بسم الله الرحمن الرحیم

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی کامپیوتر

یادگیری عمیق

نیم‌سال دوم ۰۳-۰۲

مدرس: مهدیه سلیمانی

تمرین دوم

- مهلت ارسال پاسخ تا ساعت **۲۳:۵۹** روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر تمرین‌های نظری بدون کسر نمره تا سقف ۵ روز و تمرین‌های عملی تا سقف ۱۰ روز وجود دارد. محل بارگزاری جواب تمرین‌های نظری بعد از ۳ روز و تمرین‌های عملی بعد از ۵ روز بسته خواهد شد و پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند شد.
- هم‌فکری در انجام تمرین مانعی ندارد، فقط توجه داشته باشید که پاسخ تمرین حتما باید توسط خود شخص نوشته شده باشد. همچنین در صورت همفکری در هر تمرین، در ابتدای جواب تمرین نام افرادی که با آن‌ها هم‌فکری کرده اید را حتما ذکر کنید.
- برای پاسخ به سوالات نظری در صورتی که از برگه خود عکس تهیه می‌کنید، حتما توجه داشته باشید که تصویر کاملا واضح و خوانا باشد. درصورتی که خوانایی کافی را نداشته باشد، تصحیح نخواهد شد.
- محل بارگذاری سوالات نظری و عملی در هر تمرین مجزا خواهد بود. به منظور بارگذاری بایستی تمارین تئوری در یک فایل pdf با نام HW2_[First-Name]_[Last-Name]_[Student-Id].pdf و تمارین عملی نیز در یک فایل مجزای زیپ با نام HW2_[First-Name]_[Last-Name]_[Student-Id].zip بارگذاری شوند.
- در صورت وجود هرگونه ابهام یا مشکل، در کوئرای درس آن مشکل را بیان کنید و از پیغام دادن مستقیم به دستیاران آموزشی خودداری کنید.
- طراحان این تمرین : آقایان محمد جواد محمدی، علی رازقندی، حسام اسدالله زاده

بخش نظری (۷۰ نمره)

سوال اول: (۱۶ نمره)

تابع هزینه مسئله‌ی رگرسیون خطی به صورت زیر تعریف می‌شود:

$$\mathcal{L}(\mathbf{w}) = \|y - X\mathbf{w}\|_2^2 \quad (۱)$$

همانطور که در درس دیدید، می‌توانیم چند عنصر دیگر به عنوان Regularization Term به این تابع هزینه اضافه کنیم. در این صورت خواهیم داشت:

$$\mathcal{L}(\mathbf{w}) = \|y - X\mathbf{w}\|_2^2 + \|\Gamma\mathbf{w}\|_2^2 \quad (۲)$$

که Γ یا Tikhonov matrix باید به درستی انتخاب شود. Γ در اکثر موارد یک ضریب از ماتریس همانی انتخاب می‌شود ($\Gamma = \lambda I$). دقت کنید که L_2 Regularization یک حالت خاص از Tikhonov Regularization است. حال فرض کنید می‌خواهیم از dropout در فرآیند آموزش مدل رگرسیون خطی خود استفاده کنیم. برای ورودی dـبُعدی فرض کنید که یک ویژگی با احتمال p نگه داشته شده و در غیر اینصورت zero-out خواهد شد. در این صورت تابع هزینه به شکل مقابل تغییر خواهد کرد:

$$\mathcal{L}(\hat{\mathbf{w}}) = \mathbb{E}_{D\sim Bernoulli(p)}[\|y - (D \odot X)\hat{\mathbf{w}}\|_2^2] \quad (۳)$$

که در اینجا $\hat{\mathbf{w}}$ پارامترهای پیدا شده توسط مدلی که با dropout آموزش داده شده است می‌باشد. همچنین $\odot$ ضرب element-wise می‌باشد.

الف) اگر $P = D \odot X$ که $D \sim Bernoulli(p)$. اثبات کنید مقادیر $\mathbb{E}[P]$ و $\mathbb{E}[P^T P]$ به شکل مقابل خواهند بود:

$$\mathbb{E}_D[P]_{ij} = \mathbb{E}_D[(D \odot X)_{ij}] = X_{ij}\mathbb{E}_D[D_{ij}] = pX_{ij} \qquad (۴)$$

$$\mathbb{E}_D[(P^T P)]_{ij} = \begin{cases} \sum_{k=1}^{N} \mathbb{E}_D[D_{ki}D_{kj}X_{ki}X_{kj}] = \sum_{k=1}^{N} \mathbb{E}_D[D_{ki}]\mathbb{E}_D[D_{kj}]X_{ki}X_{kj} = p^2(X^T X)_{ij} & \text{if } i \neq j \\ \sum_{k=1}^{N} \mathbb{E}_D[D_{ki}^2 X_{ki}X_{kj}] = \sum_{k=1}^{N} \mathbb{E}_D[D_{ki}^2]X_{ki}X_{kj} = p(X^T X)_{ij} & \text{if } i = j \end{cases}$$

ب) حال اثبات کنید که می‌توانیم این تابع هزینه را به شکل زیر بازنویسی کنیم:

$$\mathcal{L}(\hat{\mathbf{w}}) = \|y - pX\hat{\mathbf{w}}\|_2^2 + p(1-p)\|\hat{\Gamma}\hat{\mathbf{w}}\|_2^2 \qquad (۵)$$

به طوری که $\hat{\Gamma}$ یک ماتریس قطری بوده که عنصر jاُم قطری این ماتریس، برابر نُرم ستون jاُم ماتریس دادگان X می‌باشد.

ج) می‌دانیم که اکثر روش‌های رگولاریزیشن، در هنگام آموزش یک مقدار noise به فرآیند وارد می‌کنند که این نویز باید در هنگام inference به نحوی average-out شود تا تاثیر نویز اعمال شده از بین برود. در روش dropout در PyTorch وزن‌های لایه‌ای که روی آن dropout اعمال شده در $\frac{1}{1-p}$ ضرب شده و اسکیل می‌شوند و در زمان inference بدون تغییر استفاده می‌شوند. حال در مسئله رگرسیون خطی ثابت کنید می‌توان ۵ را با انجام چند transformation ساده روی $\hat{\mathbf{w}}$ و $\hat{\mathbf{\Gamma}}$ به صورت ۲ نوشت. در پاسخ خود ابتدا $\hat{\mathbf{w}}$ را بر حسب $\mathbf{w}$ بیان کنید و سپس $\mathbf{\Gamma}$ را برحسب $\hat{\mathbf{\Gamma}}$ به دست آورید.

د) فرض کنید Γ معکوس‌پذیر باشد. با یک تغییر متغیر سعی کنید ۲ را به صورت تابع هزینه مسئله Ridge Regression بازنویسی کنید:

$$\mathcal{L}(\tilde{\mathbf{w}}) = \|y - \tilde{X}\tilde{\mathbf{w}}\|_2^2 + \lambda\|\tilde{\mathbf{w}}\|_2^2 \qquad (۶)$$

به طور خاص حتما تغییرات $\tilde{X}$ نسبت به X و $\tilde{w}$ نسبت به w را بیان کنید.

ت) حال می‌دانیم Γ یک ماتریس **قطری** و معکوس‌پذیر است که عنصر jاُم قطری آن متناسب با نُرم ستون jاُم X است. در مورد نُرم ستون‌های ماتریس $\tilde{X}$ چه می‌توان گفت؟ چه ارتباطی میان این مسئله و Batch Normalization قابل بیان است؟

در ادامه می‌خواهیم با نوع دیگری از dropout آشنا شویم.

و) در صورتی که در یک شبکه از Dropout نوع گاوسی جمعی استفاده شود، تابع خطای آن بدین صورت است:

$$J = \frac{1}{2}(y_d - \sum_{k=1}^{n}(\omega_k + \delta_k)x_k)^2 \qquad (۷)$$

به طوریکه در آن $\delta_k \sim N(0, \alpha \cdot \omega_k^2)$ برقرار است.
مقدار امید ریاضی گرادیان تابع هدف نسبت به متغیر ω_k یعنی $E[\frac{\partial J}{\partial \omega_i}]$ را بدست آورید.

ه) سعی کنید با استفاده از این نوع Dropout تعبیری از رگولاسیون ارائه دهید. (لازم است با نوشتن روابط ریاضی این مطلب را نشان دهید و با توجه به آن توضیح دهید.)

ی) این تکنیک (Dropout نوع گاوسی-جمعی) را با روش Spatial Dropout مقایسه و تحلیل نمائید.

سوال دوم: (۷ نمره)

فرض کنید برداری به طول N دارید و قصد دارید یک لایه‌ی کانولوشن یک بعدی روی آن اعمال کنید. حاصل اعمال یک لایه کانولوشن را از طریق رابطه‌ی:

$$Z = W * X \ \rightarrow \ Z_i = \sum_{j=0}^{K-1} W_j X_{i+j} \qquad (۸)$$

به دست می‌آوریم که K اندازه‌ی فیلتر را نشان می‌دهد. اگر مقدار $\frac{\partial L}{\partial Z_i}$ را برای تمامی مقادیر i بدانیم، رابطه‌ی مربوط به $\frac{\partial L}{\partial W_j}$ را به طور دقیق برحسب آن پیدا کنید. نشان دهید این رابطه برای پیدا کردن مقادیر $\frac{\partial L}{\partial W_j}$ عملا معادل اعمال یک فیلتر کانولوشن است.

سوال سوم: (۱۳ نمره)

در این سوال قصد بررسی گرادیان پارامترهای Batch Normalization داریم.

Input: Values of x over a mini-batch: $\mathcal{B} = \{x_{1...m}\}$;
Parameters to be learned: γ, β
Output: $\{y_i = \text{BN}_{\gamma,\beta}(x_i)\}$

$$\mu_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m} x_i \qquad \text{// mini-batch mean}$$

$$\sigma^2_{\mathcal{B}} \leftarrow \frac{1}{m}\sum_{i=1}^{m}(x_i - \mu_{\mathcal{B}})^2 \qquad \text{// mini-batch variance}$$

$$\widehat{x}_i \leftarrow \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma^2_{\mathcal{B}} + \epsilon}} \qquad \text{// normalize}$$

$$y_i \leftarrow \gamma \widehat{x}_i + \beta \equiv \text{BN}_{\gamma,\beta}(x_i) \qquad \text{// scale and shift}$$

(آ) با در نظر گرفتن فرمول بالا مقدار گرادیان loss را نسبت به ورودی $(\partial l/\partial x)$ و گرادیان هر دو پارامتر قابل یادگیری در Batch Normalization را حساب کنید $(\partial l/\partial \gamma, \partial l/\partial \beta)$. برای انجام محاسبات مقدار ϵ را صفر در نظر بگیرید.

(ب) دو مورد از مزایای استفاده از Batch Normalization را توضیح دهید.

(ج) توضیح دهید طبق فرمول‌های ارائه شده اگر خروجی را بجای y، $\hat{x}$ در نظر بگیریم چه مشکلی بوجود می‌آید (توجه کنید در این حالت scale و shift نداریم)؟

(د) حال فرض کنید در مسئله‌ای دو کلاس داریم؛ به عنوان مثال دو دسته‌ی سگ و گربه. حال یک شبکه برای دسته‌بندی عکس‌های موجود آموزش می‌دهیم که در آن از لایه BN استفاده می‌شود. فرض کنید در زمان آموزش به علت ایرادی که در پیاده‌سازی وجود دارد، هر mini-batch تنها شامل یکی از این دو کلاس است.

توضیح دهید چرا وجود لایه BN در شبکه با توجه به مدل ساخته شدن mini-batch ها باعث بوجود آمدن مشکل discrepancy بین آموزش و آزمون می‌شود.

(راهنمایی: توجه کنید که در زمان ترین متغیرهای μ و σ از روی داده های هر mini-batch محاسبه می‌شوند اما در زمان تست این دو متغیر از میانگین متحرک (moving average) مقادیر محاسبه شده در زمان ترین محاسبه می‌شوند.)

سوال چهارم: (۱۲ نمره)

به سوالات زیر در حد یک یا دو خط پاسخ کوتاه دهید.

(آ) شبکه CNN را در نظر بگیرید که از بلاک‌هایی به فرم زیر استفاده می‌کند.

$$(ConvLayer) \longrightarrow (BatchNorm) \longrightarrow (Activation)$$

آیا حذف بایاس b از لایه کانولوشن در این شبکه ایجاد مشکل می‌کند؟ چرا؟

(ب) بلوک‌های مورد استفاده در شبکه بخش قبل را در نظر بگیرید. فرض کنید شبکه را آموزش داده‌ایم. آیا ضرب کردن وزن ها در یک عدد مانند α در زمان تست عملکرد شبکه را تغییر می‌دهد؟ ضرب کردن α در تمام درایه های ورودی شبکه چطور؟

(ج) مرتبه‌ی تعداد عملیات محاسباتی لازم در هر یک از لایه های زیر را برای یک ورودی به عرض w و ارتفاع h که دارای c کانال است بنویسید (محاسبات لازم برای رسیدن به جواب هر بخش را بنویسید).

(i) لایه کانولوشن با اندازه‌ی کرنل k و گام s

(ii) لایه pooling Average با اندازه‌ی کرنل k و گام s

(iii) لایه Norm Batch

(د) علت استفاده از کانولوشن 1×1 چیست؟

(ه) با توجه به اینکه عملیات pooling باعث از دست رفتن اطلاعات (loss of information) می‌شود، چرا همچنان از این عملیات در معماری های مختلف استفاده می‌شود؟

(و) یک بلوک CNN به صورت زیر را در نظر بگیرید:

3x3 Conv (stride 2) - 2x2 Pool - 3x3 Conv (stride 2) - 2x2 Pool

حال receptive field یک پیکسل خروجی این بلاک را بدست آورید.

سوال پنجم: (۱۴ نمره)

فرض کنید که تصویر RGB با ابعاد $512 \times 512 \times 3\,(height \times width \times channels)$ در ورودی داریم و می‌خواهیم دو شبکه عصبی کانولوشنی متفاوت با نام‌های شبکه A و شبکه B به آن اعمال کنیم. در ادامه مشخصات این دو شبکه را ملاحظه می‌کنید:

1. **Network A:**
   (i) Apply a depthwise separable convolution with a kernel of size 4x4, stride of 2, padding of 1.
   (ii) Apply a pointwise convolution to reduce the number of channels to 64.
   (iii) Apply a depthwise separable convolution with a kernel of size 4x4, stride of 2, padding of 1.
   (iv) Apply a pointwise convolution to reduce the number of channels to 128.
   (v) Apply a depthwise separable convolution with a kernel of size 4x4, stride of 2, padding of 1.
   (vi) Apply a pointwise convolution to reduce the number of channels to 256.
2. **Network B:**
   (i) Apply a standard convolution with a kernel of size 4x4, stride of 2, padding of 1, and 64 output channels.
   (ii) Apply a standard convolution with a kernel of size 4x4, stride of 2, padding of 1, and 128 output channels.
   (iii) Apply a standard convolution with a kernel of size 4x4, stride of 2, padding of 1, and 256 output channels.

حال، برای هر یک از این دو شبکه، به سوالات زیر پاسخ دهید:

(آ) ابعاد نقشه ویژگی بدست آمده پس از هر یک لایه‌های کانولوشنی را بدست آورید.

(ب) تعداد کل پارامترهایی که برای انجام عملیات‌های کانولوشنی وجود دارد را بدست آورید. (از بایاس صرف نظر کنید.)

(ج) تعداد کل عملیات‌های ضرب مورد نیاز برای بدست آمدن نقشه ویژگی خروجی از روی تصویر ورودی را بدست آورید.

(د) نتایج بدست آمده در سه قسمت قبلی در این دو شبکه را با هم مقایسه نمائید.

در ادامه، با توجه به این مقاله به سوالات زیر پاسخ دهید: (در صورت لزوم باید روابط ریاضی مورد نیاز نیز نوشته شود تا توضیحات شما کامل باشد).

(ه) درباره معماری مدل MobileNet توضیح دهید و استفاده از depthwise separable convolutions بعنوان جزء کلیدی در این معماری را تشریح کنید و اشاره کنید که چگونه باعث بهبود عملکرد این مدل می‌شود.

(و) این مقاله به دو ابرپارامتر مهم در این معماری اشاره می‌کند. درباره هر یک از آنها توضیح دهید و تاثیرشان را بر روی عملکرد، دقت و همچنین هزینه محاسباتی مدل توضیح دهید.

سوال ششم: (۸ نمره)

در بحث شبکه‌های عصبی کانولوشنی بنابر کاربرد مد نظر از توابع خطای متنوعی استفاده می‌شود. یکی از توابع خطا که اغلب در زمینه تشخیص چهره مورد استفاده قرار گرفته است تابع center loss می‌باشد که در حالت چند کلاسه بدین صورت تعریف می‌گردد:

$$L_c = \frac{1}{2}\sum_{i=1}^{m} ||x_i - C_{y_i}||^2 \qquad (۹)$$

که در آن فرض شده است این خطا برای یک دسته شامل m داده محاسبه شده است و x_i ویژگی‌های (عمیق) بدست آمده و همچنین C_{y_i} مرکز نمونه‌های کلاس متناظر با نمونه i ام در فضای امبدینگ هستند. با در نظر گرفتن این تابع ضرر به سوالات زیر پاسخ دهید.

(آ) توضیح دهید استفاده از این تابع ضرر به تنهایی و یا در ترکیب با سایر توابع ضرر چگونه می‌تواند به عمل دسته‌بندی تصاویر در یک شبکه عصبی کمک کند.

(ب) با فرض آنکه این تابع ضرر پس از یک لایه خطی استفاده شده باشد، رابطه به روز رسانی پارامترهای آن شبکه و همچنین مرکز دسته‌ها را بنویسید.

(ج) در این مقاله به روشی برای بهبود عمکرد این تابع ضرر برای استفاده در شبکه‌های عصبی کانولوشنی اشاره شده است. با مطالعه این مقاله این روش را تشریح نمائید.

بخش عملی (۳۰ نمره)

سوال اول

در این نوت‌بوک به دسته‌بندی و رنگ‌آمیزی عکس‌های موجود در دیتاست CIFAR10 می‌پردازیم. در بخش اول نوت‌بوک که مربوط به دسته‌بندی است شما باید معماری شبکه ResNet را پیاده سازی کنید و از این شبکه برای دسته‌بندی نمونه‌ها استفاده کنید. در بخش دوم نوت‌بوک نیز بعد از پیاده سازی معماری U-Net از آن برای تبدیل عکس‌های سیاه و سفید به عکس‌های رنگی استفاده می‌کنیم.

سوال دوم

در این نوت‌بوک، شما با داده ای از تصاویر گربه و سگ کار خواهید کرد. هدف این است که یک مدل شبکه عصبی کانولوشنال (CNN) در PyTorch پیاده سازی کرده و آموزش دهید تا بتواند این تصاویر را به درستی طبقه بندی کند. ابتدا باید داده را از اینترنت دریافت و آماده سازی کنید. سپس یک معماری CNN شامل لایه های کانولوشن، تابع فعالساز غیرخطی، لایه های پولینگ را پیاده سازی کنید. پس از آن، مدل را با دادگان آموزش، آموزش داده و روند زیان و دقت در مجموعه آموزشی و اعتبارسنجی را رسم نمایید. هدف رسیدن به دقت بالای ٪۷۵ در مجموعه اعتبارسنجی است. در انتها، دقت نهایی مدل را بر روی داده آزمون محاسبه کرده و نمونه هایی از طبقه بندی های درست و نادرست را نشان دهید. علاوه بر این، خروجی های میانی مدل را بررسی و ویژگی هایی را که توسط فیلترها آموخته شده است، تحلیل کنید.

سوال سوم

در این نوت‌بوک الگوریتم تشخیص اشیا YOLO (You Only Look Once) را با استفاده از PyTorch پیاده‌سازی خواهید کرد. در این تمرین، شما باید معماری شبکه عصبی کانولوشنال TinyYOLOv2 را با لایه های مختلف پیاده‌سازی کنید. همچنین باید خروجی مدل را برای نمایش جعبه‌های محصور کننده اشیا بر روی تصاویر ورودی پردازش کنید. این شامل فیلتر کردن جعبه‌ها بر اساس امتیازات اطمینان، محاسبه تقاطع بر اتحاد (IoU) جعبه‌ها و

حذف جعبه‌های تکراری با استفاده از الگوریتم حذف حداکثر غیرمتقاطع می(NMS) باشد. در نهایت باید نتایج نهایی مدل را بر روی مجموعه‌ای از تصاویر نمایش دهید.