Due Date : February 22nd, 24:00

Instructions

- *For all questions, show your work!*
- *Use a document preparation system such as LaTeX.*
- *Submit your answers electronically via the course gradescope*
- *TA for this assignment are :* ***Andjela Mladenovic*** *(IFT6135B) and* ***Ghait Boukachab*** *(IFT6135A).*

1. **Selection of Activation Function (10 pts)** We will compare two different activation functions in the following question. Recall the definition of $\sigma(x) = \frac{1}{1+e^{-x}}$ and $\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$.
   (a) (2 pts) Find the derivative of the sigmoid function $\sigma'(x)$ and express it in terms of the sigmoid function $\sigma(x)$.
   (b) (2 pts) Find the derivative of the $\tanh'(x)$ function and express it in terms of the $\tanh(x)$ function.
   (c) (2 pts) Upper bound the value of $\sigma'(x)$ with a constant (you can use AM–GM inequality).
   (d) (2 pts) Upper bound the value of $\tanh'(x)$ with a constant (you can use GM-HM inequality or the property that the square of real number is always non-negative).
   (e) (2 pts) Compare the two upper bounds and explain what impact would this difference have on optimization.

   Useful inequalities:

   Inequality of Arithmetic and Geometric Means (AM-GM)

$$\frac{x_1 + x_2 + \dots x_n}{n} \geq \sqrt[n]{x_1 x_2 \dots x_n} \tag{1}$$

   Inequality of Geometric and Harmonic Means (GM-HM)

$$\sqrt[n]{x_1 x_2 \dots x_n} \geq \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \dots \frac{1}{x_n}} \tag{2}$$

   The above inequalities hold for any real positive numbers $x_1, x_2, \dots x_n$ with equality if and only if $x_1 = x_2 = \dots = x_n$.

2. **Cross Entropy Properties (12 pts)**

   Cross-entropy loss function (a popular loss function) is given by:

$$\mathrm{CE}(p, x) = -x \log(p) - (1-x)\log(1-p)$$

   .

   Please refer to this loss for (a) and (b) parts.
   (a) (2 pts) **Cross Entropy and Maximum Likelihood** For this derivation, we assume that x is binary, i.e. $x \in \{0, 1\}$. Derive the cross-entropy cost function using the maximum likelihood principle for $x \in \{0, 1\}$.

(b) (2 pts) **Cross Entropy and KL divergence** Suggest a probabilistic interpretation of the cross-entropy cost function when $x \in (0, 1)$. (Hint: KL divergence between two distributions)

(c) (4 pts) **Discrete distribution - Maximum Entropy** Let X be a random variable which takes n values with probabilities $p_1, p_2, \ldots, p_n$ with $p_i > 0, \forall i$. What is the distribution that maximizes entropy $H(X) = -\sum_{i=1}^{n} p_i \log p_i$? Derive the upper bound for the entropy $H(X)$ expressed as a function of n. (Hint : use Jensen Inequality)

(d) (4 pts) **Continuous distribution (known mean μ and variance σ^2) - Maximum Entropy** Given mean μ and variance σ^2, what is the continuous distribution that maximizes differential entropy $h(X) = -\int_x f(x) \log f(x) dx$? Prove it.

3. **Output size and Parameters of Convolution Layers (5 pts)**
Consider a 3 hidden-layer convolutional neural network. Assume the input is a color image of size 128×128 in the RGB representation. The first layer convolves 64 8×8 kernels with the input, using a stride of 2 and zero-padding of 4. The second layer downsamples the output of the first layer with a 2×2 non-overlapping max pooling. The third layer convolves 128 4×4 kernels with a stride of 2 and zero-padding of 2.

   (a) (3 pts) What is the dimensionality of the output of the third layer?

   (b) (2 pts) Not including the biases, how many parameters are needed for the last layer?

4. **MLP Mixer (16 pts)**

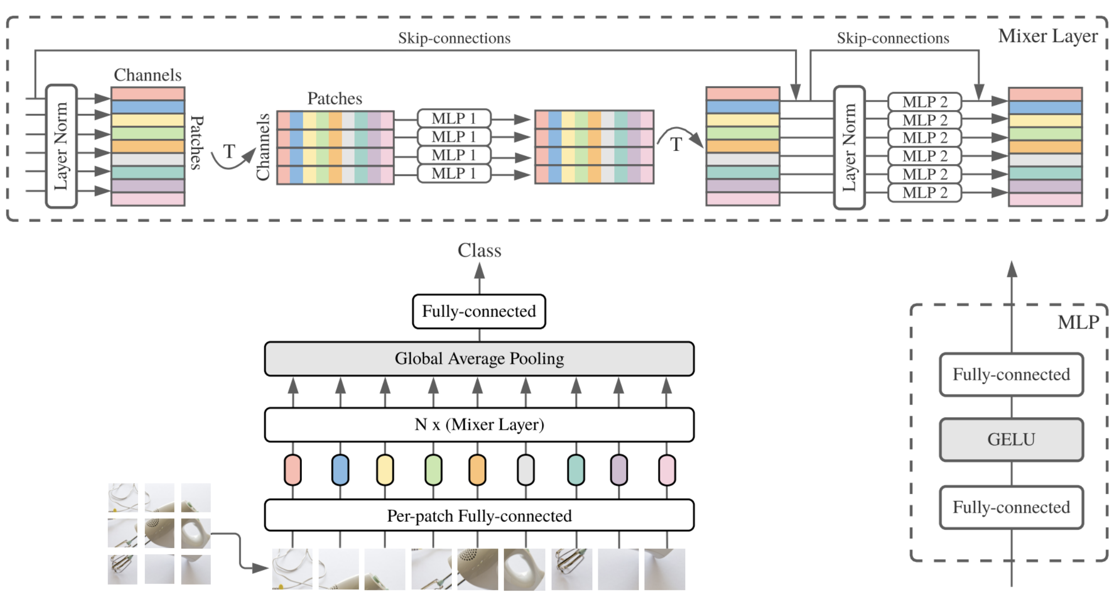

FIGURE 1 – (Borrowed from the MLPMixer paper.) MLP-Mixer consists of per-patch linear embeddings, Mixer layers, and a classifier head. Mixer layers contain one token-mixing MLP and one channel-mixing MLP, each consisting of two fully-connected layers and a GELU nonlinearity. Other components include : skip-connections, dropout, and layer norm on the channels.

(a) (2 pts) **MLP Mixer Dimensions** Let's assume that Mixer architecture is being applied to an input image of size 64×64. The Mixer's output is of size 16×128. Determine the patch resolution P, number of patches S, as hidden dimension C(channels).

(b) (2 pts) **MLP Mixer Complexity** Show that the computational complexity of the MLP Mixer is linear in terms of number of input patches.

(c) (6 pts) **Input Transformation - Channel Mixing MLP** Consider the following scenario : The original input image A is of size 9×9. We convert the input image into non-overlapping patches of size 3×3, and then linearly project all patches with the same projection matrix. The result of these operations is a matrix X of size 9×6. Then we apply the *channel-mixing MLP* that acts on rows of X, and is shared across all rows. The result of this operation is matrix U size 9×6. Now consider a modified image A such that $A_{\text{modified}} = PA$, where we define matrix P in the following manner:

$$P = \begin{bmatrix} e_{\pi(1)} \\ e_{\pi(2)} \\ \vdots \\ e_{\pi(9)} \end{bmatrix} \tag{3}$$

Here e_k is k-th basis vector and π represents the permutation of indices from $1 \dots 9$. Find all possible P such that by permuting rows of U_{modified} we can get back matrix U.

(d) (6 pts) Select one of your solutions for P and find $P_{reverse}$ such that $P_{\text{reverse}} U_{\text{modified}} = U$.

5. **Gradient Descent Convergence (12 pts)**

(a) (6 pts) **Convex Function Convergence** Consider the following function:

$$f(x) = \begin{cases} \frac{3}{4}(1-x)^2 - 2(1-x) & \text{if } x > 1 \\ \frac{3}{4}(1+x)^2 - 2(1+x) & \text{if } x < -1 \\ x^2 - 1 & \text{otherwise} \end{cases} \tag{4}$$

Show that f is a convex function. Find its unique minimizer and its gradient. Consider the following algorithm : $x_t = x_{t-1} - \eta f'(x_{t-1})$ where $\eta = 1$. Will this algorithm converge to a stationary point if it starts at point x_0, where $x_0 > 1$? Why or why not?

(b) (6 pts) **Prove Convergence of Gradient Descent to Stationary Point in Non-Convex case** Suppose we are trying to minimize the function $F(w)$ that is L-smooth. Let F_* be the minimal function value (i.e. the value at the global minima). Using $\eta = \frac{1}{L}$, prove that gradient descent will "almost" converge to a stationary point in a bounded (and polynomial) number of steps. Precisely,

$$\min_{k<K} \|\nabla F(w^{(k)})\|^2 \leq \frac{2L}{K}(F(w^{(0)}) - F_*) \tag{5}$$

Hints:

i. L-smoothness implies that:

$$F(w^{(k+1)}) \leq F(w^{(k)}) - \eta \|\nabla F(w^{(k)})\|^2 + \frac{1}{2}\eta^2 L \|\nabla F(w^{(k)})\|^2 \tag{6}$$

Combine this with $\eta = \frac{1}{L}$

ii. Use the fact that the minimum of a sequence of elements is less than the average of the sequence.