SparseVSR: Lightweight and Noise Robust Visual Speech Recognition – Extended Abstract

Pingchuan Ma[1,2] Adriana Fernandez-Lopez[1] Alexandros Haliassos[2] Honglie Chen[1] Stavros Petridis[1,2] Maja Pantic[1,2] *

[1]Meta AI [2] Imperial College London

1. Introduction

Recent advances in deep neural networks have contributed significantly to the progress in visual speech recognition (VSR). However, for VSR models to be practically applicable in real-world scenarios, they must be adaptable to multiple devices. Unfortunately, there is a significant gap between current methods and their deployment in resource-constrained devices. Sparse networks offer a promising solution due to their computational and memory efficiency, as well as their ability to be transferred and resistant to noise. In fact, studies have shown that certain pruning methods, like the lottery ticket hypothesis, can achieve comparable or even superior performance to dense models. In this paper, we consider the problem of pruning VSR networks. Thus, we explicitly look for a lightweight model computationally and memory efficient, transferable, and noise-robust. To this end, we conduct an in-depth analysis of various network pruning methods and set the first sparse model benchmark in VSR. Our sparse models reach state-of-the-art results on LRS3 with 19.5% WER at 10% sparsity and outperform the dense model up to 70% sparsity. Differently from previous works, we introduce visual noise augmentations during training to prevent overfitting and push forward the achievable performance of VSR models. We evaluate our 50% sparse model on 7 different visual noise types and achieve an overall absolute improvement of more than 2% WER compared to the dense equivalent, suggesting that sparse models are more robust to noise than the dense counterpart.

2. Methodology

Given a variable length video sequence X, we compute its corresponding transcription Y using a function $f(\cdot;\theta_T)$, where $\theta_T \in \mathbb{R}^p$ are the model parameters, p is the number of parameters of the network, T is the number of training epochs. Therefore, the dense model $Y = f(X;\theta_T)$ can be viewed as a combination of sequence-to-sequence and connectionist temporal classification models, which is primarily composed of a CNN-based frontend, a Conformer encoder, and a Transformer decoder as described in [10].

To create a sparse subnetwork, we adopt the conventional *unstructured global magnitude pruning* [6] approach, which collectively eliminates the weights with the smallest magnitudes regardless of their position within the network. After training the dense model for T epochs, pruning is applied to the parameters θ_T using a binary mask $m \in \{0,1\}^p$ to meet the target sparsity. This operation involves element-wise multiplication between the mask and model parameters ($\theta' = m \odot \theta_T$), resulting in pruned parameters θ' that meet the target sparsity, where $\odot$ denotes the element-wise product. Subsequently, rather than considering the pruned model as our final state, the model undergoes fine-tuning on relevant downstream datasets, where the goal is to learn (m, θ') such that the sparse subnetwork $f(\cdot;\theta')$ yields the lowest possible fine-tuned loss $\mathcal{L}$ on the given training data.

$$f(\cdot;\theta') = \mathcal{A}^{\mathcal{D}}_{i,r}(f(\cdot;\textsc{Frozen}(m \odot \theta_0))) \tag{1}$$

Given a training dataset $\mathcal{D}$, the algorithm is formulated in eq. (1), where $\mathcal{A}$ denotes an optimization algorithm, and i and r refer to the number of training iterations and rounds, respectively. Note, the FROZEN operation only freezes the pruned weights and keeps other weights adjustable. We split the subnetwork discovery in multiple rounds to extract better subnetworks, especially at high sparsity levels. The model starts at $\theta_0 = \theta_T$ and is iteratively pruned and re-trained for i iterations during r rounds to reach the target sparsity. Each round prunes $1/r$ from the non-zero weights [3]. We perform $r = 10$ rounds with a weight drop of 10% every time. After each round, we rewind the learning rate to its initial value and initialise the parameters to the last weights ($\theta_0 = \theta'$) for the next round until completion [12].

*Only non-Meta co-authors downloaded, accessed, and used the datasets. Only non-Meta authors conducted any of the dataset pre-processing (no dataset pre-processing took place on Meta's servers or facilities).

Method	Training Data (hours)	Noise	WER (%)	# Params.
CM-seq2seq [9]	438	X	46.9	56.4 M
KD+CTC [2]	772	X	59.8	212 M
AVHuBERT [13]	1,759	X	26.9	325 M
RAVEn [5]	1,759	X	24.4	493 M
TM-seq2seq [1]	1,362	X	58.9	54.2 M
Auto-AVSR [7]	3,448	X	20.5	250.4 M
RNN-T [11]	31,000	X	33.6	62.9 M
dense BASE-S	438	X	39.3	56.4 M
20% sparse BASE-S		X	37.3	45.4 M
50% sparse BASE-S		X	38.2	28.9 M
dense BASE-S	3,068*	✓	26.6	56.4 M
dense BASE-L		X	21.1	250.4 M
dense BASE-L		✓	20.3	250.4 M
10% sparse BASE-L		✓	19.5	225.7 M
20% sparse BASE-L		✓	20.0	201.1 M
30% sparse BASE-L		✓	19.8	176.5 M
40% sparse BASE-L		✓	20.2	151.9 M
50% sparse BASE-L		✓	19.6	127.3 M
60% sparse BASE-L		✓	19.9	102.7 M
70% sparse BASE-L		✓	20.1	78.1 M
80% sparse BASE-L		✓	21.8	53.5 M
90% sparse BASE-L		✓	26.3	28.8 M
8-bit dense BASE-L	3,068*	✓	20.3	250.4 M
distilled BASE-S		X	26.8	56.4 M
distilled BASE-S		✓	32.1	56.4 M

Table 1. Results of VSR networks on LRS3. Sparse models are generated by iterative pruning LRR [12]. Our models do not use LM. *LRS3, VoxCeleb2 and AVSpeech.

3. Results

3.1. Lightweight competitive VSR models

In Table 1, we show the performance of dense and sparse models on LRS3. We observe that our dense models are comparable to other VSR methods, using less training data. We highlight the benefits of adding noise during training, which provides data augmentation that allows the model to generalize better. Next, based on our dense BASE-L model exposed to noise, we iteratively generate the sparse models. We show that our sparse models reach state-of-the-art results with 19.5% WER at 10% sparsity, while they outperform the dense model up to 70% sparsity and remain without much degradation until 80% sparsity. We further highlight the benefits of sparse networks against other compression techniques. Table 1 shows the performance of the dense BASE-L model exposed to noise when trained for additional 75 epochs on non-uniform 8-bit quantization-aware-training, following [4]. We observe that the quantized model does not suffer degradation, but does not provide any benefits in terms of performance. In addition, we follow Ma et al. [8] to distill BASE-S student networks that share a similar amount of parameters as our sparser network without significant accuracy loss (i.e. 80% sparsity) using our dense BASE-L model as a teacher. We notice that the 80% sparse model outperforms the best distilled network by 5% WER absolute difference.

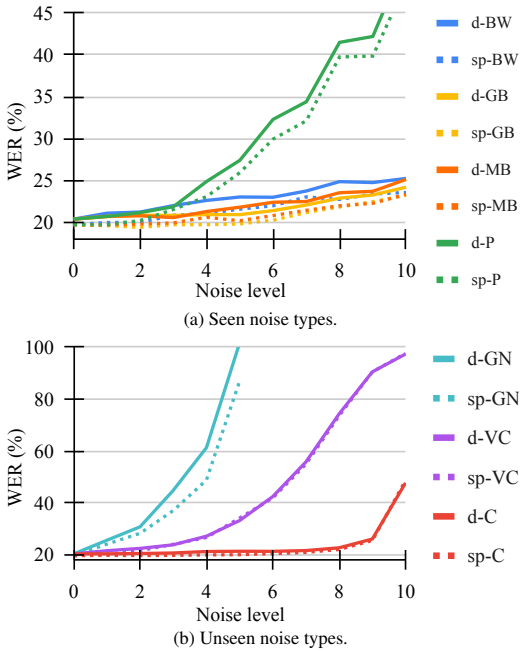

Figure 1. WER of noise-exposed dense "d" and 50% sparse "sp" BASE-L models on LRS3 corrupted by visual noises at different levels (level 0 means clean data).

3.2. Noise experiments

In order to investigate the robustness of our VSR models against visual noise, we run our 50% sparse and dense BASE-L models at different levels of noise. Note that all models are augmented with 4 visual noises, i.e. block-wise distortions (BW), Gaussian blur (GB), motion blur (MB) and pixelation (P), and 10 levels of corruption as well as clean image sequences. Results are shown in Figures 1a and 1b. Overall, we show that the gap between the sparse model and the dense model becomes increasingly larger as the level of noise increases, which indicates that the sparse model is more beneficial when image sequences are heavily corrupted. The same conclusions can be drawn when unseen noise such as JPEG video compression (VC), Gaussian noise (GN) and contrast (C) is added to the test set. In particular, when GN is added to the test set, our sparse model outperforms the dense model by 14.7% WER absolute improvement at level 5. This is probably because the sparse model filters deeper the image and only retains the relevant details, while the dense model may consider the noise. On the other hand, we show a 1.4% improvement on VC at level 1, which reduces as the level of corruption increases, indicating that our VSR models are less sensitive to video compression noise. This is likely due to the loss of important details on the mouth and lip appearance at high compression ratios of that noise type.

4. Conclusions

We revisit network pruning techniques and set the first benchmark in VSR. We present state-of-the-art results on LRS3 using VSR models at different sparsity levels. We also show no degradation up to 70% sparsity and a model parameter reduction rate up to 3.2. More importantly, we reinforce the fact that under similar conditions, sparse networks are more robust against noise than their dense counterpart.

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, et al. Deep audio-visual speech recognition. *IEEE Transactions on PAMI*, 2018. 2

[2] T. Afouras, J. S. Chung, and A. Zisserman. ASR is all you need: Cross-modal distillation for lip reading. In *Proceeding of ICASSP*, pages 2143–2147, 2020. 2

[3] T. Chen, J. Frankle, S. Chang, S. Liu, et al. The lottery ticket hypothesis for pre-trained bert networks. *in Proceedings of NeurIPS*, 33:15834–15846, 2020. 1

[4] S. Gui, H. Wang, H. Yang, C. Yu, Z. Wang, and J. Liu. Model compression with adversarial robustness: A unified optimization framework. *in Proceedings of NeurIPS*, 32, 2019. 2

[5] A. Haliassos, P. Ma, R. Mira, S. Petridis, and M. Pantic. Jointly learning visual and auditory speech representations from raw data. *arXiv preprint arXiv:2212.06246*, 2022. 2

[6] S. Han, J. Pool, J. Tran, and W. Dally. Learning both weights and connections for efficient neural network. *in Proceedings of NeurIPS*, 28:1135–1143, 2015. 1

[7] P. Ma, A. Haliassos, A. Fernandez-Lopez, H. Chen, S. Petridis, and M. Pantic. Auto-AVSR: Audio-visual speech recognition with automatic labels. *in Proceedings of ICASSP*, 2023. 2

[8] P. Ma, B. Martinez, S. Petridis, and M. Pantic. Towards practical lipreading with distilled and efficient models. In *Proceedings of ICASSP*, pages 7608–7612, 2021. 2

[9] P. Ma, S. Petridis, and M. Pantic. End-to-end audio-visual speech recognition with conformers. In *Proceedings of ICASSP*, pages 7613–7617, 2021. 2

[10] P. Ma, S. Petridis, and M. Pantic. Visual Speech Recognition for Multiple Languages in the Wild. *Nature Machine Intelligence*, pages 930–939, 2022. 1

[11] T. Makino, H. Liao, Y. Assael, B. Shillingford, et al. Recurrent neural network transducer for audio-visual speech recognition. In *Proceedings of ASRU Workshop*, pages 905–912, 2019. 2

[12] A. Renda, J. Frankle, and M. Carbin. Comparing rewinding and fine-tuning in neural network pruning. *in Proceedings of ICLR*, 2020. 1, 2

[13] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *in Proceedings of ICLR*, 2022. 2