Can Local Binary Convolutions Make Neural Networks Models Smaller?

Background

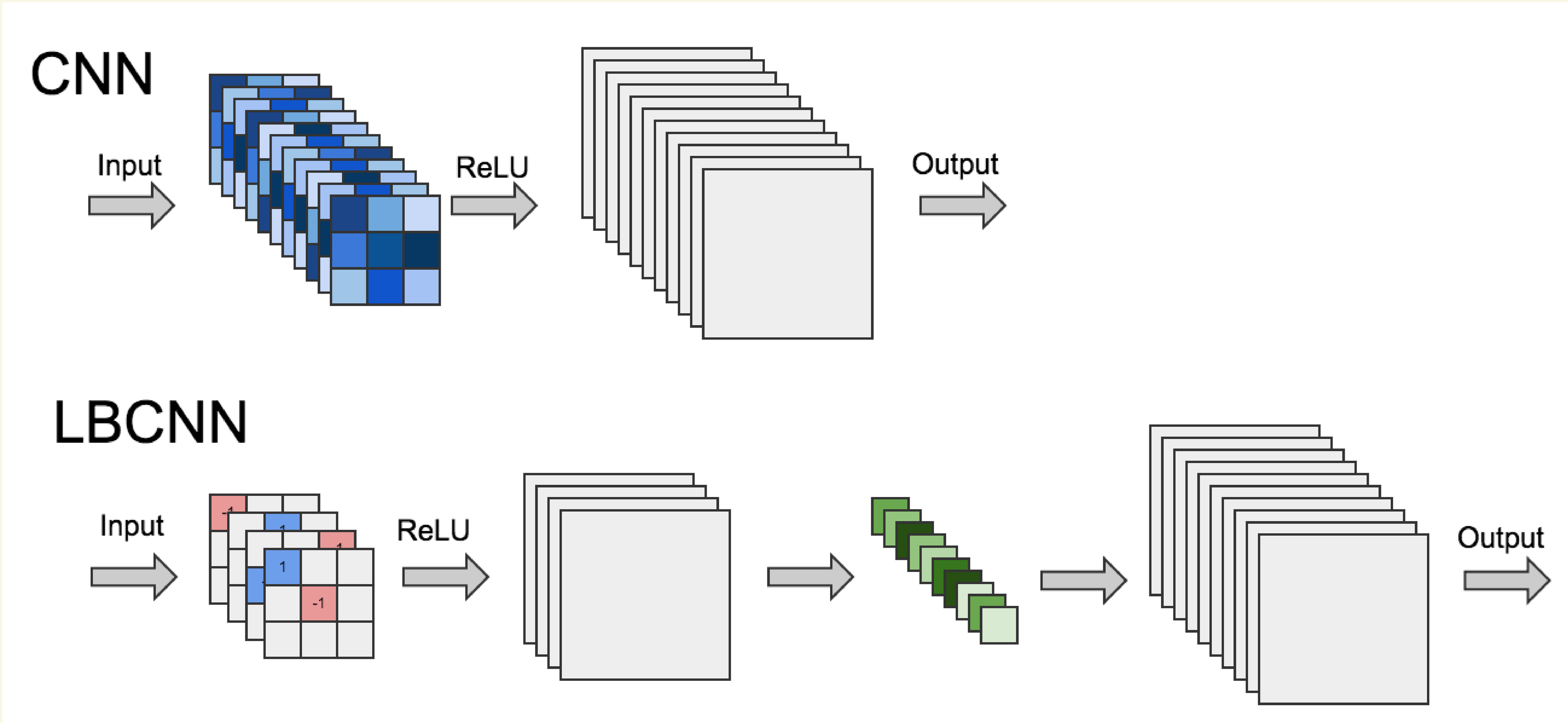

Figure 1: LBCNN construction.

- Local Binary Convolution Neural Networks(LBCNN) are models that use Local Binary Convolution(LBC) layers instead of normal convolution layers.
- Figure 1 shows the architecture of an LBC layer. LBC is a combination of 3×3 weight anchored convolution layer (-1,1,0), ReLU activation, and 1×1 learned convolution layer.
- The number of parameters:

$$\frac{Param.\ of\ CNN}{Pram.\ of\ LBCNN} = \frac{p \times h \times w \times q}{m \times q} = \frac{p \times h \times w}{m} \quad (1)$$

*(p,input channel; q,output channel; h and w,kernel size)

- Under the assumption of $m = p$, LBC can save up to $h \times w$ parameters compared to normal convolution.

Problems

- Juefei-Xu et al. [1]discussed that for every output of normal convolution layer d, there exists a vector v that makes output of LBC layer d' equals to d:

$$d' = v * \sigma_{sigmoid}(B * X) \in \mathbb{R}^{(H \times W) \times 1} = \sigma_{relu}(w * X) \in \mathbb{R}^{(H \times W) \times 1} = d \quad (2)$$

*(B,Local Binary Filters; X,input; v,1×1 filters)

- Only when $m > rank(X^{(H \times W) \times m})$ the corresponding v may exist.
- This m may always be very large, there may exist problems when directly applying LBCNN.

Proposed Method

- We proposed a half normal half LBC architecture as a trade off between accuracy and number of parameters.

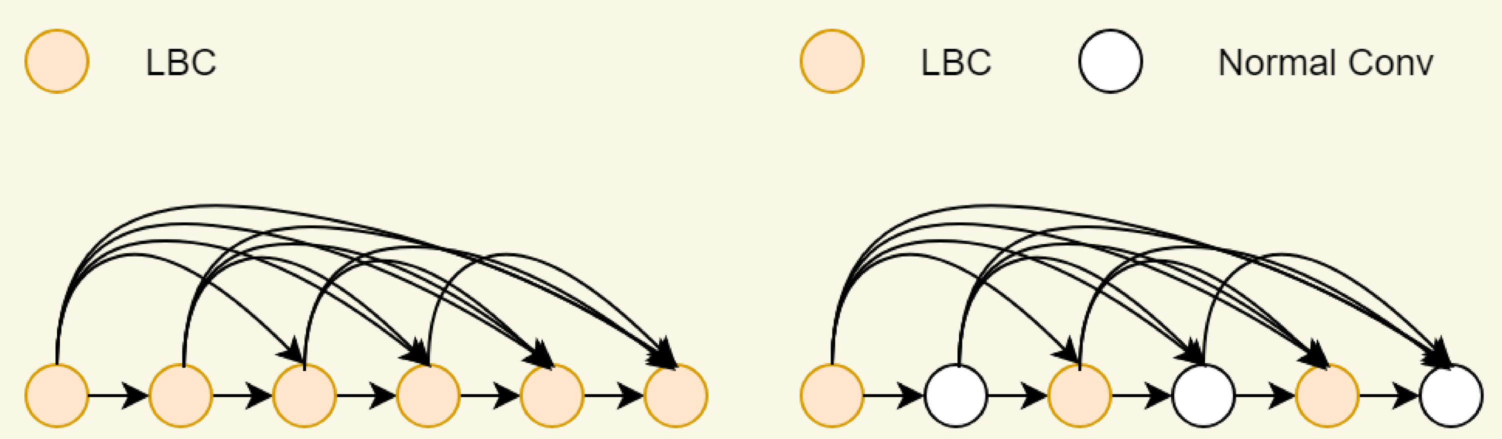

Figure 2: Basic block of DenseNet full (left) and fused (right).

- Take Densely Connected Convolution Networks(DenseNet)[2] as an example, as is shown in Figure 2.

Experiment & Results

- We test those models on ImageNet dataset.

Table 1: Accuracy of Models on ImageNet

Model	Top1	Top5	Learned Para.
DenseNet-121	75.46	92.74	6.8M
Full	69.73	89.21	4.7M
Fused	73.70	91.63	5.8M

- We design different series of Models based on DenseNet called full-m and fused-m models (where m indicates the number of LBC filters).
- Adjust m to change amount of learned parameters:

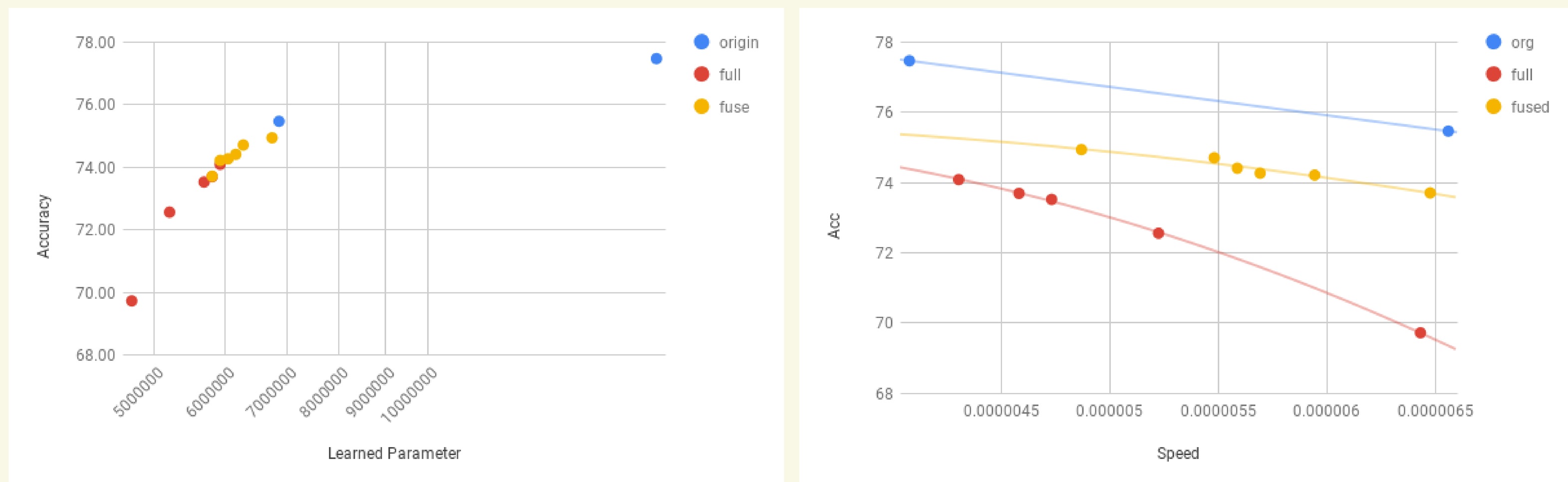

Figure 3: Relationship of Accuracy and Number of parameter (left), Relation ship of Accuracy and Performance (right) .

- According to these results, increasing m helps, while also increasing the number of parameters.
- Additional LBC layers makes model slower.

Conclusion

- We extend the method of LBCNN to larger model on ImageNet to see if it helps.
- We analyzed the disadvantage of LBCNN and proposed a half normal half LBC architecture as a trade off solution.
- LBCNN can't get better results in more complex models and this method seems effect the training speed a lot.
- To conclude, it is possible to get similar results with 3×3 convolution versus LBCNN method, while degrading performance.

Future work

- Test LBCNN on more datasets and more models.
- Test LBCNN on more tasks besides image classification such as semetic segmentation.
- Find some better approximate of normal convolution besides LBCNN.

References

Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Local Binary Convolutional Neural Networks. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2261–2269. IEEE, 2017.

Haoyu Zhang[1],Mohamed Wahib[2],Peng Chen[1 2], Satoshi Matsuoka[1 3]

zhang.h.am@m.titech.ac.jp,mohamed.attia@aist.go.jp,chen.p.aa@m.titech.ac.jp,matsu@is.titech.ac.jp

[1]Tokyo Institute of Technology, Dept. of Mathematical and Computing Science, Tokyo, Japan
[2]AIST-Tokyo Tech Real World Big-Data Computation Open Innovation Laboratory
[3]RIKEN Center for Computational Science,Hyogo,Japan