Deep Learning

Initializers

Edgar Roman-Rangel.
edgar.roman@itam.mx

Department of Computer Science.
Instituto Tecnológico Autónomo de México, ITAM.

Outline

Introduction

We have already mentioned that parameters in a Neural Network are initialized randomly.

Do you remember why?

We need to avoid co-dependency and we want to add randomness to the exploration of the parameter space.

The magnitude of parameters

It has been noticed, by fellow researchers, that having parameters with “small” magnitude helps avoid vanishing and exploiting gradients.

So, how small is small, and what distribution must we use for the random sampling?

Uniform and Gaussian distributions

$$w_i \sim \mathcal{U}(-1, 1),$$

or

$$w_i \sim \mathcal{N}(0, 1).$$

These are an older approaches, no longer used.

Glorot

a.k.a., Xavier initializer.

Xavier Glorot and Yoshua Bengio. (2010). "Understanding the difficulty of training deep feedforward neural networks". *International Conference on Artificial Intelligence and Statistics*.

$$w_i \sim \mathcal{N}\left(0, \sqrt{\frac{1}{f_a}}\right),$$

where,

$$f_a = \frac{f_i + f_o}{2},$$

and,

- f_i: number of input variables in current layer,
- f_o: number of output variables in current layer.

Often used in layers with linear, logistic, tanh, and softmax non-linear activation functions.

Glorot Uniform

There is also a version by Glorot, using a uniform distribution.

$$w_i \sim \mathcal{U}\left(-\frac{1}{\sqrt{f_i}}, \frac{1}{\sqrt{f_i}}\right).$$

Not so much used in practice.

He initializer

Kaiming He et al. (2015). "Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification". *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

$$w_i \sim \mathcal{N}\left(0, \sqrt{\frac{2}{f_i}}\right).$$

Often used in layers with ReLU and its variants non-linear activation functions.

LeCun initializer

Günter Klambauer et al. (2017). "Self-Normalizing Neural Networks". *Conference on Neural Information Processing Systems*.

$$w_i \sim \mathcal{N}\left(0, \sqrt{\frac{1}{f_i}}\right).$$

Often used in layers with SELU non-linear activation functions.

Q&A

Thank you!

edgar.roman@itam.mx