به نام خدا

دانشکده‌ی مهندسی کامپیوتر

مفاهیم پیشرفته در یادگیری ماشین

نیم‌سال دوم ۴۰۱ـ۱۴۰۰

مدرس: دکتر مهدیه سلیمانی

تمرین سری اول | عنوان تمرین | زمان تحویل: ۲۵ اسفندماه

لطفا نکات زیر را رعایت کنید:

- سوالات خود را از طریق پست مربوط به تمرین در Quera مطرح کنید.
- در هر کدام از سوالات، اگر از منابع خاصی استفاده کرده‌اید باید آن را ذکر کنید.
- پاسخ ارسالی واضح و خوانا باشد.
- تمام پاسخ‌های خود را در یک فایل با فرمت HW#_[SID]_[Fullname].zip روی کوئرا قرار دهید.
- برای ارسال هر تمرین تا ساعت ۲۳:۵۹ روز ددلاین فرصت دارید. علاوه بر آن، در هر تمرین می توانید تا سقف هفت روز از تأخیر مجاز باقیمانده‌ی خود استفاده کنید.
- برای کسب نمره کامل در این تمرین کافیست ۱۶۰ نمره را دریافت نمایید، ما بقی نمرات امتیازی می باشند (۴۰ نمره امتیازی).

سوال ۱: بهینه‌سازی در یادگیری چند وظیفه‌ای (۱۵ نمره)

در سالیان اخیر الگوریتم‌های یادگیری عمیق و یادگیری تقویتی عمیق، توانسته‌اند پیشرفت‌های فوق العاده‌ای در زمینه حل بسیاری از مسائل پیچیده کسب نمایند. با این حال این روش‌ها هنوز به شدت وابسته به حجم زیاد داده هستند و در صورت کم بودن مجموعه داده آموزشی، عملکرد آن‌ها با افت جدی روی وظیفه هدف مواجه می‌شود. یکی از راهکارهای پیشنهادی برای حل این مسئله استفاده از رویکرد یادگیری چندوظیفه‌ای است و این تکنیک به این امید استفاده می‌شود که یادگیری ساختار داخلی مشترک میان داده‌های وظایف مختلف، بتواند به بهبود عملکرد کلی مدل روی تمامی وظایف منجر شود. با این حال وجود پیچیدگی‌هایی در روش‌های بهینه‌سازی شبکه‌های عصبی باعث شده است تا این هدف آن طور که باید محقق نشود و بهینه‌سازی همزمان روی وظایف مختلف اثربخش نباشد . برای روشن‌تر شدن این موضوع، این مقاله را مطالعه نموده و به پرسش‌های زیر پاسخ دهید:

(آ) سه عاملی که در صورت بروز همزمان می‌توانند باعث خراب شدن بهینه‌سازی شوند را معرفی کنید. (۹ نمره)

(ب) الگوریتم PCGrad که برای حل این مشکل معرفی شده است را توضیح دهید. (۶ نمره)

سوال ۲: یادگیری چند وظیفه‌ای (۳۵ نمره)

همانطور که در پرسش قبل توضیح داده شد، توجه به گرادیان‌ها ابزاری موثر برای انجام یادگیری چند وظیفه‌ای [1] می‌باشد. در این سوال قصد داریم تا یک کاربرد دیگر این رویکرد را بررسی کنیم. شبکه عصبی عمیقی به صورت f_θ با پارامترهای $\theta \in \mathbb{R}^q$ برای حل یک مسئله یادگیری چندوظیفه‌ای در نظر بگیرید. فرض کنید این شبکه تا کنون برای حل $t-1$ وظیفه [2] $\{\mathcal{T}_1, \mathcal{T}_2, ..., \mathcal{T}_{t-1}\}$ آموزش دیده است. حال قصد داریم تا یک وظیفه جدید مانند $\mathcal{T}_t$ به این شبکه آموزش دهیم بدون آن که عملکرد این مدل روی وظایف قبلی دچار مشکل شود. اما از سویی به دلیل محدودیت‌های منابع، قادر به نگه‌داری کل مجموعه دادگان وظایف قبلی نیستیم و تنها می‌توانیم به تعداد محدود m داده به ازای هر وظیفه از مجموعه وظایف پیشین در یک حافظه جانبی کوچک ذخیره سازی نماییم. اگر داده‌های ذخیره شده در این حافظه از وظیفه k ام را با $\mathcal{M}_k$ نمایش دهیم، در این صورت تابع هزینه تجربی مدل روی وظیفه k به صورت زیر در نظر گرفته می‌شود:

$$\ell(f_\theta, \mathcal{M}_k) = \frac{1}{m} \sum_{(x,y)\in\mathcal{M}_k} \ell(f_\theta(x,k), y) \qquad (۱)$$

در رابطه فوق، x تصویر ورودی شبکه و y برچسب متناظر با آن را نشان می‌دهد. k نیز عددی طبیعی است که شماره وظیفه را به مدل اطلاع می‌دهد.

[1] Multi-Task Learning
[2] Task

بهینه‌سازی مستقیم روی $\{\mathcal{M}_1, \mathcal{M}_2, ..., \mathcal{M}_{t-1}\} \cup \mathcal{T}_t$ می‌تواند باعث بیش‌برازش [3] به داده‌های وظیفه t یا داده‌های اندک داخل حافظه بشود. لذا به دنبال روشی هستیم که بدون بروز این مشکل بتواند وظیفه جدید را یاد بگیرد. روش پیشنهادی به این صورت است: داده (x, y) از وظیفه شماره t را در نظر بگیرید. برای ما ایده‌آل است تا مسئله بهینه‌سازی زیر را حل نماییم:

$$\min_{\theta'} \quad \ell(f_{\theta'}(x, t), y) \tag{۲آ}$$

$$\text{s.t.} \quad \ell(f_{\theta'}, \mathcal{M}_k) \leq \ell(f_\theta, \mathcal{M}_k), \text{ for all } k < t \tag{۲ب}$$

وجود شرط‌های بهینه‌سازی در رابطه فوق باعث می‌شود تا عملکرد مدل روی وظایف قبلی حفظ شود. حال برای آن که مسئله بهینه‌سازی فوق حل شود، می‌توانیم فرض کنیم که تابع هزینه ℓ به صورت محلی تقریب خطی قابل قبولی دارد. لذا اگر به گرادیان توابع هزینه به صورت محلی نگاه کنیم و جهت این گرادیان را به ازای همه k ها کنترل نماییم، می‌توانیم کم شدن تابع هزینه را تضمین نماییم. با استفاده از این رویکرد، کافیست تا گرادیانی که برای آپدیت $f_{\theta'}(x, t), y)$ استفاده می‌شود، با گرادیان تابع هزینه سایر وظایف هم راستا باشد یا اگر احیاناً هم راستا نیست، در جهت آن گرادیانها تصویر شود. لذا شرط ۲ب را می‌توانیم با شرط زیر جایگزین کنیم:

$$\langle g \ , \ g_k \rangle := \langle \frac{\partial \ell(f_\theta(x, t), y)}{\partial \theta} \ , \ \frac{\partial \ell(f_\theta, \mathcal{M}_k)}{\partial \theta} \rangle \geq 0, \text{, for all } k < t \tag{۳}$$

لذا بهینه‌سازی نهایی را می‌توانیم به شکل زیر بنویسیم:

$$\min_{g'} \quad \|g' - g\|_2^2 \tag{۴آ}$$

$$\text{s.t.} \quad \langle g' \ , \ g_k \rangle \geq 0, \text{ for all } k < t \tag{۴ب}$$

با در نظر گرفتن توضیحات فوق به پرسش‌های زیر پاسخ دهید:

(آ) مسئله بهنیه‌سازی ۴ را به صورت یک برنامه‌ریزی مرتبه دو [4] (QP) بازنویسی نمایید. (۱۵ نمره)

(ب) دوگان معادل این مسئله را بازنویسی نمایید. فرم دوگان را تا جای ممکن ساده نمایید به طوری که بهینه‌سازی آن تنها روی یک متغیر انجام شود. نحوه به دست آوردن جواب بهینه‌سازی اولیه از روی جواب مسئله دوگان توضیح دهید. (۱۵ نمره)

(ج) از میان دو فرم اصلی و دوگان ساده شده، کدام یک حل عملیاتی ساده‌تری دارد؟ چرا؟ (۵ نمره)

به طور کلی تمامی گام‌های پاسخ خود را به صورت دقیق توضیح دهید و جزئیات تمام مراحل را با ذکر منابع استفاده شده، مشخص نمایید. به عنوان راهنمایی می‌توانید برای مطالعه مسائل QP و فرم دوگان آن‌ها به این مقاله (Dorn 1960) مراجعه کنید. (به طور خاص مطالعه صفحه ۲ این مقاله و همچنین توجه به Type II معرفی شده در صفحه ۶ می‌تواند موثر باشد)

سوال ۳: انتقال یادگیری (۴۰ نمره)

انتقال یادگیری [5] یکی از الگوهای پرطرفدار در زمینه آموزش شبکه‌های عصبی است. برای این منظور ابتدا شبکه عصبی را روی یک دیتاست حجیم آموزش می‌دهند و سپس آن را روی داده‌های وظیفه هدف به‌سازی [6] می‌کنند. از سویی بهینه‌سازی بدون قید و شرط به داده‌های وظیفه هدف، به خصوص زمانی که تعداد داده‌های مقصد محدود است، می‌تواند در بعضی شرایط باعث از دست رفتن عمومیت بخشی [7] مدل شده و اتفاقاً دقت مدل را پایین آورد. برای حل این مشکل، بهینه‌سازی وظیفه هدف را به صورت مقید در آورده یا از جملات منظم سازی در آن استفاده می‌کنند.
فرض کنید شبکه عصبی f_W با L لایه و با پارامترهای W را در اختیار داریم. ورن‌های این شبکه پیش‌آموزش دیده‌اند و از این طریق نقطه شروع $\hat{W}_i^{(0)}$ به دست آمده است. همچنین $\mathcal{L}(f_W) = \mathbb{E}_{(x,y)}[\ell(f_W(x), y)]$ رابه عنوان تابع هزینه وظیفه مقصد در نظر بگیرید و فرض کنید ℓ تابعی محدب، 1-Lipschitz و از بالا کراندار با کران C_2 است . برای جلوگیری از فاصله گرفتن وزن‌های مدل از نقطه شروع پیش‌آموزش، به‌سازی مقید زیر را پیشنهاد می‌کنیم:

$$\hat{W} = \arg\min_{W} \quad \hat{\mathcal{L}}(f_W) \tag{۵آ}$$

$$\text{s.t.} \quad \left\| W_i - \hat{W}_i^{(0)} \right\|_2 \leq D_i, \forall i = 1, ..., L \tag{۵ب}$$

[3] Overfitting
[4] Quadratic Programming
[5] Transfer Learning
[6] Fine-Tune
[7] Generalization

که منظور از $W_i \in \mathbb{R}^{d_{i-1}\times d_i}$ ماتریس وزن‌ها در لایه i ام می‌باشد (بدیهی است که ابعاد ورودی برابر d_0 خواهد بود). همچنین $\hat{\mathcal{L}}$ تابع هزینه تجربی وظیفه مقصد را روی n داده نشان می‌دهد و D_i ها هایپرپارامترهایی هستند که میزان مقید بودن به نقطه شروع اولیه را نشان می‌دهند.
در ادامه این سوال به دنبال محاسبه کرانی برای خطای عمومیت بخشی $\hat{W}$ هستیم و برای این منظور به خطای $\mathcal{L}(f_{\hat{W}}) - \hat{\mathcal{L}}(f_{\hat{W}})$ توجه کرده و از ابزارهای PAC-Bayes برای آن استفاده می‌کنیم.
با فرض محدود بودن وزن‌های اولیه ($\left\|\hat{W}_i^{(0)}\right\|_2 \leq B_i, B_i > 1, \forall i = 1, ..., L$) و با فرض محدود بودن ورودی ($\|x\|_2 \leq C_1, C_1 \geq 1$) و با در نظر گرفتن H به عنوان عرض شبکه ($H = \max d_i$) کران زیر با احتمال $1 - 2\delta$ برقرار خواهد بود:

$$\mathcal{L}(f_{\hat{W}}) \leq \hat{\mathcal{L}}(f_{\hat{W}}) + \epsilon + C_2\sqrt{\frac{\frac{36}{\epsilon^2}C_1^2 H\log(4LHC_2)(\Sigma_{i=1}^{L}\frac{\prod_{j=1}^{L}(B_j+D_j)}{B_i+D_i})^2(\Sigma_{i=1}^{L}D_i^2) + 3\ln\frac{n}{\delta} + 8}{n}} \tag{۶}$$

که منظور از ϵ یک عدد کوچک مثبت و دلخواه می باشد.
در این تمرین قصد داریم رابطه ۶ را ثابت نماییم. لذا گام به گام به صورت زیر عمل می‌کنیم:

(آ) با توجه به قضایای مطرح شده در این مقاله می‌دانیم که اگر H یک فضای فرضیه باشد، توزیع P یک توزیع پیشین مستقل از داده‌های آموزش روی این فضا باشد و Q توزیع پسین وابسته به داده‌های آموزشی باشد، در این صورت با احتمال $1 - \delta$ باند زیر معتبر است :

$$\mathbb{E}_{h\sim Q}[\mathcal{L}(h)] \leq \mathbb{E}_{h\sim Q}[\hat{\mathcal{L}}(h)] + C_2\sqrt{\frac{KL(Q|P) + 3\ln n/\delta + 8}{n}} \tag{۷}$$

با در نظر گرفتن P به صورت یک توزیع گاوسی حول $\hat{W}^{(0)}$ و Q به صورت یک توزیع گاوسی با میانگین $\hat{W}$ و واریانس‌های $\sigma^2 I$ که σ^2 یک ثابت قابل کنترل است، جمله KL را محاسبه کرده و وجود باند بالای $\frac{\sum_{L}^{i=1} D_i^2}{2\sigma^2}$ را برای آن ثابت کنید. (۱۰ نمره)

(ب) متغیر e را به صورت $e = \sigma\sqrt{2H\log(4L.HC_2)}$ در نظر بگیرید. همچنین ماتریس U را یک ماتریس تصادفی فرض کنید که هر درایه آن به صورت i.i.d. از توزیع گاوسی با میانگین صفر و واریانس σ^2 گرفته شده است. در این صورت فرض کنید رابطه زیر با احتمال $1 - \delta$ برقرار است:

$$\|f_{W+U}(x) - f_W(x)\|_2 \leq eC_1\left(\sum_{i=1}^{L}\frac{\prod_{j=1}^{L}\|W_j\|_2 + e}{\|W_i\|_2 + e}\right) \tag{۸}$$

از این رابطه استفاده کرده و به شرطی که δ به اندازه کافی کوچک باشد نشان دهید: (۱۰ نمره)

$$\mathbb{E}_{h\sim Q}[\hat{\mathcal{L}}(h)] \leq \hat{\mathcal{L}}(f_{\hat{w}}) + 2eC_1\left(\sum_{i=1}^{L}\frac{\prod_{j=1}^{L}\|W_j\|_2 + e}{\|W_i\|_2 + e}\right) \tag{۹}$$

(ج) به طریق مشابه قسمت قبل، می‌توان باند پایین مناسبی برای $\mathbb{E}_{h\sim Q}[\mathcal{L}(h)]$ به صورت زیر به دست آورد:

$$\mathbb{E}_{h\sim Q}[\mathcal{L}(h)] \geq \mathcal{L}(f_{\hat{w}}) - 2eC_1\left(\sum_{i=1}^{L}\frac{\prod_{j=1}^{L}\|W_j\|_2 + e}{\|W_i\|_2 + e}\right) \tag{۱۰}$$

حال از این موضوع و از پاسخ خود به قسمت‌های قبلی استفاده کرده و رابطه زیر را نشان دهید: (۱۰ نمره)

$$\mathcal{L}(f_{\hat{W}}) \leq \hat{\mathcal{L}}(f_{\hat{W}}) + 4eC_1\left(\sum_{i=1}^{L}\frac{\prod_{j=1}^{L}B_j + D_j + e}{B_i + D_i + e}\right) + C_2\sqrt{\frac{\sum_{i=1}^{L}\frac{D_i^2}{2\sigma^2} + 3\ln\frac{n}{\delta} + 8}{n}} \tag{۱۱}$$

(د) با در نظر گرفتن $\sigma = \frac{\epsilon}{6C_1\alpha\sqrt{2H\log(4L.HC_2)}}$ که $\alpha = (\sum_{i=1}^{L}\frac{\prod_{j=1}^{L}B_j+D_j}{B_i+D_i})$ اثبات را تمام کنید. (۱۰ نمره)

سوال ۴: (عملی) انتقال یادگیری (۲۰ نمره)

هدف از این سوال دستگرمی برای یادآوری یادگیری عمیق و تجربه انتقال یادگیری است.

(آ) در این بخش شما بایستی یک دسته‌بند را روی دادگان Cifar10 آموزش دهید. بدین جهت در فایل‌های پیوست ماژولی تحت عنوان ProtoNetBack قرار داده شده است. از این ماژول کانولوشنی در معماری شبکه خود به عنوان Backbone استفاده کنید. بدین صورت که روی خروجی این ماژول دو لایه نهان feed-forward با اندازه دلخواه قرار دهید و سپس لایه دسته‌بند را روی آنها سوار کنید. در آموزش شبکه خود سعی کنید تا از تکنیک‌های افزون‌سازی و همچنین بهینه‌ساز مناسب استفاده کنید تا دقت شبکه‌تان بهتر باشد. انتظار می‌رود دقت شبکه آموزش داده شده بر روی دادگان تست Cifar10 حداقل به میزان هفتاد درصد باشد.

(ب) در فایل‌های پیوستی دادگانی تحت عنوان Tiny Image در دو مجموعه آموزش و تست قرار داده شده اند. در این سوال شما بایستی با استفاده از معماری سوال قبل و تحت سه سناریو زیر یک دسته‌بند را برای این دادگان آموزش دهید:

۱. مساله دسته‌بندی تصاویر Tiny Image را از ابتدا روی معماری مشابه سوال قبل آموزش دهید. در واقع در این سناریو از وزن‌های به دست آمده در سوال قبلی نباید هیچ استفاده ای کنید.

۲. شبکه پیش آموزش دیده شده بر روی Cifar10 را در نظر بگیرید. برای آموزش مساله دسته‌بندی تصاویر Tiny Image در این سناریو، وزن‌های قسمت Backbone را ثابت در نظر بگیرید (فریز کنید) و صرفا به Fine-Tune وزن‌های لایه‌های Feed-Forward بپردازید.

۳. شبکه پیش آموزش دیده شده بر روی Cifar10 را در نظر بگیرید. برای آموزش مساله دسته‌بندی تصاویر Tiny Image در این سناریو، تمامی وزن‌های شبکه را مورد Fine-Tune قرار بدهید.

برای هر سه سناریو بالا نمودار تابع Loss و میزان دقت مدل در حین آموزش را نیز رسم کنید. نتایج و نمودارهای حاصل از هر سه سناریو را با یکدیگر مقایسه کرده و در مورد آن‌ها بحث کنید.

سوال ۵: (عملی) یادگیری چندوظیفه‌ای (۴۰ نمره)

در فایل‌های پیوستی زیرمجموعه‌ای از دادگان Omniglot آورده شده است. این دادگان شامل ۲۰ تصویر از رسم الخط هر یک از کاراکترهای مربوط به سی الفبای متفاوت نظیر لاتین، سانسکریت و ... است. در این شروع ابتدا این مجموعه را به دو دسته آموزش و تست تقسیم کنید. برای این منظور تعداد ۶ عدد از رسم الخط های هر کاراکتر را به عنوان داده تست و ۱۴ عدد باقی مانده را به عنوان داده آموزش جداسازی کنید. در ضمن هر یک از تصاویر دادگان را به سایز ۳۲ در ۳۲ تغییر اندازه دهید. در هر یک از موارد خواسته شده زیر نتایج و دقت روی هر یک از الفبای مختلف را به همراه مجموع تعداد پارامترهای شبکه گزارش کنید.

(آ) در این بخش برای مساله دسته بندی کاراکترها برای هر الفبا یک شبکه دسته بند را به صورت جداگانه آموزش دهید (به بیان بهتر در این بخش شما بایستی سی شبکه عصبی آموزش دهید). طراحی معماری شبکه شما مشابه سوال قبل است. در این جا نیز از ProtoNetBack به عنوان Backbone استفاده کنید و دو لایه نهان Feed-Forward با اندازه دلخواه نیز بر روی خروجی آن اضافه کنید و سپس یک لایه دسته‌بند به اندازه تعداد کاراکترهای الفبای منظور در انتهای شبکه قرار دهید.

(ب) در این بخش قصد داریم تا با استفاده از یادگیری چندوظیفه‌ای به حل مساله بپردازیم. به این منظور از معماری Multi-Head استفاده کنید. این معماری را در دو سناریو زیر مورد آزمایش قرار بدهید:

۱. قسمت Backbone و دو لایه نهان بین تمامی وظیفه‌ها مشترک هستند و تنها لایه آخر (لایه دسته بند) مختص به هر وظیفه است.

۲. قسمت Backbone و لایه نهان اول بین تمامی وظیفه‌ها مشترک هستند و لایه نهان دوم و لایه دسته‌بند مختص به هر وظیفه هستند.

سوال ۶: (عملی) متایادگیرنده مبتنی بر مدل (۴۰ نمره)

در این سوال قصد داریم مدل متایادگیرنده معرفی شده در مقاله Mishra 2017 را که یک مدل متایادگیرنده مبتنی بر مدل می‌باشد را به صورت گام به گام پیاده‌سازی نماییم.
مدل این مقاله از دو ماژول اصلی Temporal Convolution و Attention تشکیل شده است که هدف ماژول Temporal Convolution جمع‌آوری اطلاعات لازم برای پیش‌بینی مدل از تمام داده‌های Support و هدف ماژول Attention بدست آوردن و توجه به بخش‌های اطلاعات‌دار داده می‌باشد.
به صورت کلی مواردی که باید در Notebook داده شده تکمیل کنید به شرح زیر است:

(آ) آماده‌سازی دادگان خود باتوجه به پارامترهای متایادگیری و نحوه استفاده آن در مدل

(ب) پیاده‌سازی شبکه‌های مدل براساس توضیحات داده شده و تکمیلی‌تر موجود در مقاله

(ج) آموزش مدل در فاز meta-training و تست آن در فاز meta-testing

در Notebook داده شده تمام مراحل و کارهایی که باید انجام دهید با دقت تشریح شده است. دادگانی که در این سوال استفاده می‌کنید دادگان Omniglot می باشد که از دادگان رایج در مسئله متایادگیری می باشد. در Notebook داده شده روند آموزش مدل باتوجه به خطای تابع هزینه و دقت روی دادگان ارزیابی گزارش شود. حداقل دقت برای این سوال ۹۳ می باشد.