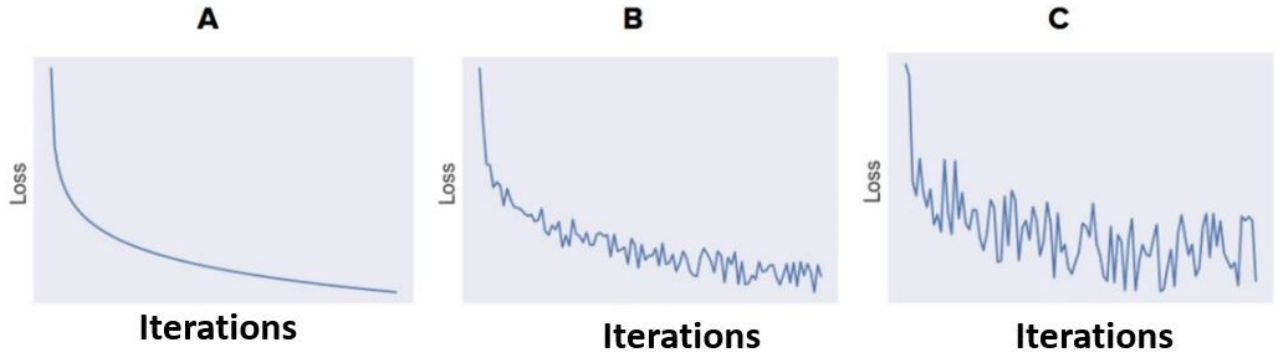
به نام خدا		
نام درس: یادگیری عمیق	نام دانشکده: دانشکده برق و کامپیوتر	
نام استاد: دکتر سمانه حسینی	نام طراح: مریم محمدی	نیمسال: ۱-۱۴۰۳-۱۴۰۲
نمره: ۱۹.۲۵	زمان تحویل: ۷ آذر	

سوال اول:

الف) مشکل exploding gradient را توضیح دهید و یک راه‌کار برای رفع این مشکل ارائه کنید. (۱.۵ نمره)

ب) ۳ مزیت استفاده از mini-batch gradient descent را در مقایسه با stochastic gradient descent با $\text{batch size} = 1$ نام ببرید. (۱.۵ نمره)

ج) در هریک از شکل‌های زیر منحنی مربوط به train loss در برابر تعداد $epoch$ها نشان داده شده است. مشخص کنید کدام نمودارها مربوط به کدام یک از الگوریتم‌های SGD ، MBGD و BGD است. (۰/۷۵ نمره)



د) نقطه زینی را تعریف کنید. سپس توضیح دهید استفاده از الگوریتم stochastic gradient descent چه عیب/مزیتی در مقابله با نقاط زینی دارد؟(۱ نمره)

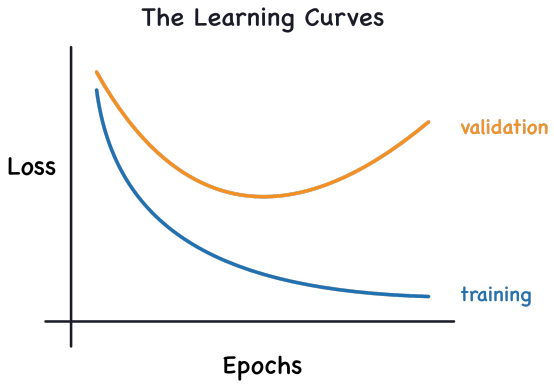
سوال دوم:

ثابت کنید که کران پایین برای cross-entropy loss با L کلاس درست به صورت زیر است (۱.۵ نمره)

$$L(\hat{y}, y) \geq L \log L, \quad y \in \{0,1\}^{n_y}.$$

سوال سوم:

نمودار زیر مربوط به آموزش یک شبکه عمیق است که مقدار تابع هزینه را در برابر تعداد Epoch نشان می‌دهد. به نظر شما این نمودار چه اتفاقی را در شبکه نشان می‌دهد؟ (۰/۷۵ نمره) هریک از موارد زیر چه تاثیری روی این پدیده خواهند داشت.



۱. عمیق‌تر کردن شبکه (۰/۷۵ نمره)
۲. استفاده از تکنیک early stopping (۰/۷۵ نمره)
۳. در نظر گرفتن تعداد $epoch$های بیشتر برای train شبکه (۰/۷۵ نمره)
۴. استفاده از تکنیک data augmentation (۰/۷۵ نمره)

سوال چهارم:
تابع softmax دارای خاصیت مطلوبی است که یک توزیع احتمال را خروجی می‌کند و اغلب به عنوان تابع فعال‌سازی در بسیاری از شبکه‌های عصبی مورد استفاده قرار می‌گیرد. یک شبکه عصبی ۲ لایه را برای طبقه‌بندی K-class با استفاده از تابع فعال‌ساز softmax و تابع Log Loss به صورت زیر در نظر بگیرید:

$$
\begin{aligned}
z^{[1]} &= W^{[1]}x + b^{[1]}, \\
a^{[1]} &= LeakyReLU(z^{[1]}, \alpha = 0.01), \\
z^{[2]} &= W^{[2]}a^{[1]} + b^{[2]}, \\
\hat{y} &= softmax(z^{[2]}), \\
L &= -\sum_{i=1}^{K} y_i \log(\hat{y}_i).
\end{aligned}
$$

که در آن ورودی x دارای ابعاد $D_x \times 1$ و $y \in \{0,1\}^K$. فرض کنید لایه مخفی دارای D_a نورون باشد یعنی یک بردار به ابعاد $D_a \times 1$ باشد. با توجه به اطلاعات مطرح شده، به سوالات زیر پاسخ دهید:
الف) مقدار $\frac{\partial \hat{y}_k}{\partial z_k^{[2]}}$ را برحسب $\hat{y}$ محاسبه کنید. (۰/۵ نمره)
ب) مقدار $\frac{\partial \hat{y}_k}{\partial z_i^{[2]}}$ را برای $i \neq k$ برحسب $\hat{y}$ محاسبه کنید.(۰/۵ نمره)
ج) فرض کنید درایه kام بردار y برابر ۱ و بقیه درایه‌هایش برابر صفر باشند، در این صورت مقدار $\frac{\partial L}{\partial z_i^{[2]}}$ را برای هر دو حالت $i = k$ و $i \neq k$ محاسبه کنید. (۱ نمره)
د) اگر $\frac{\partial L}{\partial z^{[2]}}$ را با δ_0 نشان دهیم، $\frac{\partial L}{\partial W^{[1]}}$ و $\frac{\partial L}{\partial b^{[1]}}$ را بدست آورید. (۱ نمره)
ه) برای جلوگیری از مواجه‌شدن با مشکلات پایداری عددی، می‌توانیم تابع softmax را به صورت زیر تعریف کنیم.

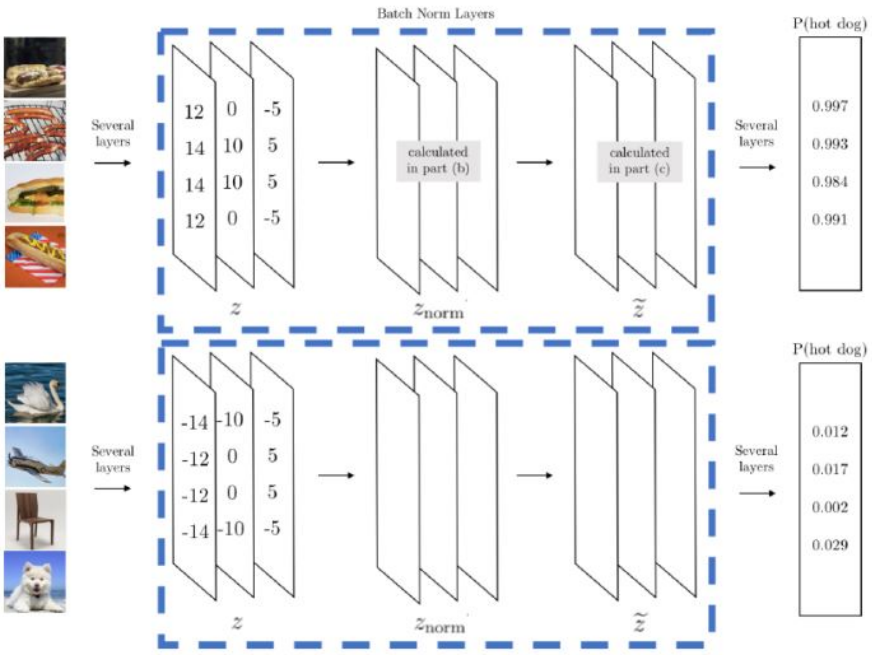
$$\hat{y}_i = \frac{\exp(z_i^{[2]} - m)}{\sum_{j=1}^{K} \exp(z_j^{[2]} - m)}, \quad m = \max z_i.$$

به نظر شما مشکل تابع اولیه softmax در محاسبات چیست و چرا فرمول اصلاح‌شده می‌تواند به حل مشکل آن کمک کند؟(۱ نمره)
سوال پنجم:

در این سوال، اهمیت توازن مناسب میان نمونه‌های مثبت و منفی در یک mini-batch مورد بررسی قرار می‌گیرد. در نظر داشته باشید که لایه Batch Normalization مقدار $z = (z^{(1)}, \cdots, z^{(m)})$ را به عنوان ورودی دریافت و مقدار $\tilde{z} = (\tilde{z}^{(1)}, \cdots, \tilde{z}^{(m)})$ را براساس روابط زیر محاسبه میکند

$$
\begin{aligned}
z^i_{norm} &= \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}, \\
\mu &= \frac{1}{m}\sum_{i=1}^{m} z^{(i)}, \\
\sigma^2 &= \frac{1}{m}\sum_{i=1}^{m} z^{(i-\mu)^2}, \\
\tilde{z}^{(i)} &= \gamma z^{(i)}_{norm} + \beta.
\end{aligned}
$$

فرض کنید یک شبکه عصبی به صورت شکل زیر را برای مسئله "تشخیص هات داگ" داشته باشیم. با در نظر گرفتن مقدار ϵ برابر صفر، به سوالات زیر پاسخ دهید



الف) مرحله انتشار رو به جلو (forward) را برای یک batch با تعداد m نمونه در نظر بگیرید. ورودی لایه Batch Normalization که به صورت $z = (z^{(1)}, \cdots, z^{(m)})$ در نظر می‌گیریم، دارای ابعاد $m = 4$ و $n = 3$ است که $n = 3$ تعداد نورون‌ها در لایه قبل از لایه Normalization Batch را نشان می‌دهد.

$$
\begin{bmatrix}
12 & 14 & 14 & 12 \\
0 & 10 & 10 & 0 \\
-5 & 5 & 5 & -5
\end{bmatrix}
$$

مقدار z_{norm} را محاسبه کنید. (۰/۷۵ نمره)

ب) ۳ مورد از مزیت‌های استفاده از لایه Batch Normalization را در شبکه ذکر کنید. (۰/۷۵ نمره)

ج) به نظر شما اگر در لایه Batch Normalization به جای $\tilde{z}$ از z_{norm} استفاده کنیم چه اتفاقی خواهد افتاد؟ (۱ نمره)

د) مشکل covariate shift را توضیح دهید و بگویید استفاده از Batch Normalization چه کمکی به رفع این مشکل می‌کند؟(۱ نمره)

سوال ششم: نشان دهید چرا اگر از تابع softmax به عنوان تابع فعال‌ساز استفاده کنیم، cross-entropy loss هرگز صفر نخواهد شد؟ (تعداد

کلاس درست را c در نظر بگیرید) (۱ نمره)
سوال هفتم: کدام مورد یا موارد زیر در مورد مشکل vanishing gradient صحیح است؟(۰/۷۵ نمره)
الف) تابع فعالیت Tanh معمولا بر تابع sigmoid ترجیح داده می‌شود زیرا مشکل vanishing gradient را ندارد.
ب) تابع فعالیت Leaky Relu کمتر از تابع sigmoid از مشکل vanishing gradient رنج می‌برد.
ج) Xavier initialization می‌تواند به رفع مشکل vanishing gradient کمک کند.
د) اضافه کردن Batch Normalization قبل از هر تابع فعالیت می‌تواند به رفع مشکل vanishing gradient کمک کند.

توضیحات:
۱. همانطور که قبلا هم اطلاع داده شد، شما مجاز هستید در طول ترم تا ۸ روز تاخیر در تحویل تکالیف داشته باشید.
۲. دانشجویان می‌توانند در حل تکالیف با دوستان خود مشورت نمایند اما در نهایت هرکس موظف است تکلیف را به صورت فردی انجام و تحویل دهد. لذا، در صورت مشاهده تکالیف کپی بین دانشجویان، نمره تمامی افراد شرکت‌کننده در آن، صفر خواهد بود.
۳. در صورت داشتن هرگونه سوال میتوانید از طریق ایمیل زیر با دستیار آموزشی مربوطه در ارتباط باشید.
mohammadi.maryam@math.iut.ac.ir