

Dynamic Encoder Size Based on Data-Driven Layer-wise Pruning for Speech Recognition

Jingjing Xu^{1,2,*}, Wei Zhou^{*,†}, Zijian Yang^{1,2}, Eugen Beck², Ralf Schlüter^{1,2}

¹Machine Learning and Human Language Technology Group, Computer Science Dept., RWTH Aachen University, Germany

²AppTek GmbH, 52062 Aachen, Germany

{jxu, zhou, zyang, schluter}@ml.rwth-aachen.de, ebeck@apptek.com

Abstract

Varying-size models are often required to deploy ASR systems under different hardware and/or application constraints such as memory and latency. To avoid redundant training and optimization efforts for individual models of different sizes, we present the dynamic encoder size approach, which jointly trains multiple performant models within one supernet from scratch. These subnets of various sizes are layer-wise pruned from the supernet, and thus, enjoy full parameter sharing. By combining score-based pruning with supernet training, we propose two novel methods, *Simple-Top-k* and *Iterative-Zero-Out*, to automatically select the best-performing subnets in a data-driven manner, avoiding resource-intensive search efforts. Our experiments using CTC on both *Librispeech* and *TED-LIUM-v2* corpora show that our methods can achieve on-par performance as individually trained models of each size category. Also, our approach consistently brings small performance improvements for the full-size supernet.

Index Terms: Speech recognition, Supernet training, Dynamic encoder, Pruning

1. Introduction

Automatic speech recognition (ASR) models run in different scenarios with different application needs and computational budgets. For some applications, inference speed is critical which often requires trading off accuracy for model latency. One of the simplest and most effective ways to reduce model latency is to reduce model size. For on-site ASR, edge devices have limited storage and memory budgets, thus also imposing constraints on the model size. Obtaining ASR models with different model sizes often requires optimizing training hyperparameters for each model individually. However, repeated training results in high computational costs. Therefore, arises the question of how to efficiently train models of different sizes.

Pruning [1, 2, 3], as a model compression technique, is commonly used to obtain neural models of smaller sizes. Pruning aims at removing unimportant weights from the network. The lottery ticket hypothesis [4, 5] discovers that there exists a sparse net in a full network that can achieve the same performance. However, pruning requires a converged base model and fine-tuning of each small model separately, so the problem of repeated training remains unresolved. The concept of supernet training is first proposed in [6]. The supernet and a fixed number of subnets fully share parameters and are simultaneously trained. After the joint training, all networks can achieve good convergence. However, how to efficiently search for the subnets during training is still challenging [7, 8].

* denotes equal contribution

† work done while at RWTH Aachen University, now at Meta

In this work, we combine the benefits of both ideas and demonstrate an efficient dynamic encoder training framework. We leverage score-based layer-wise pruning to find the optimal layer combination for the subnets, saving the computationally expensive search required by the general supernet training methods [9, 10]. Furthermore, we design an efficient two-step training pipeline. In Step 1, we propose two methods, *Simple-Top-k* and *Iterative-Zero-Out*, to effectively learn the associated layer importance scores in a data-driven way. In step 2, we generate binary masks for all subnets and exploit the sandwich rule [6] for efficient joint training of the supernet and subnets. Additionally, we explore different training techniques to mitigate the mutual training inference and further boost the word error rate (WER). We evaluate our approach by conducting experiments with the Conformer [11] connectionist temporal classification (CTC) [12] model on both *Librispeech* and *TED-LIUM-v2* datasets. The results show that with our proposed framework, multiple models with the desired number of layers can be obtained in a single training job, each with competitive WER performance. Even a slight WER improvement on the full-size model is obtained presumably due to the regularization effect of the co-trained subnets. We also investigate the selected layers for the subnets and unexpectedly find that the convolutional layers are selected the most.

2. Related Work

2.1. Supernet/Subnet Joint Training

The RNN-T cascaded encoder architecture [13, 14] utilizes the idea of auxiliary loss [15, 16], allowing direct connections between intermediate encoder layers and decoders. An advantage of this approach is that the training overhead is negligible since no additional forward pass is required for the subnets. Nevertheless, the low-level layers may not be the optimal choice for the subnets. With the same model size, there might be a better combination of layers to make up the subnets. [17] for example, comprises the subnets by choosing every other or every third layer. To avoid manual layer selection, our work utilizes a score-based pruning method to achieve automatic layer selection during training. [18] randomly select one subnet from a total of 1000 subnets at each step of supernet training, which does not guarantee to find the optimal subnet under a specific size constraint. [9, 10] use evolutionary search to find the top-performing subnets under different size constraints from a pre-defined search space. The WER and loss on the validation set are used as the ranking metric in [9] and [10], respectively. As a result, in one search procedure, all possible subnet candidates need to be forwarded once with the whole validation data. Although [9] leverages quantization to make the inference more efficient, repeating such a search procedure in each training step is computationally expensive. [19, 20] use regularization tricks

like stochastic depth [21] and auxiliary loss [15, 16] to make model pruning aware and uses layer-wise pruning after training to search for subnets. Compared to these methods, our work determines the optimal subnets based on importance scores, thereby saving resource-intensive search efforts. [22] employs unstructured gradient-based pruning criteria to determine the subnets. Unstructured pruning may result in the irregular sparse matrix which requires reconstruction original dense shape in inference, hindering the acceleration in practical use. Thus, our work applies structured pruning to avoid irregular sparsity.

2.2. Pruning

Magnitude-based pruning [23, 24] preserves weights with the highest absolute values. However, the weight magnitude does not necessarily reflect the weight importance. To address this issue, movement pruning [25] considers first-order information, which is how the weights change in training. In this work, we follow their idea for our score-based pruning. Furthermore, [26] introduces L_0 norm regularization on the non-zero elements of weights, so that the models can be pruned to a specified sparsity. There has been growing interest in applying the L_0 norm for ASR tasks [1, 2, 3]. We compare our approach with L_0 norm in Sec. 4.2.

3. Dynamic Encoder Size

In this section, we present our approach to model encoders with dynamic size based on a supernet and M subnets that share parameters with the supernet. Consider the learnable parameters $\theta = \{\theta_j\}_{j=1}^L$, where L is the total number of layers, θ_j denotes the parameters of layer j . Let $\mathcal{C} = \{k_1, \dots, k_m, \dots, k_M\}$ denotes a predefined set of expected number of layers for the M subnets n_1, n_2, \dots, n_M . \mathcal{C} is sorted in a decreasing order such that $k_1 = L$ and $k_M = k_{min} = \min_{m \in M} k_m$. For each subnet n_m , a binary pruning mask $z_m \in \{0, 1\}^L$ is learned such that $\sum_{j=1}^L z_m^j = k_m$, where z_m^j denotes the mask for layer j . $z_m^j = 0$ indicates that layer j is pruned for subnet n_m . The layer importance scores are specified as $s = \{s_l\}_{l=1}^L \in \mathbb{R}^L$. The joint training optimization problem can be formulated as:

$$\min_{\theta, \forall z_m} \mathbb{E}_{(x,y) \in \mathcal{D}} \left[\mathcal{L}_{ASR}(x, y; \theta) + \sum_{m=1}^M \lambda_m \mathcal{L}_{ASR}(x, y; \theta \odot z_m) \right],$$

where \mathcal{D} is the training data and λ_m is the tunable loss scale, \odot denotes the element-wise product.

We design an efficient 2-step training pipeline. The main goal of Step 1 is to automatically learn the layer importance score. In Step 1, the supernet and one subnet are jointly trained from scratch. The subnet is initialized with the full-size model and progressively layer-wise pruned until its number of layers reaches k_{min} . During this progressive pruning process, the layer scores learned at each intermediate size category also reflecting the layer selection for the corresponding subnet. We believe that with such a process, layer selection can be learned in a smooth way. Besides, the ASR loss of the full-sized supernet ensures that the performance of the supernet is not compromised. After the layer importance scores are learned, in Step 2, we generate binary pruning masks for each subnet n_1, n_2, \dots, n_M based on the importance score and then jointly train the supernet and all M subnets.

3.1. Step 1 - Progressive Self-Pruning

In Step 1, we progressively prune the subnet with a dynamically decreasing size, while the supernet and the subnet are jointly trained during the pruning process. This step takes about 60% of the total training time. The subnet is initialized with 0 spar-

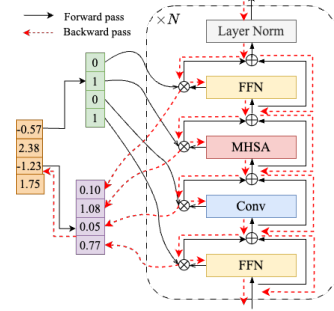


Figure 1: Illustration of Simple-Top-k, STE uses a relaxed k-hot vector to estimate the gradients of the binary mask.

sity, which is defined as the percentage of pruned layers to the total number of layers. We adopt an iterative training procedure to gradually increase the target sparsity of the subnet until it reaches the desired maximum value $\frac{L-k_{min}}{L}$. We denote I as the total number of training iterations, each with ΔT training steps. In the i -th iteration, we set the number of layers of the subnet to $k = L - \frac{(L-k_{min}) \times i}{I}$. In the following, we present two iterative self-pruning methods for the given k to learn the associated layer importance.

3.1.1. Simple-Top-k

Simple-Top-k is a differentiable top-k operator inspired by [25]. For a given number of layers k , the pruning binary mask for the subnet is z is $\{z^j | z^j = 1 \text{ if } s_j \text{ in } \text{topk}(s, k) \text{ else } 0, j = 1, 2, \dots, L\}$. Since such a binary mask z is not differentiable, Simple-Top-k uses z in the forward pass to calculate the loss. In the backward pass, as depicted in Figure 1, the straight through estimator (STE) [27] is used to approximate the gradients for the step function. We use a relaxed k-hot vector $\alpha = [\alpha_1, \dots, \alpha_L]$ where $\sum_{j=1}^L \alpha_j = k, 0 \leq \alpha_j \leq 1$ to approximate the gradient of the binary mask z . The relaxed top-k algorithm is used to derive α from s , we refer the reader to [28] for more details about the algorithm.

3.1.2. Iterative-Zero-Out

Simple-Top-k uses STE, which leads to inconsistency in forward and backward passes. Since the impact of the approximate gradient is unclear [29], we design another method called Iterative-Zero-Out, to circumvent the usage of STE. The pruning mask for the subnet here is defined as $z = [f(s_1), f(s_2), \dots, f(s_L)]$ where $f(\cdot)$ is an activation function. We use sigmoid for $f(\cdot)$ in this work. To constrain $\sum_{j=1}^L z^j = k$, we add an L1 norm term to the training objective to ensure sparsity. The total optimization loss can be formulated as

$$\mathcal{L}_{ASR}(x, y; \theta) + \lambda \mathcal{L}_{ASR}(x, y; \theta \odot z) + \gamma \left| \frac{\sum_{j=1}^L (z^j) - k}{L} \right|,$$

where γ is a tunable scale. However, by simply adding a sparsity loss, we observe that z tends to converge closer to a uniform distribution. Therefore, we adopt the zero-out idea from [24], where each iteration adopts the following procedure:

1. Jointly train the supernet with parameters θ and the subnet with parameters $\theta \odot z$ for ΔT steps
2. Zero out the smallest $L - k$ elements in z but meanwhile keep them in the computation graph so that they can still get updated in further iterations and may have a chance to be revived.

In this way, the mask z will converge very close to a k-hot vector, though not exactly.

3.2. Step 2 - Supernet/Subnets Joint Training

In Step 2, we set the binary mask $z_m = \{z_m^j | z_m^j = 1 \text{ if } s_j \text{ in } \text{topk}(s, k_m) \text{ else } 0\}$ for subnet n_m with k_m number of layers. All binary masks are kept fixed in this step. The sandwich rule, which is proposed in [6] and successfully applied in [10, 9], is employed in this step to improve efficiency. More specifically, in each training step, we jointly train the supernet, the smallest subnet, and one medium subnet randomly sampled from the remaining M-2 subnets.

Furthermore, we take advantage of layer dropout [21] to diminish the mutual interference between subnets and supernet during joint training. More precisely, we apply layer dropout to those pruned layers that have indices $\{j | s_j \text{ not in } \text{topk}(s, k_{min})\}$, similar to [17]. We also empirically observed that using around 40% of the total training time is already enough to reach good performance for all subnets.

4. Experiments

4.1. Setup

We conduct the experiments on the 960h *Librispeech* corpus [30] and the *TED-LIUM-v2* corpus [31]. We use a phoneme-based CTC model as in [32]. We use a set of 79 end-of-word augmented phonemes [33]. The acoustic model consists of a VGG front end and 12 Conformer [11] blocks. In the Conformer block, we do not apply relative positional encoding. Instead, we swap the order of the convolution module and the multi-head self-attention module as in [34] to speed up the training and inference. The model size is set to 512 for *Librispeech* and 384 for *TED-LIUM-v2* corpus. We use log Mel-filterbank features as input and specaug [35] for data augmentation. Similar to [36], one cycle learning rate scheduler is used for training. The learning rate (lr) is first linearly increased from $4 \cdot 10^{-6}$ to $4 \cdot 10^{-4}$ for 45% of the training time, then linearly decreased from $4 \cdot 10^{-4}$ to $4 \cdot 10^{-6}$ for another 45% of the time. For the rest 10% of training time, the lr linearly decays to 10^{-7} . We train 50 epochs for *TED-LIUM-v2* corpus and 30 epochs for *Librispeech*. The loss scale for each subnet in both training steps is set to 0.3. In inference, we apply Viterbi decoding with a 4-gram word-level language model. The config files and code to reproduce the results can be found online¹.

4.2. 1 Supernet + 1 Subnet

In Table 1, we compare ASR results of encoders with 24 and 48 layers trained using different methods. For the *Aux-Loss* method, we add an auxiliary CTC loss with a loss scale of 0.3 to the output of the 6-th Conformer block (corresponds to 24 layers). We can see that for the 48 layers, both *Simple-Top-k* and *Iterative-Zero-Out* achieve the same or slightly better WER. Compared to the individually trained model, *Simple-Top-k* and *Iterative-Zero-Out* achieve on-par WER for the 48-layer model and better WER for the 24-layer model. Compared to *Aux-Loss*, both *Simple-Top-k* and *Iterative-Zero-Out* gain slight WER improvement on the 48-layer model and substantially outperform the 24-layer model by a relative $\sim 8\%$ with even fewer parameters. It confirms our hypothesis in Sec. 2.1 that low-level layers are not the optimal choice for the subnet.

We also compare our proposed method with the widely-used pruning method *L₀-Norm* [26]. We use the implementation from [37]. We adapt *L₀-Norm* to our joint training scenario and apply it to the layer-wise self-pruning in Step 1. The

¹<https://github.com/rwth-i6/returnn-experiments/tree/master/2024-dynamic-encoder-size>

Table 1: ASR results comparison between different approaches for training two encoders with 48 and 24 layers on *TED-LIUM-v2* dev set. The 48-layer model has a total of 41.7 M parameters.

Training	Large	Small	
	WER[%]	Params. [M]	WER[%]
separately	7.5	20.9	8.4
Aux-Loss	7.6		8.8
<i>L₀-Norm</i>	7.7	18.7	9.5
<i>Simple-Top-k</i>	7.5	19.7	8.1
<i>Iterative-Zero-Out</i>	7.4	18.8	8.2

results in Table 1 show that there is a considerable WER degradation for the 24-layer model trained from *L₀-Norm* compared to other methods. The reason could be that the mask z is generated by adding a random variable $u \sim U(0, 1)$ distribution. At each training step, the layer outputs of the subnet are scaled by a fluctuating variable, which may cause disturbance to training.

4.3. 1 Supernet + 2 Subnets

Table 2 and Table 3 report the WER of models with 16, 32, and 48 layers on *TED-LIUM-v2* and *Librispeech* test set. On *TED-LIUM-v2* test set, *Simple-Top-k* performs best across all three model sizes. For the *Librispeech* test set, *Simple-Top-k* performs best only on large and medium-sized models. Both the proposed methods outperform the separately trained baselines for large and medium models, while performance degradation is observed for the small one. A likely reason is that the loss scale for supernet is 1 and for all subnets is 0.3, thus placing more emphasis on the supernet during training. Furthermore, we observe that the joint training can improve the WER of the supernet. One possible explanation is that the layer masking on the shared parameters introduces some regularization effect, similar to LayerDrop [19]. In addition, the models trained from *Iterative-Zero-Out* perform slightly worse than *Simple-Top-k*.

Table 2: ASR results of three encoders with 48, 32, and 16 layers on *TED-LIUM-v2* test set. The 48-layer model has a total of 41.7 M parameters.

Training	Large	Medium		Small	
	WER[%]	Params.[M]	WER[%]	Params.[M]	WER[%]
separately	8.1	28.1	8.4	14.4	9.3
Aux-Loss	7.9		8.6		10.9
<i>Simple-Top-k</i>	7.8	27.5	8.0	14.1	9.1
<i>Iterative-Zero-Out</i>	8.1	25.8	8.2	12.9	9.6

Table 3: ASR results of three encoders with 48, 32 and 16 layers on *Librispeech* test set. The 48-layer model has a total of 74.1 M parameters.

Training	Large		Medium				Small	
	WER[%]		Params. [M]	WER[%]		Params. [M]	WER[%]	
	clean	other		clean	other		clean	other
separately	3.3	7.1	49.9	3.5	7.7	25.6	3.6	8.4
Aux-Loss	3.2	6.9		3.6	7.9		4.5	9.7
<i>Simple-Top-k</i>	3.1	6.8	47.1	3.2	7.0	23.8	3.9	9.1
<i>Iterative-Zero-Out</i>	3.2	7.1	47.0	3.2	7.2	26.7	4.1	9.6

4.4. Ablation Study

4.4.1. Layer Dropout

The layer dropout in Step 2 can alleviate the mutual interference between the supernet and subnets, thus playing an important role in joint training. We explicitly study the impact of the layer dropout and present the result in Table 4. If layer dropout is applied in Step 1, we only apply it on the unselected layers, i.e., $z^j \neq 1$. Table 4 demonstrates that it is not necessary to

Table 4: ASR results of applying different values of layer dropout in Step 1 and Step 2 on Librispeech test set. Method Simple-Top-k is used to train three models with 16,32,48 layers respectively.

Layer dropout		WER [%]					
Step 1	Step 2	Large		Medium		Small	
		clean	other	clean	other	clean	other
n/a	0	3.1	6.9	3.2	7.1	4.1	9.7
	0.1	3.2	6.8	3.2	7.0	4.2	9.8
	0.3	3.1	6.8	3.2	7.0	3.9	9.1
	0.5	3.3	7.2	3.3	7.3	4.1	9.5
0.1	0.3	3.3	7.0	3.2	7.2	4.2	9.6
0.3		3.2	7.0	3.2	7.0	4.0	9.2

apply layer dropout in Step 1. The best result is achieved by only applying layer dropout with a value of 0.3 in Step 2. Additionally, we have also tried to apply dropout on the entire layer group as in [19]. However, we empirically find out that applying dropout to each layer individually performs better. This may make the training more robust since in each training step, different combinations of layers can be dropped out.

4.4.2. Number of Self-Pruning Iterations

Table 5 compares the WERs of employing a different number of pruning iterations in Step 1 in Sec. 3.1. The WERs of all three size models tend to decrease when more pruning iterations are used. As the models are trained from scratch in Step 1, the layer importance may change significantly during training. Presumably, using more iterations avoids making suboptimal decisions that select suboptimal layers in the early stages.

Table 5: ASR results of employing different number of self-pruning iterations in Step 1 on Librispeech test set. Simple-Top-k is used to train three models with 16,32,48 layers respectively.

# iterations	WER [%]					
	Large		Medium		Small	
	clean	other	clean	other	clean	other
1	3.3	7.1	3.4	7.3	3.8	9.3
2	3.2	7.1	3.2	7.2	3.9	9.5
4	3.3	7.0	3.3	7.2	4.1	9.3
8	3.1	6.9	3.2	7.2	3.9	9.2
32	3.1	6.8	3.2	7.0	3.9	9.1

4.4.3. Training Time Distribution

Table 6 shows the WER results of using different training time distributions in Step 1 and Step 2. We observe that using 60% of training time in Step 1 and 40% in Step 2 achieves the best performance. If less time is distributed to Step 1, the layer importance scores may not be learned well, leading to premature decisions. On the contrary, allocating more time in Step 1 will lead to insufficient joint training time under a fixed training budget and also lead to performance degradation.

Table 6: ASR results of using different training time distribution in Step 1 and Step 2 on Librispeech test set. Method Simple-Top-k is used to train three models with 16,32,48 layers respectively.

Step 1	Step 2	WER[%]					
		Large		Medium		Small	
		clean	other	clean	other	clean	other
50%	50%	3.3	7.1	3.3	7.2	4.0	9.3
60%	40%	3.1	6.8	3.2	7.0	3.9	9.1
70%	30%	3.1	7.1	3.2	7.2	4.4	10.9

Table 7: ASR results of four encoders with 12, 24, 36 and 48 layers trained using the sandwich rule on TED-LIUM-v2 test set. The 48-layer model has totally 41.7 M parameters.

Training	Large	Medium 1		Medium 2		Small	
	WER [%]	Params. [M]	WER [%]	Params. [M]	WER [%]	Params. [M]	WER [%]
separately	8.1	31.4	8.3	21.2	8.6	10.9	9.8
Aux-Loss	8.1	31.4	8.3	21.2	9.1	10.9	12.2
Simple-Top-k	8.0	31.1	8.1	20.6	8.6	10.0	10.2
Iterative-Zero-Out	7.9	29.3	8.1	19.3	8.8	10.7	10.3

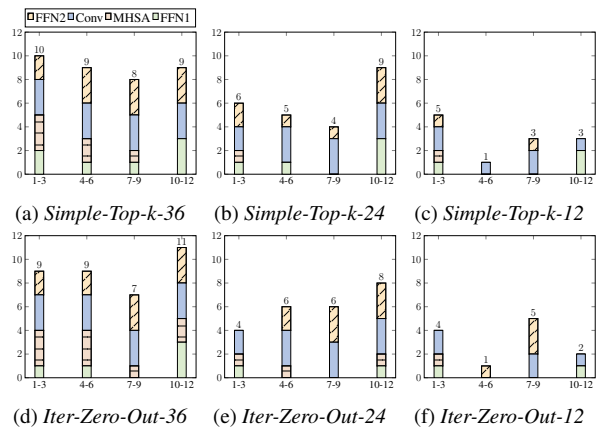


Figure 2: The distribution of selected layers for the models with 12, 24 and 36 layers shown in Table 7.

4.5. Sandwich Rule: 1 Supernet and $M > 2$ Subnets

Table 7 shows the effectiveness of applying the sandwich rule. Compared to jointly training three models, there is no increase in the training computation. In Step 2, we only need to update the 48-layer model, 12-layer model, and one randomly selected model with 24 or 36 layers. Yet we can obtain one more subnet and the WERs of all four models are still competitive.

We further analyse the selected layers for the models with 12, 24, and 36 layers in Figure 2. We note that the feed-forward layers tend to be pruned the most, followed by multi head self-attention (MHSA) layers. Also, the remaining MHSA layers tend to be distributed at the bottom layer. To our surprise, the convolutional layers are the most selected layers, indicating that it is generally more important for our phoneme CTC model.

5. Conclusion

In this work, we present an efficient training scheme to obtain models of various sizes by combining score-based model pruning and supernet training. We also propose two novel methods, Simple-Top-k and Iterative-Zero-Out, to automatically learn the optimal layer combinations for the subnets through the training process. Furthermore, we combine different training methods including layer dropout and the sandwich rule to achieve better overall performance. The experimental results on Librispeech and TED-LIUM-v2 show that for each size, models trained using our approach can match or slightly outperform models trained individually, and largely outperform the models trained with auxiliary loss. This shows that our training scheme can significantly reduce training redundancy while preserving model performance. For future work, the proposed approaches can be extended by using finer pruning granularity for subnets such as the attention head in MHSA, the dimension of feed-forward layer, etc. as in [1]. Applying in-place knowledge distillation to transfer the knowledge from supernet to subnets can also give potential improvements as reported in [22, 38].

6. Acknowledgements

This work was partially supported by the project RESCALE within the program *AI Lighthouse Projects for the Environment, Climate, Nature and Resources* funded by the Federal Ministry for the Environment, Nature Conservation, Nuclear Safety and Consumer Protection (BMUV), funding ID: 67KI32006A.

7. References

- [1] H. Jiang, L. L. Zhang, Y. Li, Y. Wu, S. Cao, T. Cao, Y. Yang, J. Li, M. Yang, and L. Qiu, "Accurate and Structured Pruning for Efficient Automatic Speech Recognition," in *Interspeech*, Dublin, Ireland, Aug. 2023, pp. 4104–4108.
- [2] H. Wang, S. Wang, W. Zhang, H. Suo, and Y. Wan, "Task-Agnostic Structured Pruning of Speech Representation Models," in *Interspeech*, Dublin, Ireland, Aug. 2023, pp. 4104–4108.
- [3] Y. Peng, K. Kim, F. Wu, P. Sridhar, and S. Watanabe, "Structured Pruning of Self-Supervised Pre-Trained Models for Speech Recognition and Understanding," in *ICASSP*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
- [4] J. Frankle and M. Carbin, "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks," in *ICLR*, New Orleans, USA, May 2019.
- [5] S. Ding, T. Chen, and Z. Wang, "Audio Lottery: Speech Recognition Made Ultra-Lightweight, Noise-Robust, and Transferable," in *ICLR*, Virtual, Apr. 2022.
- [6] J. Yu, L. Yang, N. Xu, J. Yang, and T. S. Huang, "Slimmable Neural Networks," in *ICLR*, New Orleans, USA, May 2019.
- [7] T. W. Z. Z. S. H. Han Cai, Chuang Gan, "Once-for-All: Train One Network and Specialize it for Efficient Deployment," in *ICLR*, Addis Ababa, Ethiopia, Apr. 2020.
- [8] M. Chen, H. Peng, J. Fu, and H. Ling, "Autoformer: Searching transformers for visual recognition," in *ICCV*, Montreal, Canada, 2021, pp. 12 270–12 280.
- [9] Y. Shangguan, H. Yang, D. Li, C. Wu, Y. Fathullah, D. Wang, A. Dalmia, R. Krishnamoorthi, O. Kalinli, J. Jia, J. Mahadeokar, X. Lei, M. Seltzer, and V. Chandra, "TODM: Train Once Deploy Many Efficient Supernet-Based RNN-T Compression For On-device ASR Models," arXiv:2309.01947, Nov.2023.
- [10] H. Yang, Y. Shangguan, D. Wang, M. Li, P. Chuang, X. Zhang, G. Venkatesh, O. Kalinli, and V. Chandra, "Omni-Sparsity DNN: Fast Sparsity Optimization for On-Device Streaming E2E ASR Via Supernet," in *ICASSP*, Virtual and Singapore, May 2022, pp. 8197–8201.
- [11] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented Transformer for Speech Recognition," in *Interspeech*, Shanghai, China, Oct. 2020, pp. 5036–5040.
- [12] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *ICML*, Pittsburgh, Pennsylvania, USA, Jun. 2006, pp. 369–376.
- [13] A. Narayanan, T. N. Sainath, R. Pang, J. Yu, C. Chiu, R. Prabhavalkar, E. Variiani, and T. Strohmaier, "Cascaded Encoders for Unifying Streaming and Non-Streaming ASR," in *ICASSP*, Toronto, Canada, Jun. 2021, pp. 5629–5633.
- [14] S. Ding, W. Wang, D. Zhao, T. N. Sainath, Y. He, R. David, R. Botros, X. Wang, R. Panigrahy, Q. Liang, D. Hwang, I. McGraw, R. Prabhavalkar, and T. Strohmaier, "A Unified Cascaded Encoder ASR Model for Dynamic Model Sizes," in *Interspeech*, Incheon, Korea, Sep. 2022, pp. 1706–1710.
- [15] A. Tjandra, C. Liu, F. Zhang, X. Zhang, Y. Wang, G. Synnaeve, S. Nakamura, and G. Zweig, "DEJA-VU: Double Feature Presentation and Iterated Loss in Deep Transformer Networks," in *ICASSP*, Barcelona, Spain, May 2020, pp. 6899–6903.
- [16] J. Lee and S. Watanabe, "Intermediate Loss Regularization for CTC-Based Speech Recognition," in *ICASSP*, Toronto, Canada, Jun. 2021, pp. 6224–6228.
- [17] Y. Shi, V. Nagaraja, C. Wu, J. Mahadeokar, D. Le, R. Prabhavalkar, A. Xiao, C. Yeh, J. Chan, C. Fuegen, O. Kalinli, and M. L. Seltzer, "Dynamic Encoder Transducer: A Flexible Solution for Trading Off Accuracy for Latency," in *Interspeech*, Brno, Czechia, Aug. 2021, pp. 2042–2046.
- [18] R. Wang, Q. Bai, J. Ao, L. Zhou, Z. Xiong, Z. Wei, Y. Zhang, T. Ko, and H. Li, "Lighthubert: Lightweight and Configurable Speech Representation Learning with Once-for-all Hidden-unit Bert," in *Interspeech*, Incheon, Korea, Sep. 2022, pp. 1686–1690.
- [19] A. Fan, E. Grave, and A. Joulin, "Reducing Transformer Depth on Demand with Structured Dropout," in *ICLR*, Addis Ababa, Ethiopia, Apr. 2020.
- [20] J. Lee, J. Kang, and S. Watanabe, "Layer pruning on demand with intermediate CTC," in *Interspeech*, Brno, Czechia.
- [21] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep Networks with Stochastic Depth," in *ECCV*, Amsterdam, Netherlands, Oct. 2016, pp. 646–661.
- [22] Z. Wu, D. Zhao, Q. Liang, J. Yu, A. Gulati, and R. Pang, "Dynamic Sparsity Neural Networks for Automatic Speech Recognition," in *ICASSP*, Toronto, Canada, Jun. 2021, pp. 6014–6018.
- [23] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning Both Weights and Connections for Efficient Neural Networks," in *NeurIPS*, Cambridge, MA, USA, 2015, p. 1135–1143.
- [24] C. J. Lai, Y. Zhang, A. H. Liu, S. Chang, Y. Liao, Y. Chuang, K. Qian, S. Khurana, D. D. Cox, and J. Glass, "PARP: Prune, Adjust and Re-Prune for Self-Supervised Speech Recognition," in *NeurIPS*, virtual, Dec. 2021, pp. 21 256–21 272.
- [25] V. Sanh, T. Wolf, and A. M. Rush, "Movement Pruning: Adaptive Sparsity by Fine-Tuning," in *NeurIPS*, virtual, Dec. 2020.
- [26] C. Louizos, M. Welling, and D. P. Kingma, "Learning Sparse Neural Networks through L_0 Regularization," in *ICLR*, Vancouver, Canada, Apr. 2018.
- [27] Y. Bengio, N. Léonard, and A. Courville, "Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation," arXiv:1308.3432, Aug.2013.
- [28] S. M. Xie and S. Ermon, "Reparameterizable Subset Sampling via Continuous Relaxations," in *IJCAI*, Macao, China, Aug. 2019, pp. 3919–3925.
- [29] P. Yin, J. Lyu, S. Zhang, S. J. Osher, Y. Qi, and J. Xin, "Understanding Straight-Through Estimator in Training Activation Quantized Neural Nets," in *ICLR*, New Orleans, USA, May 2019.
- [30] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, South Brisbane, Queensland, Australia, Apr. 2015, pp. 5206–5210.
- [31] A. Rousseau, P. Deléglise, and Y. Estève, "Enhancing the TED-LIUM Corpus with Selected Data for Language Modeling and More TED Talks," in *LREC*, Reykjavik, Iceland, May 2014, pp. 3935–3939.
- [32] W. Zhou, H. Wu, J. Xu, M. Zeineldeen, C. Lüscher, R. Schlüter, and H. Ney, "Enhancing and Adversarial: Improve ASR with Speaker Labels," in *ICASSP*, Rhodes Island, Greece, Jun. 2023, pp. 1–5.
- [33] W. Zhou, S. Berger, R. Schlüter, and H. Ney, "Phoneme based neural transducer for large vocabulary speech recognition," in *ICASSP*, Toronto, Canada, Jun. 2021, pp. 5644–5648.
- [34] B. Li, A. Gulati, J. Yu, T. N. Sainath, C. Chiu, A. Narayanan, S. Chang, R. Pang, Y. He, J. Qin, W. Han, Q. Liang, Y. Zhang, T. Strohmaier, and Y. Wu, "A Better and Faster end-to-end Model for Streaming ASR," in *ICASSP*, Toronto, Canada, Jun. 2021, pp. 5634–5638.
- [35] W. Zhou, W. Michel, K. Irie, M. Kitzka, R. Schlüter, and H. Ney, "The RWTH ASR System for Ted-Lium Release 2: Improving Hybrid HMM With SpecAugment," in *ICASSP*, Barcelona, Spain, May 2020, pp. 7839–7843.
- [36] W. Zhou, W. Michel, R. Schlüter, and H. Ney, "Efficient training of neural transducer for speech recognition," in *Interspeech*, Incheon, Korea, Sep. 2022, pp. 2058–2062.
- [37] M. Xia, Z. Zhong, and D. Chen, "Structured Pruning Learns Compact and Accurate Models," in *ACL*, Dublin, Ireland, May 2022, pp. 1513–1528.
- [38] V. Nagaraja, Y. Shi, G. Venkatesh, O. Kalinli, M. L. Seltzer, and V. Chandra, "Collaborative Training of Acoustic Encoders for Speech Recognition," in *Interspeech*, Brno, Czechia, Aug. 2021, pp. 4573–4577.