ResNet

May 30, 2019

This article is written as I read "Deep Residual Learning for Image Recognition" by He et al. [3]. This architecture in this paper won the ILSVRC 2015 classification task and is used as building blocks by many subsequent papers.

1 Introduction

- Theoretically, deeper networks have more capacity than shallower ones. This is because the deeper layers can implement the identify function.
- In practice however, this is not the case. It is often observed deeper networks have higher training losses. This is called the *degradation problem*.
- How this happens is as follows. You start training. The training loss drops. It gets saturated. Then, it degrades rapidly.
- The degradation problem is not caused by overfitting. Otherwise, the training loss would have become lower as we increase the depth.
- You may think this is because deeper networks suffer from with vanishing/exploding gradients. However, this has been solved by careful initialization (for examples, Xavier [1] and He [4]) and batch normalization [5].
- This kind of means that deeper network is just too hard to train. The optimizers we have at hand have a hard time making the deeper layers into the identify mapping or something better.
- To solve the degradation, the problem proposes the following *deep residual learning* framework:

  > Supposed the desired underlying mapping is $\mathcal{H}(\mathbf{x})$, we let the network fits $\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}$ instead.

  The paper says this is easier to optimize. If we want $\mathcal{H}$ to be the identity mapping, it would be easier to push $\mathcal{F}(\mathbf{x})$ than to fit a network to the identity function.
- The mapping $\mathcal{H}(\mathbf{x}) = \mathcal{F}(\mathbf{x}) + \mathbf{x}$ can be implemented by adding *shortcut connection* that bypass the layers that implement $\mathcal{F}$. The output of the shortcut connection is added directly to the output of $\mathcal{F}(\mathbf{x})$.
- The paper showed that, for the ImageNet and the CIFAR-10 datasets, the degradation problem exist in plain networks without shortcut connections. Moreover, when shortcut connections are added, the opposite outcome is true: deeper networks achieve better training losses than shallower ones.
- The shortcut connection trick enables the authors to train a 152-layer network for ImageNet and won the ILSVRC 2015 classification competition and various others.

2 Deep Residual Learning

- Let $\mathcal{H}(\mathbf{x})$ be an underlying mapping to be learned by a few layers.
- Rather than let the network learn $\mathcal{H}(\mathbf{x})$ directly, the paper let the network learn the residual function $\mathcal{F}(\mathbf{x}) := \mathcal{H}(\mathbf{x}) - \mathbf{x}$. The original function becomes $\mathcal{H}(\mathbf{x}) = \mathcal{F}(\mathbf{x}) + \mathbf{x}$.
- The transformation is motivated by the fact that solvers may have a hard time making $\mathcal{H}$ approximate theh identity function. On the other hand, it should be easier for it to drive the weights of the layers down to 0 to make $\mathcal{F}(\mathbf{x})$ close to zero.
- The transformation above is employed at every few layers.
- The paper does so every two layers. That is:

  $$\mathcal{F}(\mathbf{x}) = \mathbf{b}_2 + W_2\sigma(\mathbf{b}_1 + W_1\mathbf{x})$$

  where W_1 and W_2 denote weight matrices, $\mathbf{b}_1$ and $\mathbf{b}_2$ denote the bias vectors, and σ denotes a non-linear function, which is ReLU in the paper. Note that we phrase the W_i, $\mathbf{b}_i$ combo as a fully connected layer, but this can be a convolutional layer as well.
- The operation $\mathcal{F}(\mathbf{x}) + \mathbf{x}$ is performed by a shortcut connection from the input.
- The paper actually applies a second ReLU to the block. That is, in the end, the block ends up computing:

  $$\mathbf{y} = \sigma(\mathbf{x} + \mathbf{b}_1 + W_1\sigma(\mathbf{b}_2 + W_2\mathbf{x})).$$

- However, Gross and Wilber suggested that removing the second ReLU actually leads to small improve in test performance [2].
- Batch normalization layers is typically placed after each affine layer. However, Gross and Wilber observed that putting a batch normalization layer after the addition with the input actually hurts test performance on the CIFAR dataset.
- In the construction so far, the dimension of $\mathcal{F}(\mathbf{x})$ must match that of $\mathbf{x}$. If this is not the case, we can perform a linear projection W_s to make the dimension match:

  $$\mathcal{F}(\mathbf{x}) + W_s\mathbf{x}.$$

- The paper details a 34-layer residual network for ImageNet classification with the following details:
  - Most convolution layers have kernel size of 3×3.
  - Most convolution layers preserve the input size.
  - Image size is halved and channels doubled every 6 layers. Downsampling is done by a convolutional layer with stride 2.
  - Shortcut connections skip two convolution layers. When they skip to a downsampled version, the projection W_s is a convolution with stride 2. The paper consider two options in making the number of channel matches:
    * W_s does not increase the number of channels. Instead, 0 are appended to make the number of channels match.
    * W_s does not increase the number of channels. A 1×1 convolution is performed afterwards to make the channel match.

  The paper found that the second option is slightly better than the first one.

References

[1] Glorot, X., and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *JMLR W&CP: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics (AISTATS 2010)* (May 2010), vol. 9, pp. 249–256.

[2] Gross, S., and Wilber, K. Training and investigating residual nets.

[3] He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *CoRR abs/1512.03385* (2015).

[4] He, K., Zhang, X., Ren, S., and Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *CoRR abs/1502.01852* (2015).

[5] Ioffe, S., and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR abs/1502.03167* (2015).