}$ (#parameters) while y-axis represents $\log_{10}$ (L_1 loss).*

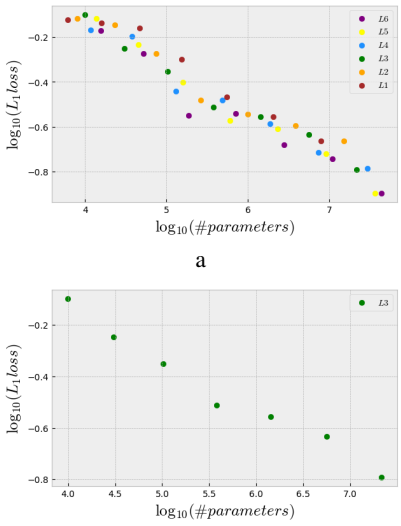

Figure 2: *Scaling results of the L_1 loss. (a) Different color represents the Transformer model with different numbers of layers, L_{num}. (b) The case of $L_{num} = 3$.*

first 7% of 500k total training steps to a peak value at 4e-4 and then linearly decayed. We have applied a dropout [21] of 0.1 on all layers and attention weights. For downstream tasks, most of the hyperparameters are the same as in pre-training, with the exception of a learning rate of 4e-3. The model is trained with a batch size of 6 using one GPU.

3.1. Power law on L_1 loss

Herein we first investigate the quantitative relationship between the model size and the L_1 loss at self-supervised pre-training stage. In Mockingjay, the L_1 loss represents the reconstruction error between prediction and ground-truth audio frames, which shows how well the Transformer learns the speech representation. For the model size, the total number of parameters in the Transformer covers over five orders of magnitude in scale, ranging from 10^3 to 10^7. As described in section 2.3, the precise number of parameters is defined by different combination of L_{num} and H_{dim}, where there are 42 different combinations thus 42 different model sizes in total.

Training these 42 Transformer models with different sizes following section 2.2, we will obtain 42 values for their L_1 loss. Then to verify the power-law scaling property, a log-log plot between the model size (as x-axis) and the L_1 loss (as y-axis) will be created. According to the Formula 1, a well-fit straight line is expected in the log-log plot if the power law holds [7]. Figure 1-a shows our results on the L_1 loss. As we can see, the general trend obeys a linear relationship thus we can easily fit a straight line in Figure 1-b. In particular, the line is $y = -0.2x + 0.65$, which gives the power relationship that $L(N) = (1778.28/N)^{(0.2)}$. Figure 2 explores in more details, where different color represents the Transformer with different values of L_{num} and Figure 2-b shows the case of $L_{num} = 3$.

Although the power-law scaling property holds for the L_1 loss in a general trend, there are some fluctuations in Figure 1 and Figure 2 which prevent points lie exactly at the straight line. The reason is twofold: 1) The intrinsic randomness within the deep learning optimization process (stochastic gradient descent) and the random audio masking process in Mockingjay. 2) A further improvement on the hyper-parameters is possible while fixing the model size. In particular, we suspect that a smoother scaling law relationship could be derived when using optimal aspect ratio as it did in [7]. To illustrate this, Figure 3 shows the relationship between the L_1 loss, the model size and its aspect ratio, i.e., H_{dim}/L_{num}. As we can see from the top-right corner of the figure, smaller values of L_1 loss may be obtained by reducing the aspect ratio for large models. In that sense, the fluctuation in Figure 1 and Figure 2 is probably attributed to the non-optimality of the pre-defined aspect ratios in section 2.3.

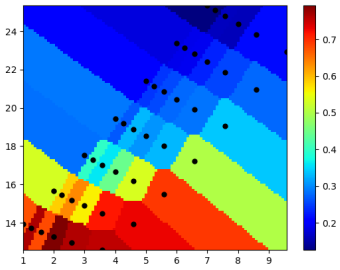

Figure 3: *Relationship between the L_1 loss (color), the aspect ratio in the x-axis, $\log_2(H_{dim}/L_{num})$ and the model size in the y-axis, $\log_2(\#parameters)$.*

3.2. Downstream tasks

With the self-supervised Transformer models from section 3.1, we extract outputs from the last layer of Transformer models as the feature representation, and feed into a one or two-layer classifier to perform downstream tasks. The quantitative relationship between the model size and the accuracy in downstream tasks is the focus of this section.

First, we conduct the speaker recognition task on the LibriSpeech 100 hours subset. A simple one-layer RNN classifier for speaker recognition is trained, receiving the feature representation from the Transformer models with different sizes. Re-

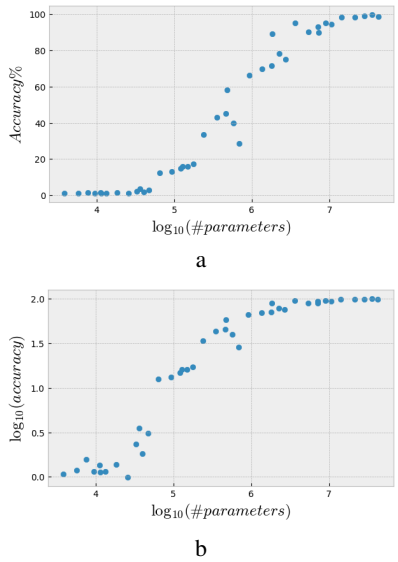

Figure 4: *Scaling results of the speaker recognition task. The x-axis is for* $\log_{10}$ *(#parameters) and y-axis represents (a) Accuracy and (b)* $\log_{10}$ *(Accuracy).*

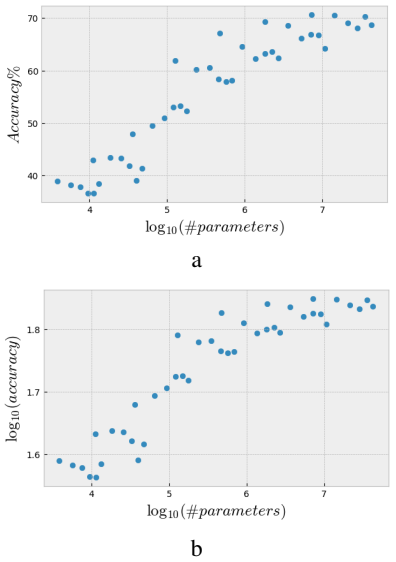

Figure 5: *Scaling results of the phoneme classification task. The x-axis is for* $\log_{10}$ *(#parameters) and y-axis represents (a) Accuracy and (b)* $\log_{10}$ *(Accuracy).*

sults are reported in Figure 4-a, where the x-axis represents the log value of the number of parameters and the y-axis is the accuracy performance. It is easy to notice the positive correlation between the model size and its performance in accuracy. The log-log plot is in Figure 4-b, where the accuracy value seems to approach saturation when the model size increases, instead of exhibit the power-law relationship.

To further examine the relationship between the model size and its accuracy, we have conducted another downstream task, phoneme classification. A two-layer phoneme classifier is trained using the LibriSpeech train-clean-360 subset. Phoneme sequences are first aligned using the Montreal Forced Aligner [22]. Results on the LibriSpeech test-clean subset are presented in Figure 5-a. The log-log plot for the phoneme classification is in Figure 5-b. Similar to the speaker recognition task, the positive correlation between the model size and its accuracy is obvious, but the top performance tends to get saturated in large models.

With the results in Figure 4 and Figure 5, there is a positive correlation between the model size and its performance in downstream tasks, but not exactly the power-law relationship. We have one hypothesis on this. When the model size gets constantly increased, the size of training data becomes the bottleneck to saturate its performance. Further work could verify the hypothesis and explore the case when training data is sufficiently enough.

3.3. Few annotated data

Apart from the power-law scaling property between the model size and its loss, [7] shows that large models are more sample-efficient than small models, reaching the same level of performance with fewer training data. To verify this, we design experiments at the phoneme classification task, where only 3.6 hours (1%) of transcribed speech data is available.

Figure 6 shows the results of the model accuracy using only 1% of training data, compared to the case of using 360 hours (100%) training data. As we can see, large speech models still perform reasonably well (with the accuracy around 60%) when the annotated data is insufficient. It demonstrates a clear advantage over smaller models since large models can use training data more efficiently. Compared to the case of 360-hour training data, the performance drop of the model with the same size is $2 \sim 10\%$, which is relatively small given the reduction of training data is 99%. Therefore, the large self-supervised speech models provide a good option to compensate the lack of annotated data.

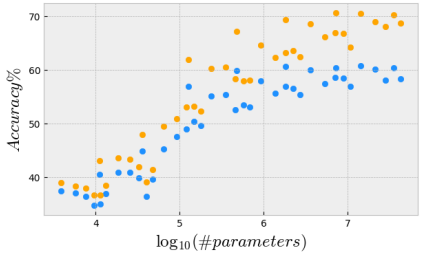

Figure 6: *Scaling results when few annotated training data is available. The x-axis is for* $\log_{10}$ *(#parameters) and y-axis represents the accuracy (%). Orange points show the results for 360-hour training data while blue points for 3.6 hours (1%).*

4. Conclusion

In this paper, we investigate how the model size of self-supervised speech networks influence their performance, i.e., the L_1 loss and accuracy on downstream tasks. Given our experimental results, the power-law scaling property of the L_1 loss is first verified and largely holds for speech models. The relationship between the model size and its accuracy is positively correlated, but no exactly obey the power law. Finally, we demonstrate that large self-supervised speech models can be more data efficient than small models when annotated training data is scarce.

5. References

[1] A. Kolesnikov, L. Beyer, X. Zhai, J. Puigcerver, J. Yung, S. Gelly, and N. Houlsby, "Big transfer (bit): General visual representation learning," in *European Conference on Computer Vision (ECCV)*, vol. 12350. Springer, 2020, pp. 491–507.

[2] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. Van Der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 181–196.

[3] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[4] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, pp. 5753–5763, 2019.

[5] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," *arXiv preprint arXiv:1909.08053*, 2019.

[6] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[7] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[8] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[9] W. Xie, A. Nagrani, J. S. Chung, and A. Zisserman, "Utterance-level aggregation for speaker recognition in the wild," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5791–5795.

[10] A. T. Liu, S.-w. Yang, P.-H. Chi, P.-c. Hsu, and H.-y. Lee, "Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6419–6423.

[11] A. T. Liu, S. Li, and H. Lee, "TERA: self-supervised learning of transformer encoder representation for speech," *CoRR*, vol. abs/2007.06028, 2020.

[12] Y. Chung, W. Hsu, H. Tang, and J. R. Glass, "An unsupervised autoregressive model for speech representation learning," in *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, G. Kubin and Z. Kacic, Eds. ISCA, 2019, pp. 146–150.

[13] S. Ling, Y. Liu, J. Salazar, and K. Kirchhoff, "Deep contextualized acoustic representations for semi-supervised speech recognition," in *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*. IEEE, 2020, pp. 6429–6433.

[14] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[15] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. R. Glass, "An unsupervised autoregressive model for speech representation learning," in *INTERSPEECH*, 2019.

[16] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition." in *INTERSPEECH*, 2019.

[17] Y. Zhang, J. Qin, D. S. Park, W. Han, C.-C. Chiu, R. Pang, Q. V. Le, and Y. Wu, "Pushing the limits of semi-supervised learning for automatic speech recognition," *arXiv preprint arXiv:2010.10504*, 2020.

[18] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, R. Bajwa, S. Bates, S. Bhatia, N. Boden, A. Borchers *et al.*, "In-datacenter performance analysis of a tensor processing unit," in *Proceedings of the 44th annual international symposium on computer architecture*, 2017, pp. 1–12.

[19] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations (ICLR)*, 2015.

[21] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[22] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal forced aligner: Trainable text-speech alignment using kaldi," in *Interspeech*, 2017, pp. 498–502.