Lipschitz Constraint and Spectral Norm in Deep Learning

Jiaru Zhang

2020.7.5

Introduction

Recall: the sufficient condition for invertible ResNet

Theorem

Let

$$F(x) = x + g(x) \tag{1}$$

be a residual layer, then it is invertible if the **Lipschitz constant** $\mathrm{Lip}\,(g)$ *satisfies*

$$\mathrm{Lip}\,(g) < 1, \tag{2}$$

The invertible ResNet is implemented by enforcing the **spectral norm** $\|W_i\|_2 < 1$ for each layer i.

What are they? Any other applications?

* Jens Behrmann, Will Grathwohl, Ricky T. Q. Chen, David Duvenaud, Joern-Henrik Jacobsen. Invertible Residual Networks. In ICML, 2019

Lipschitz Constraint

Consider a model f_w mapping x into y:

$$y = f_w(x) \tag{3}$$

We hope it is insensitive to the perturbation of the input. For a perturbation ξ, we want

$$\|f_w(x+\xi) - f_w(x)\| \tag{4}$$

to be small. So we introduce the **Lipschitz Constraint** here:

Lipschitz Constraint

For a function f_w, if $\exists C(w)$, $\forall x, \xi$, we have

$$\|f_w(x+\xi) - f_w(x)\| \leq C(w) \cdot \|\xi\|, \tag{5}$$

then f_w is Lipschitz continuous or satisfies Lipschitz constraint.

Lipschitz Constraint in Neural Networks

- We have seen insensitivity means Lipschitz continuity. Additionally, we hope $C(w)$ as small as possible for a model f_w.
- Consider a single layer in a neural network:

$$f_w(x) = f(Wx + b), \tag{6}$$

  where W and b are parameter matrix and vector, $f(\cdot)$ is the activation function.
- If ξ is small enough,

$$\begin{aligned} \|f_w(x+\xi) - f_w(x)\| &= \|f(W(x+\xi)+b) - f(Wx+b)\| && (7) \\ &= \left\|\frac{\partial f}{\partial x} W\xi\right\| && (8) \\ &:\leq C(w) \cdot \|\xi\| && (9) \end{aligned}$$

- For popular activation functions (e.g. relu, sigmoid, tanh, ...), $\left\|\frac{\partial f}{\partial x}\right\|$ are all bounded.

Lipschitz Constraint in Neural Networks

- So we only need to maintain

$$\|W\xi\| \leq C(W) \cdot \|\xi\|, \tag{10}$$

  and answer the question: **What is the smallest C ?**

- Now we introduce the definition of **Spectral Norm**:

Definition (Spectral Norm)

For a matrix W, we define its Spectral Norm as

$$\|W\|_2 = \max_{\xi \neq 0} \frac{\|W\xi\|}{\|\xi\|}. \tag{11}$$

Note that it is a generalization of l_2 norm for vectors.

- Now we can write the formula (10) as

$$\|W\xi\| \leq \|W\|_2 \cdot \|\xi\|. \tag{12}$$

Relationship with l_2 regularization

Consider the Frobenius norm of a matrix, which is easier to compute:

$$\|W\|_F = \sqrt{\sum_{i,j} w_{ij}^2}. \tag{13}$$

By Cauchty inequality, we have

$$\|Wx\| \leq \|W\|_F \cdot \|x\|. \tag{14}$$

Therefore,

1. The Frobenius norm also satisfies the formula (10). It is a looser bound.
2. We can add it as a regularization term into loss, i.e.,

$$loss = loss(y, f_x(x)) + \lambda \|W\|_F^2 \tag{15}$$

   This is exactly the l_2 regularization!

Spectral norm regularization

We can use power iteration[1] method to estimate it:

Power Iteration

Repeat sufficient times:

1. $u = Wv$
2. $v = W^T u$
3. $\sigma = \frac{\|u\|}{\|v\|}$

Return σuv. Note that one iteration is adequate in the experiments.

Using it as a regularization item, paper[2] proposes the spectral norm regularization.

1 https://en.wikipedia.org/wiki/Power_iteration

2 Yoshida, Yuichi and Miyato, Takeru. Spectral norm regularization for improving the generalizability of deep learning. arXiv preprint arXiv:1705.10941, 2017.

Enforcing Lipschitz Continuity

- Spectral norm regularization penalizes the spectral norm by adding explicit regularization term.
- What about manually adjusting the parameters?
- A new method[3]

$$W := \frac{1}{\max\left(1, \frac{\|W\|_2}{\lambda}\right)} W, \tag{16}$$

where λ is the expected bound of the Lipschitz constant.

- W is replaced as formula (16) shows in each updating
- Further, it is used in convolution layers by convolutional power iteration[4].
- i-ResNet uses similar method.

---

[3] Gouk, H., Frank, E., Pfahringer, B., and Cree, M. Regularisation of neural networks by enforcing lipschitz continuity. arXiv preprint arXiv:1804.04368, 2018.

[4] Farzan Farnia, Jesse Zhang, David Tse. Generalizable Adversarial Training via Spectral Normalization. In ICLR, 2019

Spectral Normalization

- Spectral normalization[5] normalizes the spectral norm so that it satisfies the Lipschitz constraint $C(w) = 1$:

$$W := \frac{W}{\|W\|_2} \tag{17}$$

- The method is used in training GANs to stabilize the training of the discriminator.

5 Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In ICLR, 2018.

Conclusion

- Restriction on Spectral norm is a popular regularisation method to make models insensitive.
- It seems that the performance decreasing of i-ResNet is a sudden.