Regularization of Neural Networks using DropConnect Supplementary Material

1. Preliminaries

Definition 1 (DropConnect Network)**.** *Given data set S with ℓ entries: $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_\ell\}$ with labels $\{y_1, y_2, \ldots, y_\ell\}$, we define DropConnect network as a mixture model:*

$$\mathbf{o} = \sum_m p(M) f(\mathbf{x}; \theta, M) = \mathbf{E}_m \left[f(\mathbf{x}; \theta, M) \right] \quad (1)$$

Each network $f(x; \theta, M)$ has weights $p(M)$ and network parameters are $\theta = \{W_s, W, W_g\}$. W_s are the softmax layer parameters, w are the DropConnect layer parameters and W_g are the feature extractor parameters. Further more, m is the DropConnect layer mask.

Remark 1. *when each element of M_i has equal probability of being on and off ($p = 0.5$), the mixture model has equal weights for all sub-models $f(\mathbf{x}; \theta, M)$, otherwise the mixture model has larger weights in some sub-models than others.*

Reformulate cross-entropy loss on top of soft-max into a single parameter function that combines soft-max output and labels. Same as logistic.

Definition 2 (Logistic Loss)**.** *The following loss function defined on k-class classification is call logistic loss function:*

$$A_y(o) = -\sum_i y_i \ln \frac{\exp o_i}{\sum_j \exp(o_j)} = -o_i + \ln \sum_j \exp(o_j)$$

where y is binary vector with i^{th} bit set on

Lemma 1. *Logistic loss function A has the following properties:*

1. $A_y(0) = \ln k$
2. $-1 \leq A'_y(o) \leq 1$
3. $A''_y(o) \geq 0$.

The first one says $A(0)$ is depend on some constant related with number of labels. The second one says A is Lipschitz function with $L = 1$. The third one says A is a convex function w.r.t x.

Definition 3 (Rademacher complexity)**.** *For a sample $S = \{x_1, \ldots, x_\ell\}$ generated by a distribution D on set X and a real-valued function class $\mathcal{F}$ in domain X, the empirical Rademacher complexity of $\mathcal{F}$ is the random variable:*

$$\hat{R}_\ell(\mathcal{F}) = \mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{\ell} \sum_{i=1}^{\ell} \sigma_i f(x_i) \right| \mid x_1, \ldots, x_\ell \right]$$

where sigma $= \{\sigma_1, \ldots, \sigma_\ell\}$ are independent uniform $\{\pm 1\}$-valued (Rademacher) random variables. The Rademacher complexity of $\mathcal{F}$ is $R_\ell(\mathcal{F}) = \mathbf{E}_S \left[\hat{R}_\ell(\mathcal{F}) \right]$.

Theorem 1 ((Koltchinskii and Panchenko, 2000))**.** *Fix $\delta \in (0,1)$ and let $\mathcal{F}$ be a class of functions mapping from M to $[0,1]$. Let $(M_i)_{i-1}^{\ell}$ be drawn independently according to a probability distribution D. Then with probability at least $1-\delta$ over random draws of samples of size ℓ, every $f \in \mathcal{F}$ satisfies:*

$$\begin{aligned} \mathbf{E}[f(M)] &\leq \hat{\mathbf{E}}[f(M)] + R_\ell(\mathcal{F}) + \sqrt{\frac{\ln(2/\delta)}{2\ell}} \\ &\leq \hat{\mathbf{E}}[f(M)] + \hat{R}_\ell(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}} \end{aligned}$$

2. Bound Derivation

Theorem 2 ((Ledoux and Talagrand, 1991))**.** *Let $\mathcal{F}$ be class of real functions. If $\mathcal{A}$: $\mathbf{R} \to \mathbf{R}$ is Lipschitz with constant L and satisfies $\mathcal{A}(0) = 0$, then $\hat{R}_\ell(\mathcal{A} \circ F) \leq 2L\hat{R}(\mathcal{F})$*

Lemma 2. *Let $\mathcal{F}$ be class of real functions and $\mathcal{H} = [\mathcal{F}_j]_{j=1}^k$ be a k-dimensional function class. If $\mathcal{A}$: $\mathbf{R}^k \to \mathbf{R}$ is a Lipschitz function with constant L and satisfies $\mathcal{A}(0) = 0$, then $\hat{R}_\ell(\mathcal{A} \circ \mathcal{H}) \leq 2kL\hat{R}_\ell(\mathcal{F})$*

Lemma 3 (Classifier Generalization Bound)**.** *Generalization bound of a k-class classifier with logistic loss function is directly related Rademacher complexity of that classifier*

$$\mathbf{E}[A_y(o)] \leq \frac{1}{\ell} \sum_{i=1}^{\ell} A_{y_i}(o_i) + 2k\hat{R}_\ell(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2\ell}}$$

Proof. From Lemma 1, Logistic loss function $(A - c)(x) \in \mathcal{A}$ due to $(A-c)'(x) \leq 1$ and $(A-c)(0) = 0$ with some constant c. By Lemma 2: $\hat{R}_\ell((A-c) \circ \mathcal{F}) \leq 2k\hat{R}_\ell(\mathcal{F})$ □

Lemma 4. *For all neuron activations: sigmoid, tanh and relu, we have:* $\hat{R}_\ell(a \circ \mathcal{F}) \leq 2\hat{R}_\ell(\mathcal{F})$

Lemma 5 (Network Layer Bound)**.** *Let* $\mathcal{G}$ *be the class of real functions* $R^d \to R$ *with input dimension* $\mathcal{F}$*, i.e.* $\mathcal{G} = [\mathcal{F}_j]_{j=1}^d$ *and* $\mathcal{H}_B$ *is a linear transform function parameterized by* W *with* $\|W\|_2 \leq B$*, then* $\hat{R}_\ell(\mathcal{H} \circ \mathcal{G}) \leq \sqrt{d}B\hat{R}_\ell(\mathcal{F})$

Proof.

$$\begin{aligned}
&\hat{R}_\ell(\mathcal{H} \circ \mathcal{G}) \\
= \;& \mathbf{E}_\sigma \left[\sup_{h \in \mathcal{H}, g \in \mathcal{G}} \left| \frac{2}{\ell} \sum_{i=1}^{\ell} \sigma_i h \circ g(x_i) \right| \right] \\
= \;& \mathbf{E}_\sigma \left[\sup_{g \in \mathcal{G}, \|W\| \leq B} \left| \left\langle W, \frac{2}{\ell} \sum_{i=1}^{\ell} \sigma_i g(x_i) \right\rangle \right| \right] \\
\leq \;& B\mathbf{E}_\sigma \left[\sup_{f^j \in \mathcal{F}} \left\| \left[\frac{2}{\ell} \sum_{i=1}^{\ell} \sigma_i^j f^j(x_i) \right]_{j=1}^d \right\| \right] \\
= \;& B\sqrt{d}\mathbf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left| \frac{2}{\ell} \sum_{i=1}^{\ell} \sigma_i f(x_i) \right| \right] = \sqrt{d}B\hat{R}_\ell(\mathcal{F})
\end{aligned}$$

□

Remark 2. *Given a layer in our network, we denote the function of all layers before as* $\mathcal{G} = [\mathcal{F}_j]_{j=1}^d$*. This layer has the linear transformation function* $\mathcal{H}$ *and activation function* a*. By Lemma 4 and Lemma 5, we know the network complexity is bounded by:*

$$\hat{R}_\ell(\mathcal{H} \circ \mathcal{G}) \leq c\sqrt{d}B\hat{R}_\ell(\mathcal{F})$$

where $c = 1$ *for identity neuron and* $c = 2$ *for others.*

Lemma 6. *Let* $\mathcal{F}_M$ *be the class of real functions that depend on* m*, then* $\hat{R}_\ell(\mathbf{E}_M[\mathcal{F}_M]) \leq \mathbf{E}_M\left[\hat{R}_\ell(\mathcal{F}_M)\right]$

Proof.

$$\begin{aligned}
\hat{R}_\ell(\mathbf{E}_M[\mathcal{F}_M]) &= \hat{R}_\ell\left(\sum_M p(M)\mathcal{F}_M\right) \leq \sum_M \hat{R}_\ell(p(M)\mathcal{F}_M) \\
&\leq \sum_M |p(M)|\hat{R}_\ell(\mathcal{F}_M) = \mathbf{E}_M\left[\hat{R}_\ell(\mathcal{F}_M)\right]
\end{aligned}$$

because of common fact: 1) $\hat{R}_\ell(c\mathcal{F}) = |c|\hat{R}_\ell(\mathcal{F})$ and 2) $\hat{R}_\ell(\sum_i \mathcal{F}_i) \leq \sum_i \hat{R}_\ell(\mathcal{F}_i)$ □

Theorem 3 (DropConnect Network Complexity)**.** *Consider the DropConnect neural network defined in Definition 1. Let* $\hat{R}_\ell(\mathcal{G})$ *be the empirical Rademacher complexity of the feature extractor and* $\hat{R}_\ell(\mathcal{F})$ *be the empirical Rademacher complexity of the whole network. In addition, we assume:*

1. *weight parameter of DropConnect layer* $|W| \leq B_h$
2. *weight parameter of* s*, i.e.* $|W_s| \leq B_s$ *(L2-norm of it is bounded by* $\sqrt{dk}B_s$*).*

Then we have:

$$\hat{R}_\ell(\mathcal{F}) \leq p\left(2\sqrt{k}dB_s n\sqrt{d}B_h\right)\hat{R}_\ell(\mathcal{G})$$

Proof.

$$\begin{aligned}
\hat{R}_\ell(\mathcal{F}) &= \hat{R}_\ell(\mathbf{E}_M[f(\mathbf{x};\theta,M)]) \\
&\leq \mathbf{E}_M\left[\hat{R}_\ell(f(\mathbf{x};\theta,M))\right] && (2) \\
&= \mathbf{E}_M\left[\hat{R}_\ell(s \circ a \circ h_m \circ g)\right] \\
&\leq (\sqrt{dk}B_s)\sqrt{d}\mathbf{E}_M\left[\hat{R}_\ell(a \circ h_m \circ g)\right] && (3) \\
&= 2\sqrt{k}dB_s\mathbf{E}_M\left[\hat{R}_\ell(h_m \circ g)\right] && (4)
\end{aligned}$$

where $h_m = (M \circ W)v$. Equation (2) is based on Lemma 6, Equation (3) is based on Lemma 5 and Equation (4) follows from Lemma 4.

$$\begin{aligned}
&\mathbf{E}_M\left[\hat{R}_\ell(h_m \circ g)\right] \\
= \;& \mathbf{E}_{m,\sigma}\left[\sup_{h \in \mathcal{H}, g \in \mathcal{G}} \left| \frac{2}{\ell}\sum_{i=1}^{\ell} \sigma_i w^T D_M g(x_i) \right| \right] && (5) \\
= \;& \mathbf{E}_{m,\sigma}\left[\sup_{h \in \mathcal{H}, g \in \mathcal{G}} \left| \left\langle D_M w, \frac{2}{\ell}\sum_{i=1}^{\ell} \sigma_i g(x_i) \right\rangle \right| \right] \\
\leq \;& \mathbf{E}_M\left[\max_w \|D_M w\|\right]\mathbf{E}_\sigma\left[\sup_{g^j \in \mathcal{G}} \left\| \left[\frac{2}{\ell}\sum_{i=1}^{\ell} \sigma_i g^j(x_i)\right]_{j=1}^n \right\| \right] && (6) \\
\leq \;& B_h p\sqrt{nd}\left(\sqrt{n}\hat{R}_\ell(\mathcal{G})\right) = pn\sqrt{d}B_h\hat{R}_\ell(\mathcal{G})
\end{aligned}$$

where D_M in Equation (5) is an diagonal matrix with diagonal elements equal to m and inner product properties lead to Equation (6). Thus, we have

$$\hat{R}_\ell(\mathcal{F}) \leq p\left(2\sqrt{k}dB_s n\sqrt{d}B_h\right)\hat{R}_\ell(\mathcal{G})$$

□

Remark 3. *Theorem 3 implies that p is an additional regularizer we have added to network when we convert a normal neural network to a network with DropConnect layers. Consider the following extreme cases:*

1. $p = 0$*: the network generalization bound equals to 0, which is true because classifier does not depends on input any more*
2. $p = 1$*: reduce to normal network*

Symbol	Description	Related Formula
y	Data Label, can either be integer label for bit vector(depends on context)	
x	Network input data	
$g(.)$	Feature extractor function with parameter W_g	
v	Feature extractor network output	$v = g(x, W_g)$
M	DropConnect connection information parameter (weight mask)	
$h(.)$	DropConnect transformation function with parameter W, M	
u	DropConnect output	$u = h(v; W, M)$
$a(.)$	DropConnect activation function	
r	DropConnect after activation	$r = a(u)$
$s(.)$	Dimension reduction layer function with parameter W_s	
o	Dimension reduction layer output (network output)	$o = s(r; W_s)$
θ	All parameter of network expect weight mask	$\theta = \{W_s, W, W_g\}$
$f(.)$	Overall classifier(network) output	$o = f(x; \theta, M)$
λ	Weight penalty	
$A(.)$	Data Loss Function	$A(o - y)$
$L(.)$	Over all objective function	$L(x, y) = \sum_i A(o_i - y_i) + 1/2\lambda\|W\|_2^2$
n	Dimension of feature extractor output	
d	Dimension of DropConnect layer output	
k	number of class	$dim(y) = k$

Table 1. Symbol Table

References

V. Koltchinskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30:2002, 2000.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces*. Springer, New York, 1991.