HW2 Handwritten Assignment

Lecturor: Pei-Yuan Wu
TAs: Yuan-Chia Chang(Problem 5), Chun-Lin Huang(Problem 1, 2, 3, 4)

October 2023, First Edition

Problem 1 (Convolution)(0.5%)

As mentioned in class, image size may change after convolution layers. Consider a batch of image data with shape $(B, W, H, input\_channels)$, how will the shape change after the following convolution layer:

$$Conv2D\,(input\_channels, output\_channels, kernel\_size = (k_1, k_2), stride = (s_1, s_2), padding = (p_1, p_2))$$

For simplicity, the padding tuple means that p_1 pixels are padded on both left and right sides, and p_2 pixels are padded on both top and bottom sides.

Problem 2 (Batch Normalization)(1%)

Batch normalization [**?**] is a popular trick for training deep networks nowadays, which aims to preserve the distribution within hidden layers and avoids vanishing gradient issue. The alogrithm can be written as below:

Algorithm 1 Batch Normalization

Input Feature from data points over a mini-batch $B = (x_i)_{i=1}^m$
Output $y_i = BN_{\gamma,\beta}(x_i)$

```
procedure BatchNormalize(B, γ, β)
    μ_B ← (1/m) Σ_{i=1}^m x_i                          ▷ mini-batch mean
    σ_B^2 ← (1/m) Σ_{i=1}^m (x_i − μ_B)^2              ▷ mini-batch variance
    for i ← 1 to m do
        x̂_i ← (x_i − μ_B) / sqrt(σ_B^2 + ε)            ▷ normalize
        y_i ← γ x̂_i + β                                ▷ scale and shift
    end for
    return
end procedure
```

During training we need to backpropagate the gradient of loss ℓ through this transformation, as well as compute the gradients with respect to the parameters γ, β. Towards this end, please write down the close form expressions for $\frac{\partial \ell}{\partial x_i}$,

$\frac{\partial \ell}{\partial \gamma}$, $\frac{\partial \ell}{\partial \beta}$ in terms of x_i, μ_B, σ_B^2, $\hat{x}_i$, y_i (given by the forward pass) and $\frac{\partial \ell}{\partial y_i}$ (given by the backward pass).

- Hint: You may first write down the close form expressions of $\frac{\partial \ell}{\partial \hat{x}_i}$, $\frac{\partial \ell}{\partial \sigma_B^2}$, $\frac{\partial \ell}{\partial \mu_B}$, and then use them to compute $\frac{\partial \ell}{\partial x_i}$, $\frac{\partial \ell}{\partial \gamma}$, $\frac{\partial \ell}{\partial \beta}$.

Problem 3 (Constrained Mahalanobis Distance Minimization Problem)(1.5%)

1. Let $\Sigma \in R^{m\times m}$ be a symmetric positive semi-definite matrix, $\mu \in R^m$. Please construct a set of points $x_1, ..., x_n \in R^m$ such that

$$\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)(x_i-\mu)^T = \Sigma, \quad \frac{1}{n}\sum_{i=1}^{n} x_i = \mu$$

   - Find the relation between set of points and (μ, Σ) and (μ, Σ) is known

2. Let $1 \le k \le m$, solve the following optimization problem (and justify with proof):
   minimize $\quad Trace(\Phi^T \Sigma \Phi)$
   subject to $\quad \Phi^T \Phi = I_k$
   variables $\quad \Phi \in R^{m\times k}$

References

[1] Sergey Ioffe and Christian Szegedy (2015), “Batch Normalization: Accelerating Deep Network Training b y Reducing Internal Covariate Shift”, Arxiv:1502.03167

Problem 4 (Convergence of K-means Clustering) (1.5%)

In the K-means clustering algorithm, we are given a set of n points $x_i \in \mathbb{R}^d, i \in \{1, \ldots, n\}$ and we want to find the centers of k clusters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_k)$ by minimizing the average distance from the points to the closest cluster center. In general, $n \ge k$. Define function $\mathcal{C} : \{1, \cdots, n\} \to \{1, 2, \cdots, k\}$ assigns one of k clusters to each point in the data set such that $\mathcal{C}(i) = q$ if the i-th data point x_i is assigned to the q-th cluster where $i \in \{1, 2, \cdots, n\}$ and $q \in \{1, 2, \cdots, k\}$

Formally, we want to minimize the following loss function

$$L(\mathcal{C}, \boldsymbol{\mu}) = \sum_{i=1}^{n} \left\| x_i - \mu_{\mathcal{C}(i)} \right\|_2^2 = \sum_{q=1}^{k} \sum_{i:\mathcal{C}(i)=q} \left\| x_i - \mu_q \right\|_2^2$$

The K-means algorithm:

Algorithm 2 K-means algorithm

Initialize cluster center $\mu_j, j = 1, 2, \cdots, k$ (k random x_n from data set)
Repeat:

1. Fix $\boldsymbol{\mu}$, update $\mathcal{C}(i)$ for each i that minimizes L. Formally, consider a data point x_i, and let $\mathcal{C}(i)$ be the assignment from the previous iteration and $C^*(i)$ be the new assignment obtained as: $C^*(i) = \arg\min_{j=1,\cdots,k} \|x_i - \mu_j\|_2^2$
2. Fix $\mathcal{C}$, update the centers μ_j which satisfies

$$|\{i : \mathcal{C}(i) = j\}| \, \mu_j = \sum_{i:\mathcal{C}(i)=j} x_i,$$

for each j, where $|\{i : \mathcal{C}(i) = j\}|$ is the number of elements of set $\{i : \mathcal{C}(i) = j\}$.(i.e. Set the cluster centres to be the means of the points in each cluster.)

The algorithm stops when no change in loss function occurs during the assignment step.

Suppose that the algorithm proceeds from iteration t to $t + 1$.

1. Consider the points $z_1, z_2, \cdots, z_m$, where $m \geq 1$. and for $i \in \{1, 2, \cdots, m\}$, $z_i \in \mathbb{R}^d$. Let $\bar{z} = \frac{1}{m} \sum_{i=1}^{m} z_i$ be the mean of these points, and let $z \in \mathbb{R}^d$ be an arbitrary point in the same (d-dimensional) space. Then

$$\sum_{i=1}^{m} \|z_i - z\|_2^2 \geq \sum_{i=1}^{m} \|z_i - \bar{z}\|_2^2$$

2. Show that $L(\mathcal{C}^{t+1}, \boldsymbol{\mu}^t) \leq L(\mathcal{C}^t, \boldsymbol{\mu}^t)$ i.e. The first step in K-means clustering
3. Show that $L(\mathcal{C}^{t+1}, \boldsymbol{\mu}^{t+1}) \leq L(\mathcal{C}^{t+1}, \boldsymbol{\mu}^t)$ i.e. The second step in K-means clustering. (Hint: Use the result of (a))
4. Use the result in (b) and (c) to show that the loss of k-means clustering algorithm is monotonic decreasing.(Hint: Show that the sequence $\{l_t\}$, where $l_t = L(\mathcal{C}^t, \boldsymbol{\mu}^t)$, which is monotone decreasing ($l_{t+1} \leq l_t, \forall t$) and bounded below ($l_t \geq 0$). Then, we use monotone convergence theorem for sequences, $\{l_t\}$ converges.)
5. Show that the k-means clustering algorithm converges in finitely many steps.

Problem 5 (Gradient Descent Convergence) (1.5%)

Suppose the function $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable. Also, f is β-smoothness and α-strongly convex.

$$\beta - smoothness : \beta > 0, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq \beta \|\boldsymbol{x} - \boldsymbol{y}\|_2$$

$$\alpha{-}strongly\ convex : \alpha > 0, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \geq \frac{\alpha}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

Then we propose a gradient descent algorithm

1. Find a initial $\boldsymbol{\theta}^0$.
2. Let $\boldsymbol{\theta}^{n+1} = \boldsymbol{\theta}^n - \eta \nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)\ \forall n \geq 0$, where $\eta = \frac{1}{\beta}$.

The following problems lead you to prove the gradient descent convergence.

1. Prove the property of β-smoothness function

$$\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \leq \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

   (a) Define $g : \mathbb{R} \to \mathbb{R}, g(t) = f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))$. Show that $f(\boldsymbol{x}) - f(\boldsymbol{y}) = \int_0^1 g^{'}(t)\, dt$.

   (b) Show that $g^{'}(t) = \nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y})$.

   (c) Show that $|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})| \leq \int_0^1 |(\nabla f(\boldsymbol{y} + t(\boldsymbol{x} - \boldsymbol{y})) - \nabla f(\boldsymbol{y}))^T(\boldsymbol{x} - \boldsymbol{y})|\, dt$.

   (d) By Cauchy-Schwarz inequality and the definition of β-smoothness, show that $|f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y})| \leq \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$, hence we get

$$f(\boldsymbol{x}) - f(\boldsymbol{y}) - \nabla f(\boldsymbol{y})^T(\boldsymbol{x} - \boldsymbol{y}) \leq \frac{\beta}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2$$

2. Let $\boldsymbol{y} = \boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})$ and use 1., prove that

$$f(\boldsymbol{x} - \frac{1}{\beta}\nabla f(\boldsymbol{x})) - f(\boldsymbol{x}) \leq -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2$$

   and

$$f(\boldsymbol{x}^*) - f(\boldsymbol{x}) \leq -\frac{1}{2\beta}\|\nabla f(\boldsymbol{x})\|_2^2,$$

   where $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} f(\boldsymbol{x})$.

3. Show that $\forall n \geq 0$,

$$\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 = \|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2 + \eta^2\|\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)\|_2^2 - 2\eta\nabla_{\boldsymbol{\theta}} f(\boldsymbol{\theta}^n)^T(\boldsymbol{\theta}^n - \boldsymbol{\theta}^*),$$

   where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$.

4. Use 2. and the definition of α-strongly convex to prove $\forall n \geq 0$

$$\|\boldsymbol{\theta}^{n+1} - \boldsymbol{\theta}^*\|_2^2 \leq (1 - \frac{\alpha}{\beta})\|\boldsymbol{\theta}^n - \boldsymbol{\theta}^*\|_2^2,$$

where $\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} f(\boldsymbol{\theta})$.

5. Use the above inequality to show that $\boldsymbol{\theta}^n$ will converge to $\boldsymbol{\theta}^*$ when n goes to infinity.

Version Description

1. First Edition: Finish Problem 1 to 5