LXMERT Model Compression for Visual Question Answering

Maryam Hashemi **Ghazaleh Mahmoudi*** **Sara Kodeiri*** **Hadi Sheikhi*** **Sauleh Eetemadi**
Iran University of Science and Technology
{m_hashemi94, gh_mahmoodi, sara_kodeiri, ha_sheikhi}@comp.iust.ac.ir, sauleh@iust.ac.ir

Abstract

Large-scale pretrained models such as LXMERT are becoming popular for learning cross-modal representations on text-image pairs for vision-language tasks. According to the lottery ticket hypothesis, NLP and computer vision models contain smaller subnetworks capable of being trained in isolation to full performance. In this paper, we combine these observations to evaluate whether such trainable subnetworks exist in LXMERT when fine-tuned on the VQA task. In addition, we perform a model size cost-benefit analysis by investigating how much pruning can be done without significant loss in accuracy. Our experiment results demonstrate that LXMERT can be effectively pruned by 40%-60% in size with 3% loss in accuracy.

1 Introduction and Related Work

Over the past few years, many single-modal pretrained models have been proposed. Inspired by this, the vision-and-language pretraining seeks to learn joint representations using visual and textual content to improve the efficiency of vision-language tasks.

Both single-modality and cross-modality pretrained models often have hundreds of millions of parameters. Unfortunately, training these over-parametrized models can be prohibitively time-consuming and costly, making them impractical for resource-limited devices. However, cross-modality pretrained models suffer more from the increased model size due to the higher input space dimension. With the task of Visual Question Answering (VQA) (Antol et al., 2015) in mind, and its ultimate goal of being helpful to the visually impaired, decreasing V+L model size makes it feasible to use them in limited-resource devices.

To address this problem, model compression techniques such as pruning have been developed. Deep Learning recently enjoyed welcoming a new powerful pruning method: The Lottery Ticket Hypothesis (LTH) (Frankle and Carbin, 2019). LTH has been shown great success in various fields. It could be a powerful tool to understand the parameter redundancy in the current pretrained V+L models. Thus, we aim to apply LTH to LXMERT(Tan and Bansal, 2020), one of the best-performing two-stream V+L models, to fill this gap. We evaluate our work on VQA (Antol et al., 2015) and compare it with DistillVLM(Fang et al., 2021), which leverages the knowledge distillation technique to compress large visual-linguistic models.

Similar to this work, Gan et al. (2021) study LTH for UNITER(Chen et al., 2020). However, UNITER is a single stream V+L model, and LXMERT is a two-stream model; our results are consistent with theirs.

2 Methodology

In this section, we briefly explain the LXMERT architecture and LTH. Then, we describe how we use LTH to compress the pretrained LXMERT model.

LXMERT is a Transformer-based model which takes two inputs: image and text. Internally, LXMERT consists of two types of encoders: single-modality encoders for each modality and a cross-modality encoder using bidirectional cross attention to exchange information and align entities across the modalities.

The Lottery Ticket Hypothesis (Frankle and Carbin, 2019) shows that by preserving the original weight initializations from the unpruned network, you can train a network with the topology of the pruned network and achieve the same or better test accuracy within the same number of training iterations.

In order to apply LTH to the LXMERT model, we use iterative magnitude pruning. Therefore, we fine-tune LXMERT on the VQA task and iteratively prune 10% of the lowest magnitude weights across the entire model, excluding embedding and output layers. We keep pruning until our model

*These authors contributed equally.

Method	test-dev				test-std			
	Yes/No	Number	other	Overall	Yes/No	Number	other	Overall
DistillVLM	-	-	-	69.6	-	-	-	69.8
LXMERT	88.24	54.45	63.05	72.45	88.29	54.37	63.18	72.63
LXMERT (low-magnitude)	86.95 ±0.95	52.60 ±1.87	60.96 ±1.76	70.72 ±1.44	87.07 ±1.12	52.28 ±1.66	61.02 ±1.83	70.87 ±1.51
LXMERT (high-magnitude)	74.11 ±0.91	42.81 ±1.36	50.5 ±0.19	59.35 ±0.61	74.23 ±0.81	42.99 ±0.87	50.71 ±0.26	59.62 ±0.55
LXMERT (random)	69.26 ±0.29	39.84 ±0.93	45.96 ±0.83	54.86 ±0.52	69.27 ±0.18	40.34 ±0.66	46.33 ±0.79	55.19 ±0.45

Table 1: Performance of subnetworks at 50% weights pruning on VQA v2, which reported for both test-dev and test-std. Test-dev is used for debugging and validation experiments. Test-standard is the default test data for the VQA competition. We test each experiment for three different seeds and report the mean and standard deviation of VQA accuracy across three seeds.

loses roughly half the weights. We use the default settings and hyperparameters of LXMERT (Tan and Bansal, 2020) to finetune on the VQA v2.0 dataset.

3 Experimental Setups and Results

The experiments are designed to investigate the effectiveness and stability of LTH on LXMERT in addition to cost-benefit analysis of the number of parameters in the model. We conduct experiments on the widely-used VQA v2.0 (Goyal et al., 2017) dataset built based on the MS-COCO (Lin et al., 2014) image corpus.

3.1 Effectiveness and Stability

The following steps are performed to compress the LXMERT model.

1. The pretrained LXMERT model plus the VQA classifier's randomly initialized weights are saved.
2. The model is fine-tuned on the 3,129 most frequent answers in the VQA v2.0 dataset.
3. Iterative magnitude pruning is applied to find the low-magnitude subnetwork (pruning 50% of the low-magnitude weights). The high-magnitude subnetwork is computed as a compliment of the low-magnitude subnetwork with equal size. A random subnetwork with an equal size is generated for comparison.
4. The saved weights are restored for all three subnetworks.
5. The high-magnitude, low-magnitude, and random subnetworks are fine-tuned and evaluated on the VQA Task using three different seeds for initializing the VQA model to ensure the stability of the results.

Results of subnetworks at 50% weights pruning on VQA v2.0 are summarized in Table 1 where DistillVLM (Fang et al., 2021) is also listed for comparison. Row 2 to row 5 reports respectively full finetuned LXMERT, low-magnitude, high-magnitude, and random subnetworks. Low-magnitude pruning achieves 97% of full finetuned LXMERT accuracy in overall for both test-dev and test-std and shows marginal improvement over DistillVLM as the baseline. By comparing performance across the subnetworks, random and high magnitude subnetworks perform far worse than low-magnitude subnetwork. Surprisingly, the results demonstrate high-magnitude subnetwork performing better than random subnetwork. This could be a LXMERT specific phenomenon and required further investigation.

3.2 Cost-Benefit Analysis

We experiment with low-magnitude subnetwork by pruning 10% of the weights all the way up to 90% of the weights in 10% increments. Accuracy of these pruned models on VQA v2.0 are reported in Figure 1. Our results indicate a significant loss of accuracy after 50% to 60% pruning.

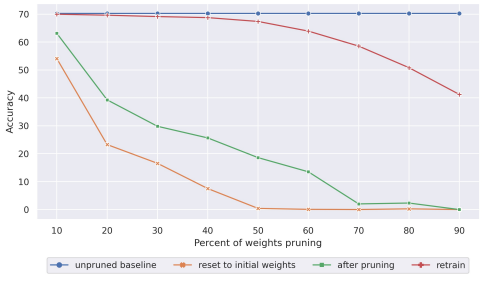

Figure 1: Model size cost-benefit analysis.

4 Conclusion

We confirm that LTH pruning is an effective method for pruning V+L pretrained models. We mainly focused on LXMERT, a two-stream V+L pretrained model, but our findings are consistent with Gan et al. (2021)'s results while using UNITER, a single-stream V+L pretrained model.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 2015 Inter, pages 2425–2433.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2021. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1428–1438.

Jonathan Frankle and Michael Carbin. 2019. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*.

Zhe Gan, Yen-Chun Chen, Linjie Li, Tianlong Chen, Yu Cheng, Shuohang Wang, and Jingjing Liu. 2021. Playing lottery tickets with vision and language. *arXiv preprint arXiv:2104.11832*.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Hao Tan and Mohit Bansal. 2020. LXMert: Learning cross-modality encoder representations from transformers. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*, pages 5100–5111.