Supplementary Materials for the Paper "Neural Architecture Optimization"

[1]**Renqian Luo**[†,*] [2]**Fei Tian**[†] [2]**Tao Qin,** [1]**Enhong Chen,** [2]**Tie-Yan Liu**
[1]University of Science and Technology of China, Hefei, China
[2]Microsoft Research, Beijing, China
[1]lrq@mail.ustc.edu.cn, cheneh@ustc.edu.cn
[2]{fetia, taoqin, tie-yan.liu}@microsoft.com

1 Details of Model Training on Image Classification and Language Modelling

CIFAR-10 contains $50k$ and $10k$ images for training and testing. We randomly choose 5000 images within training set as the dev set for measuring the performance of each candidate network in the optimization process of NAO. Standard data pre-processing and augmentation, such as whitening, randomly cropping 32×32 patches from unsampled images of size 40×40, and randomly flipping images horizontally are applied to original training set. The CNN models are trained using SGD with momentum set to 0.9, where the arrange of learning rate follows a single period cosine schedule with $l_{max} = 0.024$ proposed in [1]. For the purpose of regularization, We apply stochastic drop-connect on each path, and an l_2 weight decay of 5×10^{-4}. All the models are trained with batch size 128.

Penn Treebank (PTB) [2] is one of the most widely adopted benchmark dataset for language modeling task. We use the open-source code of ENAS [3] released by the authors and exactly follow their setups. Specifically, we apply variational dropout, l_2 regularization with weight decay of 5×10^{-7}, and tying word embeddings and softmax weights [4]. We train the models using SGD with an initial learning rate of 10.0, decayed by a factor of 0.9991 after every epoch staring at epoch 15. To avoid gradient explosion, we clip the norm of gradient with the threshold value 0.25. For WikiText-2, we use embdedding size 700, weight decay of 5×10^{-7} and variational dropout 0.15. Others unstated are the same as in PTB, such as weight tying.

2 Search Space

In searching convolutional cell architectures without weight sharing, following previous works of [5, 6], we adopt 11 possible ops as follow:

- identity
- 1×1 convolution
- 3×3 convolution
- $1 \times 3 + 3 \times 1$ convolution
- $1 \times 7 + 7 \times 1$ convolution
- 2×2 max pooling
- 3×3 max pooling
- 5×5 max pooling

[*]The work was done when the first author was an intern at Microsoft Research Asia.
[†]The first two authors contribute equally to this work.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

- 2×2 average pooling
- 3×3 average pooling
- 5×5 average pooling

When using weight sharing, we use exactly the same 5 operators as [3]:

- identity
- 3×3 separable convolution
- 5×5 separable convolution
- 3×3 average pooling
- 3×3 max pooling

In searching for recurrent cell architectures, we exactly follow the search space of ENAS [3], where possible activation functions are:

- tanh
- relu
- identity
- sigmoid

3 Best Architecture discovered

Here we plot the best architecture of CNN cells discovered by our NAO algorithm in Fig. 1.

Furthermore we plot the best architecture of recurrent cell discovered by our NAO algorithm in Fig. 2.

References

[1] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016.

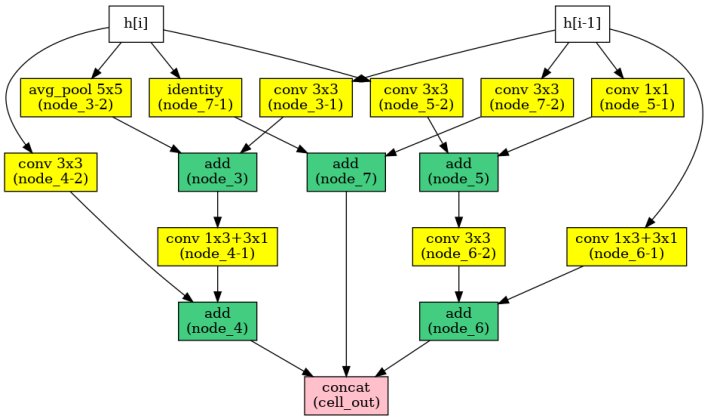
[2] Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. The penn treebank: annotating predicate argument structure. In *Proceedings of the workshop on Human Language Technology*, pages 114–119. Association for Computational Linguistics, 1994.

[3] Hieu Pham, Melody Y Guan, Barret Zoph, Quoc V Le, and Jeff Dean. Efficient neural architecture search via parameter sharing. *arXiv preprint arXiv:1802.03268*, 2018.

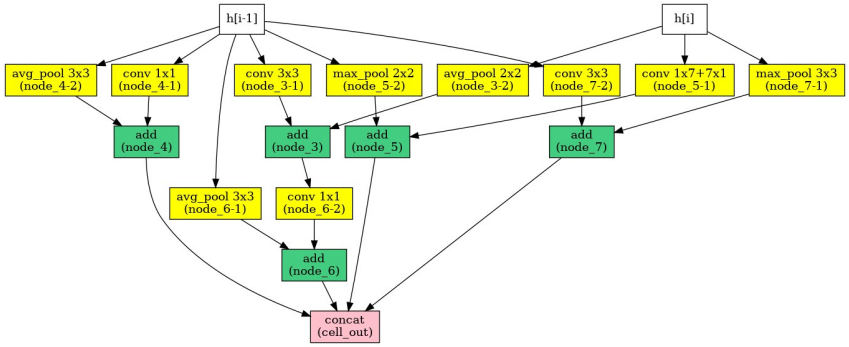
[4] Ofir Press and Lior Wolf. Using the output embedding to improve language models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 157–163. Association for Computational Linguistics, 2017.

[5] Esteban Real, Alok Aggarwal, Yanping Huang, and Quoc V Le. Regularized evolution for image classifier architecture search. *arXiv preprint arXiv:1802.01548*, 2018.

[6] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018.



(a) Normal Cell



(b) Reduction Cell

Figure 1: Basic NAONet building blocks. NAONet normal cell (left) and reduction cell (right).

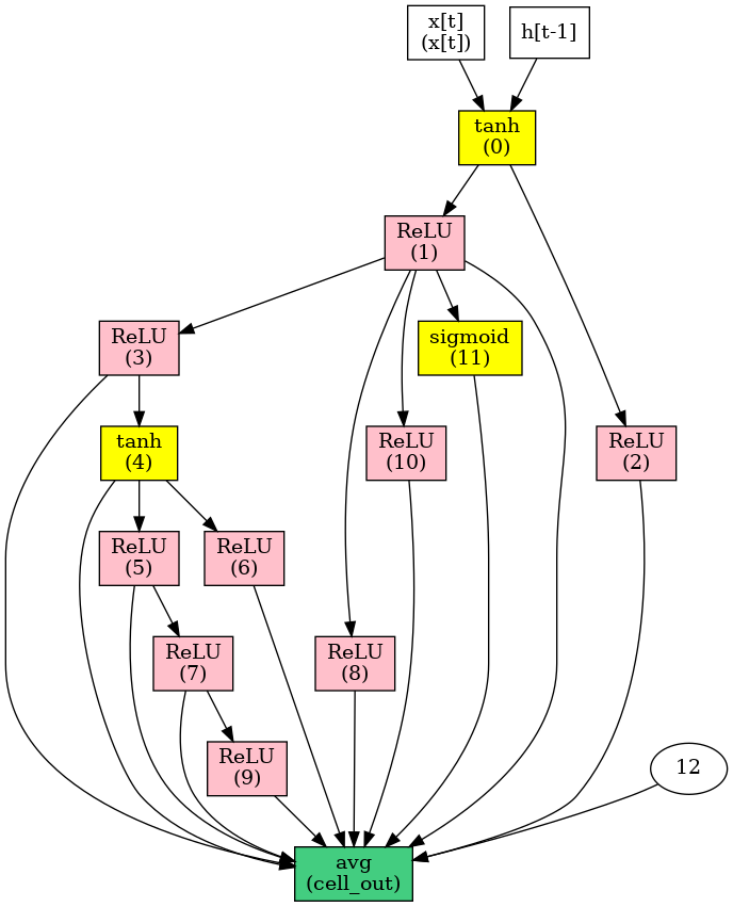


Figure 2: Best RNN Cell discovered by NAO for Penn Treebank.