Deep Learning: Assignment Two

Aditi Nair (asn264) and Akash Shah (ass502)

April 4, 2017

1 Batch Normalization

1. Let $x_1, ...x_n$ be scalar features. Then we define the mean μ_n as:

$$\mu_n = \frac{1}{n}\sum_{i=1}^{n} x_i$$

and the variance σ_n^2 as:

$$\sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_n)^2$$

Now to normalize the feature x_i (for $1 \leq i \leq n$), we compute:

$$\hat{x_i} = \frac{x_i - \mu_n}{\sigma_n}$$

Then for all $\hat{x_i}$, the expected value is 0:

$$\sum_{i=1}^{n}\hat{x_i} = \sum_{i=1}^{n}\frac{x_i - \mu_n}{\sigma_n} = \frac{1}{\sigma_n}\sum_{i=1}^{n}(x_i - \mu_n) = \frac{1}{\sigma_n}\left[\left(\sum_{i=1}^{n} x_i\right) - n\cdot\mu_n\right] = \frac{1}{\sigma_n}\left[\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i\right] = 0$$

Since $\sum_{i=1}^{n}\hat{x_i} = 0$, the expected value $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\hat{x_i}$ is also 0.

Then for all $\hat{x_i}$ the variance is 1 since:

$$\frac{1}{n}\sum_{i=1}^{n}(\hat{x_i} - \hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}(\hat{x_i} - 0)^2 = \frac{1}{n}\sum_{i=1}^{n}\hat{x_i}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \mu_n}{\sigma_n}\right)^2$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - \mu_n)^2}{\sigma_n^2} = \frac{1}{n\cdot\sigma_n^2}\sum_{i=1}^{n}(x_i - \mu_n)^2$$

$$= \frac{n}{n\cdot\sum_{i=1}^{n}(x_i - \mu_n)^2}\sum_{i=1}^{n}(x_i - \mu_n)^2 = 1$$

2. For scalar features $x_1, ...x_n$ the output of the BN module can be written as:

$$y_i = BN_{\gamma,\beta}(x_i) = \gamma\hat{x_i} + \beta$$

with

$$\hat{x_i} = \frac{x_i - \mu_n}{\sqrt{\sigma_n^2 + \epsilon}}$$

μ_n and σ_n are defined as above. For numerical stability, the BN algorithm adds ϵ to σ_n^2 in the denominator before taking the square root.

$$GRADIENT FORMULAS$$

2 Convolution

1.
2.
3.

3 Variants of Pooling

1.
2.
3.

4 t-SNE

1.
2.

5 Sentence Classification

5.1 ConvNet

5.2 RNN

5.3 Extra credit experiments of fastText

5.4 Extra credit question

6 Language Modeling