ASSIGNMENT 2: THEORY OF CNNS AND REGULARIZATION [IFT6135]

JOSEPH D. VIVIANO

1. Convolutions

2. Convolutional Neural Networks

(a) Since we have one zero padding in layer three, so the size of layer 3 is $128 \times 6 \times 6$. The image has 3 color , therefore the last layer is a fully connected layer of the size

$$3.128.6.6 = 13824.$$

.

(b) For the last convolution the size of kernel is 4×4 and we have 128 kernels of 3 colors, so the number of parameters is as follows

$$\#W = 4.4.128.3 = 6144.$$

.

3. Kernel Configurations for CNNs

(a): i : We denote the dimension of input as $W_1 \times H_1$ and the dimension of output is $W_2 \times H_2$, the kernel size is $K \times K$, number of zero padding is P and stride size is S, then we easily can see that

$$W_2 = \frac{W_1 - K + 2P}{S} + 1, \tag{3.1}$$

by using this formula we get

$$32 = \frac{64 - 8 + 2P}{S} + 1$$

. So if we set $P = 3$ and $S = 2$, the this convolution operation works.

ii : If we have dilatation of size D, then

$$W_2 = \frac{W_1 - K + 2P + (W_1 - 1)D}{S} + 1, \tag{3.2}$$

So here we have

$$32 = \frac{64 - K + 2P + 63.6}{2} + 1,$$

Date: March 2018.

. So if we set $K = 400$ and $P = 10$, then our convolution operation works.

(b): For the pooling layer if the the kernel size is $K = 4 \times 4$ with no overlapping which means the stride size is$S = 4$, Then the pooling operation works.

(c): Here we have $K = 8$, $W_1 = 32$ and $S = 4$. By replacing the values in the formula(**??**), we have

$$W_2 = \frac{32 - K}{4} + 1 = 7$$

. So the output is of the size 7×7.

(d): *i* here we have $W_2 = 4$, $W_1 = 8$ and $P = 0$, so by using the formula in (**??**), we get

$$4 = \frac{8 - K + 0}{S} + 1,$$

. By setting $K = 2$ and $S = 2$, the operation will work.

ii : we have $W_2 = 4$, $W_1 = 8$, $P = 2$ and $D = 1$, so by applying (**??**), we obtain

$$4 = \frac{8 - K + 4 + 7}{S} + 1$$

by putting $K = 13$ and $S = 2$, the operation works.

iii : By substituting the values in (**??**), we have

$$4 = \frac{8 - K + 2}{S} + 1,$$

so we choose $K = 4$ and $S = 2$.

4. Dropout as Weight Decay

5. Dropout as a Geometric Ensemble

Consider the case of a single linear layer model with a softmax output. Prove that weight scaling by 0.5 corresponds exactly to the inference of a conditional probability distribution proportional to a geometric mean over all dropout masks.

First, observe the single linear layer with softmax output with n input variables represented by the vector v with dropout mask d:

$$P(y = \mathrm{y}|v; d) = \mathbf{softmax}\left(W^T(d \odot v) + b\right)_y \tag{5.1}$$

and the ensemble conditional probability distribution which represents the geometric mean over all dropout masks:

$$p_{ens}(y = \mathrm{y}|v; d) \propto \left(\prod_{i=1}^{N} \hat{y}_v^{(i)} \right)^{\frac{1}{N}} . \tag{5.2}$$

Aren't they nice? Recall the alternative formulation of the softmax:

$$\textbf{softmax}_i = \frac{e^{x_i}}{\sum_{k=1}^{K} e^{x_k}} \tag{5.3}$$

Which we now rewrite, subbing in our vector representation of the softmax and replacing e^x with $exp(x)$:

$$\textbf{softmax}_y = \frac{exp\left(W_y^T(d \odot v) + b\right)}{\sum_{k=1}^{K} exp\left(W_{y'}^T(d \odot v) + b\right)} \tag{5.4}$$

Now we show that the ensemble predictor is defined by re-normalizing the geometric mean over all the individual ensemble members' predictions:

$$P_{ens}(y = \mathrm{y}|v) = \frac{\tilde{P}_{ens}(y = \mathrm{y}|v)}{\sum y' \tilde{P}_{ens}(y = \mathrm{y'}|v)} \tag{5.5}$$

Where each $\tilde{P}_{ens}$ is the geometric mean over all dropout masks for a single y:

$$\tilde{P}_{ens}(y = \mathrm{y}|v) = \sqrt[2^n]{\prod_{d \in \{0,1\}^n} P(y = \mathrm{y}|v; d)}. \tag{5.6}$$

Now we simply sub in our definition of $softmax$ for P:

$$\tilde{P}_{ens}(y = \mathrm{y}|v) = \sqrt[2^n]{\prod_{d \in \{0,1\}^n} \frac{exp\left(W_y^T(d \odot v) + b\right)}{\sum_{k=1}^{K} exp\left(W_{y'}^T(d \odot v) + b\right)}}. \tag{5.7}$$

Since the denominator is a constant under this normalization scheme we ignore it and simplify:

$$\tilde{P}_{ens}(y = \mathrm{y}|v) \propto 2^n \sqrt{\prod_{d \in \{0,1\}^n} exp\left(W_y^T(d \odot v) + b\right)} \tag{5.8}$$

We convert the product to the sum by taking *exp* of the entire equation:

$$\tilde{P}_{ens}(y = \mathrm{y}|v) \propto exp\left(\frac{1}{2^n} \sum_{d \in \{0,1\}^n} W_y^T(d \odot v) + b\right) \tag{5.9}$$

And finally the sum and exponent n cancel:

$$\tilde{P}_{ens}(y = \mathrm{y}|v) \propto exp\left(\frac{1}{2} W_y^T(d \odot v) + b\right) \tag{5.10}$$

Finally, we sub this back into our earlier formulation of the softmax to show that the weights W are scaled by $\frac{1}{2}$:

$$\mathbf{softmax}_y = \frac{exp\left(\frac{1}{2} W_y^T(d \odot v) + b\right)}{\sum_{k=1}^{K} exp\left(\frac{1}{2} W_{y'}^T(d \odot v) + b\right)} \tag{5.11}$$

Therefore, weight scaling by 0.5 is exactly equivilant to a conditional probability distribution proportional to a geometric mean over all dropout masks.

6. Normalization

Université de Montréal
Email address: joseph@viviano.ca