به نام خدا

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی برق

یادگیری عمیق

پاییز ۱۴۰۱
استاد: دکتر فاطمی‌زاده

گردآورندگان: ـ

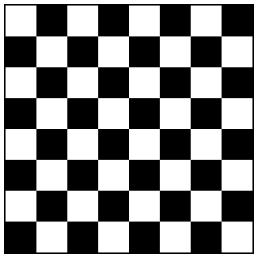
تمرین سوم | مهلت ارسال: چهارشنبه ۷ دی

- مهلت ارسال پاسخ تا ساعت ۲۳:۵۹ روز مشخص‌شده است.
- در طول ترم امکان ارسال با تاخیر پاسخ همه‌ی تمارین تا سقف ۲۰ روز و در مجموع ۲۰ روز، وجود دارد. پس از گذشت این مدت، پاسخ‌های ارسال‌شده پذیرفته نخواهند بود. همچنین، به ازای هر روز تأخیر غیر مجاز ۱۰ درصد از نمره تمرین به صورت ساعتی کسر خواهد شد.
- هم‌کاری و هم‌فکری شما در انجام تمرین مانعی ندارد اما پاسخ ارسالی هر کس حتما باید توسط خود او نوشته شده باشد. (دقت کنید در صورت تشخیص مشابهت غیرعادی برخورد جدی صورت خواهد گرفت.)
- در صورت هم‌فکری و یا استفاده از هر منابع خارج درسی، نام هم‌فکران و آدرس منابع مورد استفاده برای حل سوال مورد نظر را ذکر کنید.
- لطفا تصویری واضح از پاسخ سوالات نظری بارگذاری کنید. در غیر این صورت پاسخ شما تصحیح نخواهد شد.
- نتایج و پاسخ های خود را در یک فایل با فرمت zip به نام HW۳-Name-StudentNumber در سایت Quera قرار دهید. برای بخش عملی تمرین نیز لینگ گیت‌هاب که تمرین و نتایج را در آن آپلود کرده‌اید قرار بدهید. دقت کنید هر سه فایل نوتبوک تکمیل شده بخش عملی را در گیت‌هاب قرار دهید .
- لطفا تمامی سوالات خود را از طریق کوئرای درس مطرح بکنید (برای اینکه تمامی دانشجویان به پاسخ‌های مطرح شده به سوالات دسترسی داشته باشند و جلوی سوالات تکراری گرفته شود، به سوالات در بسترهای دیگر پاسخ داده نخواهد شد).
- دقت کنید کدهای شما باید قابلیت اجرای دوباره داشته باشند، در صورت دادن خطا هنگام اجرای کدتان، حتی اگه خطا بدلیل اشتباه تایپی باشد، نمره صفر به آن بخش تعلق خواهد گرفت.

سوالات تئوری (۲۰۰ نمره)

۱. (۵۰ نمره)

(آ) یک تصویر به شکل صفحه شطرنجی ۸ × ۸ مطابق شکل زیر را در نظر بگیرید.



فرض کنید خانه‌های به رنگ سیاه در این جدول را با عدد صفر و خانه‌های به رنگ سفید را با عدد ۲۵۵ نشان دهیم.

میخواهیم عملکرد فیلتر زیر را روی این تصویر بررسی کنیم.

$$\frac{1}{9}\begin{bmatrix}1 & 1 & 1\\ 1 & 1 & 1\\ 1 & 1 & 1\end{bmatrix}$$

با اعمال این فیلتر روی تصویر نتیجه را بدست آورید. فرض کنید پیکسل های موجود در لبه‌ی تصویر بدون تغییر باقی می‌مانند.

(ب) توضیح دهید این فیلتر به طور تقریبی چه عملی روی تصویر انجام می‌دهد.

۲. (۵۰ نمره) شبکه عصبی کانولوشنی‌ای با لایه های ستون چپ جدول در نظر بگیرید. برای هر کدام از لایه‌های ذکر شده در جدول، ابعاد خروجی لایه و همچنین تعداد پارامترهای هر لایه را وارد نمایید. ابعاد خروجی را به صورت $H \times W \times C$ که به ترتیب نشان دهنده ارتفاع، عرض و عمق خروجی می‌باشد، نمایش دهید. همچنین نحوه نمایش هر لایه مطابق زیر می‌باشد:

- CONVk-$N(S, P)$: یک لایه کانولوشی با N فیلتر، هر کدام به ابعاد $k \times k \times D$ که D عمق لایه قبلی می‌باشد و همچنین با گام (stride) برابر با S و تعداد پدینگ (padding) P. همچنین، در صورت ذکر نشدن مقادیر P و S هر دو را برابر با ۱ در نظر بگیرید.
- POOL-n: نمایش دهنده یک لایه max-pooling با ابعاد $n \times n$، گام n و پدینگ ۰ می‌باشد.
- FLATTEN: ورودی هموار می‌کند، معادل torch.nn.flatten/tf.layers.flatten
- FC-N: نمایش دهنده یک لایه fully-connected با N نورون می‌باشد.

Layer	Output Dimension	Number of Parameters
Input	$32 \times 32 \times 3$	0
CONV3-10		
ReLU		
POOL-2		
CONV3-20(3,2)		
ReLU		
POOL-2		
FLATTEN		
FC-10		

۳. (۵۰ نمره) شبکه عصبی کانولوشنی یک بعدی‌ای مطابق شکل زیر در نظر بگیرید که ورودی‌های آن، ۵ متغیر $x_۱, x_۲, x_۳, x_۴, x_۵$ می‌باشند و خروجی آن $\hat{y}$ می‌باشد که با استفاده از آن مقدار تابع هزینه مطابق رابطه ذکر شده در شکل محاسبه می‌شود.

(آ) کدام یک از متغیرهای داده شده، پارامترهای شبکه می‌باشند؟

(ب) مقادیر $\frac{\partial L}{\partial w_1}$، $\frac{\partial L}{\partial w_2}$ و $\frac{\partial L}{\partial a}$ را بر حسب y، $\hat{y}$ و v_i ها به دست آورید.

(ج) با فرض اینکه داشته باشیم:

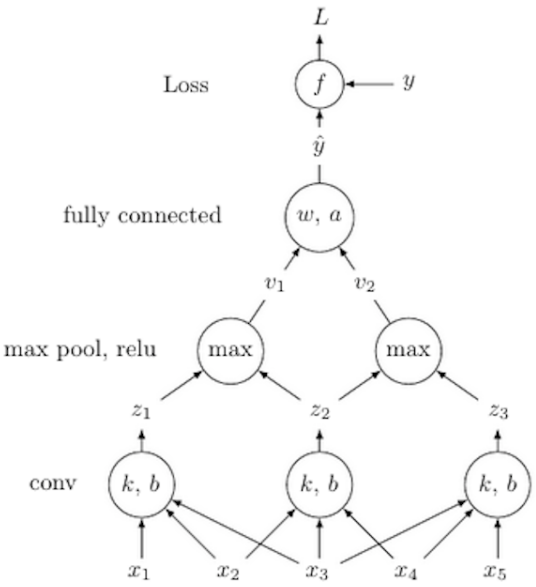
$$\frac{\partial L}{\partial v_1} = \delta_1, \quad \frac{\partial L}{\partial v_2} = \delta_2$$

مطلوب است محاسبه مقادیر $\frac{\partial L}{\partial z_i}$ بر حسب مقادیر z_i و δ_i.

(د) با فرض اینکه داشته باشیم:

$$\frac{\partial L}{\partial z_1} = \alpha_1, \quad \frac{\partial L}{\partial z_2} = \alpha_2, \quad \frac{\partial L}{\partial z_3} = \alpha_3$$

مطلوب است محاسبه مقادیر $\frac{\partial L}{\partial k_1}$، $\frac{\partial L}{\partial k_2}$، $\frac{\partial L}{\partial k_3}$، $\frac{\partial L}{\partial k_4}$و $\frac{\partial L}{\partial b}$ بر حسب x_i و α_i ها.



$$L = \frac{1}{2}(y - \hat{y})^2$$

$$\hat{y} = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} + a$$

$$\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} = \begin{bmatrix} \max\{z_1, z_2, 0\} \\ \max\{z_2, z_3, 0\} \end{bmatrix}$$

$$\begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} k_1 & k_2 & k_3 & 0 & 0 \\ 0 & k_1 & k_2 & k_3 & 0 \\ 0 & 0 & k_1 & k_2 & k_3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} + \begin{bmatrix} b \\ b \\ b \end{bmatrix}$$

شکل ۱: شبکه عصبی کانولوشنی یک بعدی

(ه) در حالت کلی، فرض کنید یک لایه کانولوشن یک بعدی با رابطه زیر داریم:

$$\begin{bmatrix} z_1 \\ \vdots \\ z_m \end{bmatrix} = \begin{bmatrix} k_1 & \dots & k_d & & \\ & k_1 & \dots & k_d & \\ & & & \ddots & \\ & & & k_1 & \dots & k_d \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} b \\ \vdots \\ b \end{bmatrix}$$

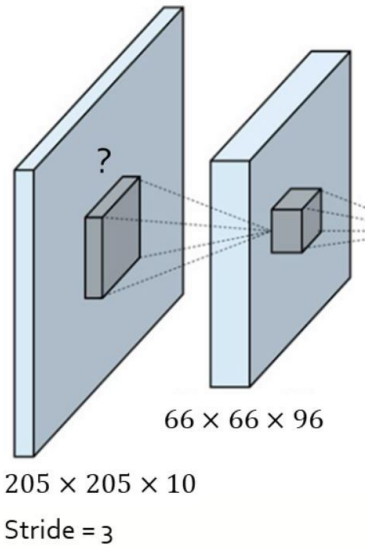
و همچنین می‌دانیم که:

$$\frac{\partial L}{\partial z_i} = \alpha_i$$

مطلوب است محاسبه $\frac{\partial L}{\partial k_j}$ و $\frac{\partial L}{\partial b}$ بر حسب x_i و α_i ها.

۴. (۵۰ نمره)

(آ) با توجه به ابعاد ورودی و خروجی نشان داده شده در شکل زیر، سایز کرنل مورد استفاده در این عملیات کانولوشنی را بدست آورید. لازم به ذکر است که ابعاد ورودی مشخص شده در شکل، با احتساب $zero - padding$ داده شده اند.



(ب) تعداد پارامترهای قابل آموزش یا همان *Learnable* موجود در این لایه کانولوشنی را تعیین نمایید.
(راهنمایی: به پارامتر بایاس موجود در هر کرنل نیز در محاسبات خود توجه داشته باشید.)

(ج) تعداد عملیات ضرب موردنیاز برای بدست آوردن خروجی را محاسبه کنید. (ضرب های در صفر را نیز در شمارش تعداد ضرب ها لحاظ نمایید.)

سوالات عملی (۲۰۰ نمره)

۱. (۱۰۰ نمره) هدف از این تمرین مقایسه ی عملکرد ساختارهای مبتنی بر MLP و CNN در حل یک تسک طبقه بندی مشابه است. قصد داریم با استفاده از تصاویر موجود در این دادگان و ساختاری مبتنی بر MLP، یک‌بار عمل طبقه‌بندی را انجام دهیم. سپس با شبکه‌ای با تعداد پارامتر های تقریبا مشابه ولی با کمک ساختار های پیچشی [۱] بار دیگر آموزش را تکرار کنیم تا با مزایا و معایب هر کدام از ساختارها آشنا شویم. نمونه‌ای از تصاویر این دادگان در شکل ۱ قابل مشاهده است.

شکل ۲

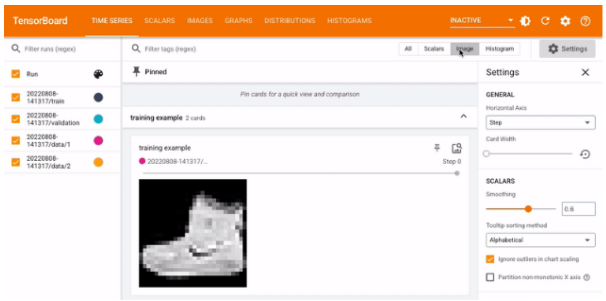
(آ) برای آشنایی بیشتر با دادگان، یک تصویر از هر کلاس را به عنوان نمونه رسم نمایید. سپس ۷۰ درصد نمونه‌ها را به عنوان آموزشی و ۳۰ درصد باقی مانده را به عنوان نمونه تست جدا کنید.

(ب) یک شبکه MLP طراحی کرده و آن را آموزش دهید (می توانید در طراحی خود از لایه Dropout نیز استفاده نمایید تا از بیش‌برازش [۲] شدن شبکه جلوگیری کنید). توجه نمایید که در طول آموزش بهترین مدل را ذخیره کنید. سعی کنید که تعداد پارامترهای شبکه ی شما در حدود پارامترهای شبکه ی پیشنهادی شما در قسمت CNN باشد تا بتوانیم مقایسه ی درستی انجام دهیم.

[۱] Convolutional
[۲] Overfit

- خطا، دقت، Recall و Precision را گزارش نمایید.
- ماتریس درهم ریختگی [۳] را رسم نمایید.
- تعداد پارامترهای شبکه را گزارش کنید.

(ج) تنسوربورد [۴] ابزاریست که به ما امکان مشاهده ی چگونگی تغییر تابع خسارت در طول زمان و یا نحوه ی تغییر وزن ها را فراهم می کند. با استفاده از این ابزار نمودارهای خسارت و دقت را برای داده‌های آموزش و تست رسم نمایید.



شکل ۳

(د) در این قسمت یک شبکه متشکل از لایه های کانولوشنی با معماری دلخواه طراحی کنید به نحوی که تعداد پارامترهای آن از پارامترهای شبکه MLP قسمت قبل بیشتر نباشد (می‌توانید در طراحی خود از لایه‌ی دراپ‌اوت یا نرمال‌سازی دسته‌ای [۵] نیز استفاده نمایید). هایپر پارامترها را مشابه قسمت قبلی انتخاب کنید. توجه نمایید که در طول آموزش بهترین مدل را ذخیره کنید.

- خسارت، دقت، Recall و Precision را گزارش نمایید.
- ماتریس درهم ریختگی را رسم نمایید.
- تعداد پارامترهای شبکه را گزارش کنید.
- با استفاده از تنسوربورد نمودارهای خطا و دقت را برای داده‌های آموزش و تست رسم نمایید.

(ه) نتیجه بدست آمده را با نتایج قسمت قبل مقایسه نمایید.

(و) در شبکه CNN بهتر است به جای لایه دراپ‌اوت از دراپ‌اوت بلوکی [۶] استفاده گردد. علت این موضوع را بیان کنید و لایه دراپ اوت قسمت (آ) را با این لایه جایگزین کرده و مجدداً شبکه را آموزش دهید.

(ز) در درس با فاکتوریزیشن کرنل‌ها آشنا شدیم (تکنیکی که بر اساس آن به طور مثال یک فیلتر ۳×۳ به دو فیلتر ۱×۳ و ۳×۱ متوالی تبدیل می‌شود).

- معماری شبکه ی کانولوشنی خود را به این منظور بروز کنید و مجددا شبکه را آموزش دهید.
- تعداد پارامترها را با قسمت (آ) مقایسه کنید.
- به طول کلی استفاده از این تکنیک چه مزایایی دارد؟

۲. (۱۰۰ نمره) در بسیاری از مواقع برای کاربری‌های کوچک، استفاده از مدل‌های بزرگ شبکه عصبی مقرون به صرفه نیست و برای مثال با محدودیت‌هایی در استفاده از منابع محاسباتی مواجه هستیم. در چنین مواقعی مجبور هستیم تا با فدا کردن مقداری از دقت، به استفاده از مدل‌های کوچک‌تر روی بیاوریم و دادگان خود را روی مدلی کوچک‌تر آموزش دهیم.

---
[۳] Confusion Matrix
[۴] Tensorboard
[۵] Batch Normalization
[۶] Block Dropout

طبیعتا با کوچک‌تر کردن مدل، یادگیری بسیاری از خواص پیچیده دادگان برای مدل دشوار می‌شود. در اینجا البته یکی از کارهایی که در راستای بهبود دقت مدل می‌توانیم انجام دهیم، پیاده‌سازی روش Knowledge-Distillation و استفاده از یک مدل بزرگ‌تر به عنوان آموزگار است. در این روش از یک مدل بزرگ‌تر که روی دادگان مرجع از پیش آموزش دیده‌است استفاده می‌کنیم تا مدل کوچکتر را آموزش دهیم. در مقاله‌ای از هینتون [۷] این روش به طور دقیق‌تر بررسی شده‌است و در اینجا می‌توانید توضیحات بیشتر در مورد پیاده‌سازی این روش را مطالعه کنید.

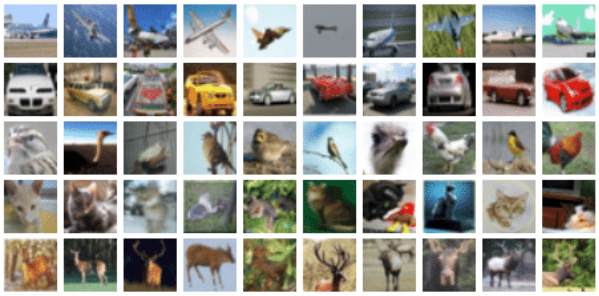
به طور خلاصه، در این روش، لاجیت‌های[۸] مدل بزرگ‌تر به عنوان برچسب و به جای برچسب‌های اصلی داده ورودی در دادگان، به مدل کوچکتر داده می‌شود تا آن را یادبگیرد. تابع خسارت نهایی به صورت زیر است:

$$\mathcal{L}(x;W) = (1-\alpha) * \mathcal{H}(y, \sigma(z_s; T=1)) + \alpha * \tau^2 \mathcal{H}(\sigma(z_t; T=\tau), \sigma(z_s, T=\tau))$$

که در اینجا $\mathcal{H}$ مقدار Cross-Entropy ، σ تابع Softmax و z مقدار لاجیت مدل است. α ضریبی برای ترکیب تابع خسارت عادی و distiller است. T نیز temperature اعمال شده روی لاجیت پس از softmax است که این مفهوم در مقاله هینتون معرفی شده است:

$$\sigma(z_i; T) = \frac{exp\left(\frac{z_i}{T}\right)}{\sum_j \exp\left(\frac{z_j}{T}\right)}$$

در این تمرین می‌خواهیم یک شبکه عصبی با معماری ResNet۱۸ برای مسئله طبقه‌بندی CIFAR-۱۰ (شکل ۳) تمرین دهیم.



شکل ۴

(آ) یک مدل از پیش آموزش دیده ResNet۵۰ روی ImageNet را آماده کنید. سپس لایه Fully-Connected نهایی آن را با یک لایه با سایز مناسب برای مسئله CIFAR-۱۰ تعویض کنید. حال پارامترهای دیگر شبکه را ثابت نگه‌دارید و تنها لایه آخر را روی دادگان CIFAR-۱۰ تمرین دهید و ارزیابی کنید. [۹]

(ب) حال مدلی که در قسمت قبل آماده کردید را به عنوان مدل آموزگار انتخاب کنید و یک مدل ResNet۱۸ را از صفر روی CIFAR-۱۰ آموزش دهید و ارزیابی کنید. برای انتخاب هایپرپارامتر τ و α آزمون انجام دهید و تا جای ممکن بهترین مقدار را انتخاب کنید. (توجه کنید که لزومی ندارد برای انتخاب هایپر پارامتر، هر بار فرایند آموزش را با تعداد epoch زیاد انجام دهید.)

(ج) حال یک‌بار مدل ResNet۱۸ را از صفر و بدون آموزگار روی CIFAR-۱۰ آموزش دهید و ارزیابی کنید. دلیل تفاوت را توضیح دهید.

(د) در صورتی که در قسمت آ به جای آموزش لایه آخر، کل مدل را Fine-tune می‌کردیم چه اتفاقی می‌افتاد؟ این آزمایش را انجام دهید و ارزیابی کنید و دلیل تفاوت را گزارش کنید.

---

[۷] Distilling the Knowledge in a Neural Network

[۸] مقادیر خروجی نهایی یک شبکه پس از عبور از لایه Fully-Connected را لاجیت می‌نامند.

[۹] به این عمل linear-tuning می‌گوییم