Scaling Laws for Discriminative Speech Recognition Rescoring Models

Yile Gu, Prashanth Gurunath Shivakumar, Jari Kolehmainen, Ankur Gandhe, Ariya Rastrow, and Ivan Bulyko

Amazon Alexa

{yilegu,psshvak,jkolehm,aggandhe,arastrow,ibbulyko}@amazon.com

Abstract

Recent studies have found that model performance has a smooth power-law relationship, or scaling laws, with training data and model size, for a wide range of problems. These scaling laws allow one to choose nearly optimal data and model sizes. We study whether this scaling property is also applicable to second-pass rescoring, which is an important component of speech recognition systems. We focus on RescoreBERT as the rescoring model, which uses a pre-trained Transformer-based architecture fined tuned with an ASR discriminative loss. Using such a rescoring model, we show that the word error rate (WER) follows a scaling law for over two orders of magnitude as training data and model size increase. In addition, it is found that a pre-trained model would require less data than a randomly initialized model of the same size, representing effective data transferred from pre-training step. This effective data transferred is found to also follow a scaling law with the data and model size.

Index Terms: automatic speech recognition, rescoring, pre-training, scaling laws.

1. Introduction

State-of-the-art automatic speech recognition systems (ASR) perform second-pass rescoring in which the n-best hypotheses, generated by the first-pass, are reranked to improve accuracy [1, 2, 3, 4, 5]. The need for second-pass rescoring stems from the architectural constraints of running a low-latency and streaming first-pass. To ensure improved performance and better WER, minimum WER (MWER) loss is typically applied [6, 3, 5, 7] when training a second-pass rescoring model.

Research has found that model performance has predictable and favorable scaling properties with respect to training data size, model size, and compute across a variety of modalities, including language [8, 9, 10], vision [11, 12], and acoustics [13]. These scaling laws not only provide supportive evidence in favor of "large models", but they also provide researchers with the ability to determine the suitable configuration based on training data, model size, and computation.

In spite of some recent studies showing that increasing model size results in better ASR rescoring performance [14, 15, 16, 17], there has not yet been a systematic study on the scaling properties of second-pass rescoring with data and model size. Understanding the scaling properties of adopting optimal data and model size would be critical, as (a) acquiring the transcribed data required for discriminative training is costly, and (b) when deployed in production, larger models require more expensive hardware to maintain low latency.

Additionally, pre-training's impact on ASR rescoring performance is not well understood. In spite of their success at reducing the amount of annotated data needed for downstream language understanding tasks, pre-trained language models, such as BERT [18], have not been widely applied to second-pass ASR rescoring. Recent studies have examined how pre-trained models can improve ASR rescoring [19, 20, 21, 16, 22, 23, 24], but there is no systematic study of how pre-training affects rescoring performance for different models and data sizes, and whether a scaling law, proposed in a previous work [25], can capture the effect.

In this work, we fill the gaps in the literature regarding rescoring models employed in speech recognition systems, and systematically study the scaling properties for both randomly initialized and pre-trained discriminative rescoring models. We use the recently proposed RescoreBERT [16] rescoring model, which is based on a BERT architecture pre-trained on large corpora. The RescoreBERT model encodes the full context of the hypothesis using a bi-directional self-attention architecture. We demonstrate that WER follows a power-law relationship with training data size and model size for over two orders of magnitude of the range studied. Furthermore, we underscore the importance of pre-training for second-pass rescoring, and show that effective data transferred from pre-training allows a model to require less training data to achieve the same performance. A power-law relationship can also be used to describe the effective data transferred.

2. Experimental Setup

2.1. RescoreBERT

We use RescoreBERT [16] as the rescoring model to explore the scaling law. As illustrated in Figure 1, RescoreBERT model uses a BERT model with a feed-forward network attached at the BERT encoder classification (CLS) token embedding (CLS) to predict a single second-pass score for a given n-best hypothesis. This score is linearly interpolated with first pass score for re-ranking. The final score s_i used for re-ranking is therefore,

$$s_i = s_i^f + w \cdot s_i^s, \tag{1}$$

where s_i^f and s_i^s are the scores from first and second passes, respectively, and w is a hyper-parameter.

Following [16], we train a RescoreBERT model using a minimum WER (MWER) discriminative loss [7]. The training minimizes expected word error rate calculated on given n-best
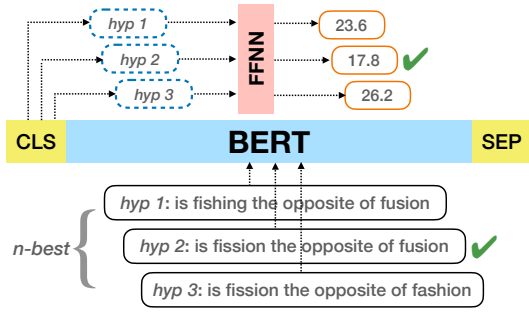
Figure 1: *Illustration of RescoreBERT. Each hypothesis is individually encoded by BERT and represented by CLS; it is then followed by a feed-forward neural network to compute a sentence-level second-pass LM score. The scores are then interpolated with first pass scores for re-ranking. The figure is reproduced from the original RescoreBERT paper [16].*

hypotheses:

$$P_i = \frac{e^{-s_i}}{\sum_{j=1}^{n} e^{-s_j}} \quad (2)$$

$$\bar{\epsilon}_H = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i \quad (3)$$

$$\mathcal{L}_{\text{MWER}} = \sum_{i=1}^{n} P_i \cdot (\epsilon_i - \bar{\epsilon}_H). \quad (4)$$

P_i is the posterior probability of a hypothesis i, and ϵ_i is the edit distance from the ground truth transcription. The MWER loss $\mathcal{L}_{\text{MWER}}$ represents the expected number of relative word errors, with $\bar{\epsilon}_H$ being the averaged word errors across the n-best list.

2.2. Experiments Performed

In the original RescoreBERT paper [16], discriminative training is performed as a fine-tuning step using pre-trained BERT models. In this study, we train both pre-trained and randomly initialized models to study the effects of pre-training.

The effective batch sizes used for training are 512, 512, 512, 2048 sets of n-best hypotheses for the 5M/17M/170M/700M models, respectively. Learning-rate decay and Adam optimizer with default parameters are used, with initial learning rates of 10^{-5} for 5M/17M/170M and 5×10^{-6} for 700M models. Training is applied with different model sizes and data sizes. We apply early stopping based on a development set, and report WER based on the test set.

2.3. Model and Data Sizes

We use four variants of BERT models with different sizes as outlined in Table 1. The models are pre-trained with MLM (Masked Language Model) objective, first with Wikipedia and mC4 [26], and then internal catalog data (up to around 1 trillion tokens in total pre-training data).

In the discriminative training phase of RescoreBERT models, we utilize internal datasets consisting of de-identified user interactions with a conversational agent in English. We use an RNN-T model [27] as the first-pass model to generate n-best hypotheses. The train, dev and test splits, in this study, contain 95300, 30 and 10 of utterances, respectively. To study the effects of training data, different fractions of training are used. We use internal data for this study, as there is no readily available public speech data at this large size to study scaling laws.

Table 1: *Summary of architecture details of BERT models.*

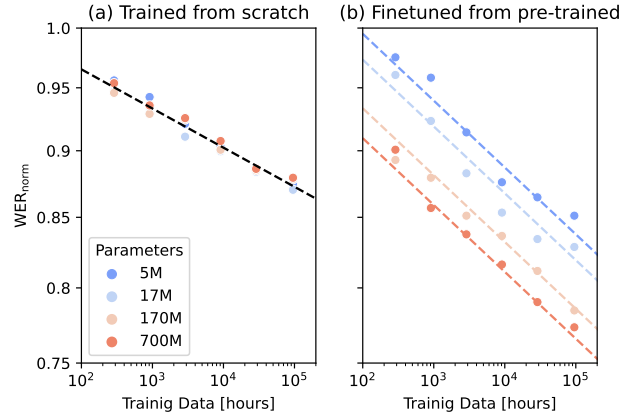| Model parameters (excluding embeddings)	5M	17M	170M	700M
Hidden Size	320	768	1024	1536
Number of Layers	4	4	16	20
Number of Attention Heads	16	16	16	16
Intermediate Layer Dimension	1200	1200	3072	6144
Embedding Parameters	49M	118M	157M	460M

Figure 2: *Normalized WER WER_{norm} (defined in Equation (5)) versus training data for (a) models trained from scratch (b) models finetuned from pre-trained models. Dashed lines denote model predictions from Equation (7) for (a) and Equation (8) for (b). Both axes are on log scale.*

However, the authors believe that the phenomena described in this paper should apply to any similar set of speech data.

3. Results and Discussion

3.1. Normalized WER (Word Error Rate)

To showcase the effectiveness of second-pass rescoring, we report the results in WER_{norm}, which is defined as

$$\text{WER}_{\text{norm}} = \frac{\text{WER}_{2P} - \text{WER}_{\text{oracle}}}{\text{WER}_{1P} - \text{WER}_{\text{oracle}}} \quad (5)$$

where WER_{1P} and WER_{2P} are WER before and after second-pass rescoring. $\text{WER}_{\text{oracle}}$ is oracle WER, which provides a lower bound for the minimum WER achievable from second-pass rescoring. For the test in this study, $\text{WER}_{\text{oracle}}$ is 66% relative lower than WER_{1P}, meaning that if WER_{norm} is 0.5, second-pass rescoring would provide 33% relative reduction in WER over first pass. WER_{norm} becomes 0 when WER from rescoring approaches oracle, and becomes 1.0 when rescoring fails to improve over first pass.

3.2. Training from Scratch

We first study the effects of training data and model sizes on WER when trained from scratch. The results can be found in Figures 2(a) and 3(a). We observe that WER_{norm} is independent from model sizes, and has a power-law relationship with training data. It indicates that for the range of data in this study,
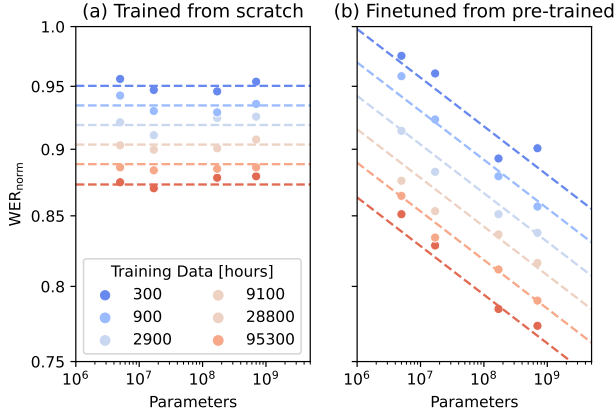
Figure 3: *Same data as in Figure 2 but plotted with a different x-axis. Normalized WER* WER_{norm} *(defined in Equation* (5)) *versus model parameters for (a) models trained from scratch (b) models finetuned from pre-trained models. Dashed lines denote model predictions from Equation* (7) *for (a) and Equation* (8) *for (b). Both axes are on log scale.*

the model is limited by training data. Because the training data here require human annotation and are hence limited, this study is limited to this low-data regime, but we expect that this regime would be the case for most speech recognition applications; while a model size of 5M is at the lower end of what's typical for a BERT model, 95, 300 of training data are at the upper end of what's usually used for speech recognition systems. It is in contrast to previous work on language modeling [8] where text data is easily accessible, which allows it to explore both scenarios of being limited by training data and limited by model sizes.

For this regime where data is limited, the previous work [8] proposed the following equation to describe the relationship between test set performance and training data size,

$$L = \left(\frac{D_C}{D} \right)^{\alpha_D}, \qquad (6)$$

where L is the test loss, D is training data size, and D_C is a model parameter and represents the critical value of D where the contribution of data to the loss function is equal to 1.0.

We found that this equation also describes the data well for WER_{norm}, as evidenced by the goodness of the fit in dashed lines in Figures 2(a) and 3(a). The equation is now,

$$WER_{norm} = \left(\frac{D_C}{D} \right)^{\alpha_D}, \qquad (7)$$

where D_C is 8.82 and α_D is 0.0146, both of which are estimated from fitting the data of $log(WER_{norm})$ vs $log(D_C)$. The small value for D_C indicates that a small amount of data is sufficient to bring WER_{norm} below 1.0. This makes sense for our data, since if the second pass just outputs the same scores for all the n-best hypotheses, due to linear interpolation with the first pass scores, WER would be the same as the first pass, and WER_{norm} would be equal to 1.0.

3.3. Finetuned from Pre-trained Models

We then study how WER is changed when the model is finetuned from a pre-trained model. As shown in Figures 2 and 3

pre-training model helps reduce WER as expected.

As shown in Figure 3(b), a larger model would now lead to more improvement, as opposed to what is found for models trained from scratch where the model performance is independent to the model size (as in Figure 3(a)). It shows that for pre-trained models, the improvement from increasing the model size is not due to the larger capacity of the model to better learn from finetuning data, but rather it is its ability to better memorize and leverage the knowledge from pre-training. Without studying the scaling laws of both randomly initialized and pre-trained models, it would be hard to distinguish the effects from the two, underscoring the importance of such exercises.

In Figure 2, to achieve the same WER for a given model size, one would require less training data due to pre-training. For example, for a 5M model to achieve WER_{norm} of 0.875, data required would be $\sim$ 30k if trained from scratch, but only $\sim$ 1k hour if pre-trained. Furthermore, this delta in data required depends on model size (as discussed in the previous paragraph) and also training data itself (reflected by the change of slope between Figures 2 (a) and (b)).

This reduction in data required due to pre-training and its relationship to finetuning data and model size are consistent with a previous work [25] studying scaling for transfer learning. In the paper, it introduces the concept of effective data transferred from pre-training, D_T, to capture this delta, and found that it can be captured by $D_T = kD^\alpha N^\beta$. Hence, for a given data size, due to pre-training, its effective data size would become $(D + D_T)$, or $(D + kD^\alpha N^\beta)$.

As also explained in the paper [25], in this low data regime, one can ignore the contribution of the original D as the effective data from transfer is much greater than the amount of data finetuned on, $D_T \gg D$, which is also demonstrated in the example before where D_T and D are 30k and 1k respectively.

Hence, the effective data size $(D + kD^\alpha N^\beta)$ can be simplified to $kD^\alpha N^\beta$. Plugging into Equation (7), and we have the equation as follows,

$$WER_{norm} = \left(\frac{D_C}{kD^\alpha N^\beta} \right)^{\alpha_D}. \qquad (8)$$

D_C and α_D have the same values as in Equation (7). As shown with dashed lines in Figures 2(b) and 3(b), the equation captures the data well, where k is 2.27×10^{-11}, α is 1.71, and β is 1.24. In contrast to the previous work [25] where the pre-training and finetuning steps are trained with the same loss and different domains of data, it shows that this equation also holds even when the pre-training and finetuning have different training objectives. The ratio of α and β suggest that a 10x increase in model size would be worth approximately a 5.3x increase in data size. Hence, the scaling law here offers useful insights as one decides between obtaining more training data and increasing model sizes to improve the model performance.

3.4. Applicability of the Scaling Law

Hence, we have proposed scaling laws for both randomly initialized model and pre-trained models as in Equations (7) and (8), respectively. Even though the data sizes (from 300 to 95300) and model sizes (from 5M to 700M) represent most speech recognition systems, it is still helpful to discuss the limitation in the applicability of the scaling laws described here.

First, as also described in the previous paper [28], scaling laws would break down at both extreme ends of the data size spectrum. When there is very little data, the model can perform as well as random guessing, and in our case the second

pass cannot further improve over the first pass, yielding 1.0 for WER_{norm}; at the other extreme, there would be a non-zero lower bound error past which the models will not be able to improve with more data (or model parameters). This lower bound error or irreducible error would include Bayers error and noises in the data.

Second, as shown in Section 3.2, for randomly initialized models, the regime in this study is limited by the amount of data. Once we increase by a sufficient amount of data, the model size would start to influence WER_{norm} even for randomly initialized models, and scaling laws for both randomly initialized models and pre-trained models likely need to be modified, as shown in the previous paper [25].

4. Conclusions

Using RescoreBERT as the rescorer, we demonstrate that scaling laws are also applicable for discriminative speech recognition rescoring models, for over two orders of magnitude of range studied. For randomly initialized models, WER is found to have power-law relationship with training data size, and independent from model sizes, indicating that it is operating at a regime where the data is limited. Due to the large data in this study (almost 100k), we expect that this regime would be the case for most speech recognition applications.

For pre-trained models, it is found that WER also has a power-law relationship with training data and model sizes. Noticeably, different from randomly initialized models, WER now decreases with increasing model size, as larger model has the ability to better memorize and leverage the knowledge from pre-training. In addition, it is found that a pre-trained model would require less data than a randomly initialized model of the same size, representing the effective data transferred from pre-training. This effective data transferred also follows a scaling law with the data and model sizes.

While we focus on RescoreBERT as the discrminative rescoring model to study the scaling laws, it would be interesting to see whether the scaling laws proposed in this study hold true for other types of discriminative rescoring models, especially when they are conditioned on first-pass audio/lattice [3, 4, 5, 29].

5. References

[1] Y. Xia, F. Tian, L. Wu, J. Lin, T. Qin, N. Yu, and T.-Y. Liu, "Deliberation networks: Sequence generation beyond one-pass decoding," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] T. N. Sainath, R. Pang, D. Rybach, Y. He, R. Prabhavalkar, W. Li, M. Visontai, Q. Liang, T. Strohman, Y. Wu, I. McGraw, and C.-C. Chiu, "Two-pass end-to-end speech recognition," in *Proc. Interspeech*, 2019, pp. 2773–2777.

[3] K. Hu, T. N. Sainath, R. Pang, and R. Prabhavalkar, "Deliberation model based two-pass end-to-end speech recognition," in *Proc. IEEE ICASSP*, 2020, pp. 7799–7803.

[4] K. Hu, R. Pang, T. N. Sainath, and T. Strohman, "Transformer based deliberation for two-pass speech recognition," in *Proc. IEEE Spoken Language Technology Workshop*, 2021, pp. 68–74.

[5] A. Gandhe and A. Rastrow, "Audio-attention discriminative language model for ASR rescoring," in *Proc. IEEE ICASSP*, 2020, pp. 7944–7948.

[6] T. Hori, C. Hori, S. Watanabe, and J. Hershey, "Minimum word error training of long short-term memory recurrent neural network language models for speech recognition," in *Proc. IEEE ICASSP*, 2016, pp. 5990–5994.

[7] R. Prabhavalkar, T. N. Sainath, Y. Wu, P. Nguyen, Z. Chen, C.-C. Chiu, and A. Kannan, "Minimum word error rate training for attention-based sequence-to-sequence models," in *Proc. IEEE ICASSP*, 2018, pp. 4839–4843.

[8] 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.

[9] B. Ghorbani, O. Firat, M. Freitag, A. Bapna, M. Krikun, X. Garcia, C. Chelba, and C. Cherry, "Scaling laws for neural machine translation," *arXiv preprint arXiv:2109.07740*, 2021.

[10] M. A. Gordon, K. Duh, and J. Kaplan, "Data and parameter scaling laws for neural machine translation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 5915–5922.

[11] T. Henighan, J. Kaplan, M. Katz, M. Chen, C. Hesse, J. Jackson, H. Jun, T. B. Brown, P. Dhariwal, S. Gray *et al.*, "Scaling laws for autoregressive generative modeling," *arXiv preprint arXiv:2010.14701*, 2020.

[12] X. Zhai, A. Kolesnikov, N. Houlsby, and L. Beyer, "Scaling vision transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 104–12 113.

[13] J. Droppo and O. Elibol, "Scaling laws for acoustic models," *arXiv preprint arXiv:2106.09488*, 2021.

[14] E. Variani, T. Chen, J. Apfel, B. Ramabhadran, S. Lee, and P. Moreno, "Neural oracle search on n-best hypotheses," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7824–7828.

[15] K. Hu, B. Li, and T. N. Sainath, "Scaling up deliberation for multilingual ASR," in *2022 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2023, pp. 771–776.

[16] L. Xu, Y. Gu, J. Kolehmainen, H. Khan, A. Gandhe, A. Rastrow, A. Stolcke, and I. Bulyko, "RescoreBERT: Discriminative speech recognition rescoring with BERT," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6117–6121.

[17] H. Huang and F. Peng, "An empirical study of efficient ASR rescoring with transformers," *arXiv preprint arXiv:1910.11450*, 2019.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, Minneapolis, Minnesota, Jun. 2019, pp. 4171–4186.

[19] J. Shin, Y. Lee, and K. Jung, "Effective sentence scoring method using BERT for speech recognition," in *Asian Conference on Machine Learning*. PMLR, 2019, pp. 1081–1093.

[20] J. Salazar, D. Liang, T. Q. Nguyen, and K. Kirchhoff, "Masked language model scoring," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 2699–2712.

[21] T.-W. Wu, I.-F. Chen, and A. Gandhe, "Learning to rank with BERT-based confidence models in ASR rescoring," in *Proc. Interspeech 2022*, 2022, pp. 1651–1655.

[22] K. Hu, T. N. Sainath, Y. He, R. Prabhavalkar, T. Strohman, S. Mavandadi, and W. Wang, "Improving deliberation by text-only and semi-supervised training," *arXiv preprint arXiv:2206.14716*, 2022.

[23] H. Futami, H. Inaguma, M. Mimura, S. Sakai, and T. Kawahara, "ASR rescoring and confidence estimation with ELECTRA," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2021, pp. 380–387.

[24] D. Fohr and I. Illina, "BERT-based semantic model for rescoring n-best speech recognition list," in *INTERSPEECH 2021*, 2021.

[25] D. Hernandez, J. Kaplan, T. Henighan, and S. McCandlish, "Scaling laws for transfer," *arXiv preprint arXiv:2102.01293*, 2021.

[26] S. Soltan, S. Ananthakrishnan, J. FitzGerald, R. Gupta, W. Hamza, H. Khan, C. Peris, S. Rawls, A. Rosenbaum, A. Rumshisky *et al.*, "AlexaTM 20B: Few-shot learning using a large-scale multilingual seq2seq model," *arXiv preprint arXiv:2208.01448*, 2022.

[27] Y. He, T. N. Sainath, R. Prabhavalkar, I. McGraw, R. Alvarez, D. Zhao, D. Rybach, A. Kannan, Y. Wu, R. Pang *et al.*, "Streaming end-to-end speech recognition for mobile devices," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6381–6385.

[28] J. Hestness, S. Narang, N. Ardalani, G. Diamos, H. Jun, H. Kianinejad, M. Patwary, M. Ali, Y. Yang, and Y. Zhou, "Deep learning scaling is predictable, empirically," *arXiv preprint arXiv:1712.00409*, 2017.

[29] P. Pandey, S. D. Torres, A. O. Bayer, A. Gandhe, and V. Leutnant, "Lattention: Lattice-attention in ASR rescoring," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 7877–7881.