CS 6.5: Theories of Deep Learning
Problem Sheet 4

Prof. Jared Tanner

November 26, 2019

Adversarial attacks for neural networks

Adversarial examples are intentionally designed optical illusions, where such inputs to learned models cause the model to make a mistake. Mathematically, given a point $\mathbf{x} \in \Omega$ drawn from class y, a scalar $\epsilon > 0$, and a metric d, we say that $\mathbf{x}$ admits an adversarial example in the metric d if there exists a point $\mathbf{x}^* \in \Omega$ with $Class(\mathbf{x}^*) \neq y$, and $d(\mathbf{x}, \mathbf{x}^*) \leq \epsilon$. In practice d is chosen as ℓ^p-norms with ℓ^∞ being the most popular choice, which limits the absolute change that can be made to any one dimension of $\mathbf{x}$.

1. Task1: Write a short report summarizing the fast gradient sign method (FGSM) for adversarial attacks[1]. Your report should be written in the format and style of a NIPS Proceedings, abridged to not exceed 2 pages. Latex style files and an exemplar template are provided on the course page, and are similar to last exercise.

2. Task2: One Layer Net: Consider the neural net defined as $\hat{y} = SM(\mathbf{W}\mathbf{x})$ trained with the cross-entropy loss $L(\mathbf{x}, y)$, where SM denotes softmax activation. Let $\mathbf{x}^*$ be the adversarial image of x resulting from FGSM attack with constant ϵ. Prove that $\forall \epsilon > 0$ we have $L(\mathbf{x}^*, y) \geq L(\mathbf{x}, y)$

3. Task3: Two Layer Net: Consider the neural net defined as $\hat{y} = SM(\mathbf{V}\sigma\mathbf{W}\mathbf{x})$ trained with the cross-entropy loss $L(\mathbf{x}, y)$, where $\mathbf{V}, \mathbf{W}$ are weights, SM denotes softmax activation and σ is ReLU activation. Suppose every element of $\mathbf{W}\mathbf{x}$ is non-zero, if $\epsilon < \frac{|\mathbf{W}\mathbf{x}|_{min}}{\|\mathbf{W}\|_\infty}$, then prove that $L(\mathbf{x}^*, y) \geq L(\mathbf{x}, y)$,

   given the fact that for $j = 1, 2, \ldots; sign(\mathbf{W}\mathbf{x})_j = sign(\mathbf{W}\mathbf{x}^*)_j$

[1] https://arxiv.org/pdf/1412.6572.pdf

Solution:

1. The loss for a example $\mathbf{x}$ and true class s can be expressed as:

$$\begin{aligned} L(\mathbf{x}, y) &= \text{crossentropy}(\text{softmax}(\mathbf{Wx}), y) \\ &= -ln(\text{softmax}(\mathbf{Wx})_s) \\ &= -ln\Big[\frac{\exp(\mathbf{Wx})_s}{\exp(\mathbf{Wx})_2 + \exp(\mathbf{Wx})_2 + \ldots + \exp(\mathbf{Wx})_k}\Big] \end{aligned} \tag{1}$$

now each element of vector $\mathbf{x}^*$ is expressed as:

$$\begin{aligned} x_i^* &= x_i + \epsilon sign(\frac{\partial L(\mathbf{x}, y)}{\partial x_i}) \\ &= x_i + \epsilon a_i \\ &= x_i + \epsilon sign(\sum_j^k \exp(\mathbf{Wx})_j w_{ji} - (\sum_j^k \exp(\mathbf{Wx})_j) w_{si}) \end{aligned} \tag{2}$$

Assuming the hypothesis is true we have to prove:

$$\begin{aligned} &\frac{\exp(\mathbf{Wx})_s}{\sum_j^k \exp(\mathbf{Wx})_j} \geq \frac{\exp(\mathbf{Wx}^*)_s}{\sum_j^k \exp(\mathbf{Wx}^*)_j} \\ \implies &\frac{\exp(\mathbf{Wx}^*)_s}{\exp(\mathbf{Wx})_s} \leq \sum_j^k \text{softmax}(\mathbf{Wx})_j \frac{\exp(\mathbf{Wx}^*)_j}{\exp(\mathbf{Wx})_j} \\ \implies &\exp(\epsilon \mathbf{Wa})_s \leq \sum_j^k \text{softmax}(\mathbf{Wx})_j \exp(\epsilon \mathbf{Wa})_j \end{aligned} \tag{3}$$

where $\mathbf{a} = [a_1 a_2 \ldots]^T$. By property of softmax and Jensen's inequality the RHS can be lower bounded by:

$$RHS \geq \exp(\sum_j^k \epsilon \text{ softmax}(\mathbf{Wx})_j (\mathbf{Wa})_j) \tag{4}$$

and hence we just need to prove

$$\begin{aligned} &\sum_j^k \text{softmax}(\mathbf{Wx})_j (\mathbf{Wa})_j \geq (\mathbf{Wa})_s \\ \implies &\sum_j^k \exp(\mathbf{Wx})_j (\mathbf{Wa})_j - \big(\exp(\mathbf{Wx})_1 + \exp(\mathbf{Wx})_2 \ldots\big)(\mathbf{Wa})_s \\ &\geq 0 \end{aligned} \tag{5}$$

where the result follows from (2) and fact that $\mathbf{x} sign(\mathbf{x}) > 0$

2. Let $\mathbf{T} = \mathbf{V}\sigma\mathbf{W}$ i.e., $y = \mathbf{T}\mathbf{x}$ and define the following index set (and using property given in the problem):

$$A = \{i : \mathbf{W}\mathbf{x}_i > 0\} = \{i : \mathbf{W}\mathbf{x}_i^* > 0\}. \tag{6}$$

Here $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{V} \in \mathbb{R}^{k \times l}$ and $\mathbf{W} \in \mathbb{R}^{l \times n}$. Then we can express the operator $\mathbf{T}$ as a linear operator:

$$\begin{aligned}(\mathbf{T}\mathbf{x})_j &= \sum_t^l v_{jt}\sigma(w_{t1}x_1 + w_{t2}x_2 + \ldots + w_{tn}x_n) \\ &= \sum_{t \in A} v_{jt}(w_{t1}x_1 + w_{t2}x_2 + \ldots + w_{tn}x_n)\end{aligned} \tag{7}$$

The loss for a example $\mathbf{x}$ and true class s can be expressed as:

$$\begin{aligned}L(\mathbf{x}, y) &= \text{crossentropy}(\text{softmax}(\mathbf{T}\mathbf{x}), y) \\ &= -ln(\text{softmax}(\mathbf{T}\mathbf{x})_s) \\ &= -ln\left[\frac{\exp(\mathbf{T}\mathbf{x})_s}{\exp(\mathbf{T}\mathbf{x})_2 + \exp(\mathbf{T}\mathbf{x})_2 + \ldots + \exp(\mathbf{T}\mathbf{x})_k}\right],\end{aligned} \tag{8}$$

which reduces to problem 1 with one linear layer.