به نام خدا

درس یادگیری عمیق

تمرین سری دوم

استاد درس : دکتر محمدرضا محمدی

دستیاران : مهدی خورشا، بهداد نادری فرد،

مرتضی حاجی‌آبادی

دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر

نیمسال دوم تحصیلی ۱۴۰۳ - ۱۴۰۴

مهلت تحویل : ۱۴۰۴/۰۱/۱۵

لطفا به نکات موجود در سند قوانین انجام و تحویل تمرین ها دقت فرمایید.

سوالات تئوری

۱.  ابتدا مقاله Shap-CAM را مطالعه کنید و سپس به سوالات زیر پاسخ دهید(۱۰نمره).

(آ) روش‌های Grad-CAM و Shap-CAM را از نظر ویژگی‌های زیر مقایسه کنید:

- نحوه محاسبه اهمیت ویژگی‌ها (ویژگی‌های محلی یا جهانی)
- وابستگی به ساختار مدل و نیاز به گرادیان
- دقت در شناسایی نواحی مهم تصویر
- حساسیت به تغییرات کوچک در ورودی

برای پاسخ به این بخش میتوانید از این مقاله استفاده کنید.

(ب) فرض کنید مدلی که برای طبقه‌بندی تصاویر استفاده می‌کنید، نسبت به تغییرات ناچیز در ورودی حساس است.

- آیا انتظار دارید Grad-CAM و Shap-CAM رفتار مشابهی داشته باشند؟ چرا؟
- کدام روش می‌تواند پایداری بیشتری داشته باشد؟ توضیح دهید.

۲.  به سوالات زیر در مورد شبکه های عصبی پیچشی[1] پاسخ دهید(۱۰ نمره)

(آ) مفهوم به اشتراک‌گذاری پارامترها در شبکه‌های عصبی پیچشی چیست و چه تاثیری در روند آموزش مدل دارد؟

(ب) توضیح دهید برای هریک از سناریوهای زیر شبکه‌های عصبی پیچشی مناسب هستند یا خیر:

- نظارت بر یک گونه‌ی خاص از گرگ در حیات وحش با پهپاد
- استخراج متن از درون صوت
- شناسایی عمل انجام شده درون ویدیو
- داوری انجام حرکت میل‌زنی در مسابقات زورخانه‌ای

(ج) معادله‌ی تلفیق (fusion) لایه‌ی batchNorm2D درون یک لایه‌ی Conv2D را بنویسید و توضیح دهید این عمل چه تاثیری در عملکرد مدل دارد.

(د) یکی از کاربردهای مدل‌های چندوجهی[2] مانند ChatGPT وظیفه‌ی VQA[3] (پرسش و پاسخ تصویر) است. در این وظیفه مدل تصویر و سوالی درباره‌ی تصویر از کاربر می‌گیرد و باید جواب متناسبی برای آن تولید کند. یکی از نقاط ضعف این مدل‌ها، پاسخ دادن به سوالاتی است که از جزئیات ریز و درحاشیه‌ی تصویر پرسیده می‌شوند. این نوع مدل‌ها در پاسخ به سوالات مربوط به تصویر حاوی یک شی برجسته در وسط توانایی خوبی دارند. برای پوشش دادن این ضعف یکی از کارهایی که می‌توان کرد، تشخیص درست ناحیه‌ی مورد پرسش و برش آن برای ورود به مدل است. فرض کنید ما یک مدل Question-image matching توسعه داده‌ایم. اما این مدل تنها میان مفهوم کلی سوالات ورودی و تصویر انطباق انجام می‌دهد و درباره‌ی ناحیه‌ی آن اشاره‌ای نمی‌کند. با استفاده از مفاهیمی که تاکنون خوانده‌اید راه‌حلی برای این مسئله پیشنهاد دهید.

[1] Convolutional
[2] Multi-Modal
[3] Visual question answering

۳.  تعداد پارامتر، ضرب و جمع و هم‌چنین میدان دید موثر لایه‌های شبکه‌ی عصبی با ورودی تصاویر رنگی از ابعاد ۲۵۶ در ۲۵۶ زیر را به تفصیل محاسبه کنید (لطفا اعداد اعشاری را به پایین گرد کنید)(۱۵ نمره)

- `Layer1 : nn.Conv2d(in_channels=3, out_channels=32, kernel_size=(7,7), stride=1, padding='same')`
- `bn1 : nn.BatchNorm2d(32)`
- `Layer2 : nn.Conv2d(in_channels=32, out_channels=64, kernel_size=(5,5), stride=2, padding='valid')`
- `bn2 : nn.BatchNorm2d(64)`
- `Layer3 : nn.AvgPool2d(kernel_size=(2,2), stride=2)`
- `Layer4 : nn.Conv2d(in_channels=64, out_channels=128, kernel_size=(3,3), stride=1, dilation=2, padding='valid')`
- `bn3 : nn.BatchNorm2d(128)`
- `Layer5 : nn.Conv2d(in_channels=128, out_channels=128, kernel_size=(3,3), stride=1, dilation=1, padding='valid')`
- `bn4 : nn.BatchNorm2d(128)`
- `Layer6 : nn.AvgPool2d(kernel_size=(2,2), stride=2)`
- `Layer7 : nn.Conv2d(in_channels=128, out_channels=256, kernel_size=(3,3), stride=1, padding='valid')`
- `bn5 : nn.BatchNorm2d(256)`
- `Layer8 : nn.AvgPool2d(kernel_size=(2,2), stride=2)`
- `fc1 : nn.Linear(in_features=43264, out_features=1024)`
- `fc2 : nn.Linear(in_features=1024, out_features=1024)`
- `dropout : nn.Dropout(p=0.5)`
- `fc3 : nn.Linear(in_features=1024, out_features=10)`

سوالات عملی

۴.  برای انجام این سوال به پوشه‌ی HW2_TM مراجعه کرده و درون فایل نوتبوک پیوست شده، سعی کنید جاهای خالی را پر کنید. برای این سوال از تصاویری که درون همان پوشه قرار داده شده‌اند استفاده کنید.

در این سوال به یکی از مسائل مهم بینایی کامپیوتر به نام تطبیق کلیشه پرداخته‌ایم. در این مسئله دو نوع ورودی به نام‌های تصویر کلیشه و تصویر جست‌وجو داریم که هدف یافتن تصویر کلیشه درون تصویر جست‌وجو و برجسته‌سازی آن با رسم مستطیل به دور شی یافته شده است. یکی از ابتدائی‌ترین روش‌های انجام این مسئله این است که تصویر جست‌وجو را به نواحی‌ای تقسیم‌بندی کرده و شباهت هر یک را با تصویر کلیشه بسنجیم. اما انجام این کار دارای چالش‌های فراوانی است اعم از: کند بودن فرایند، احتمال وجود تغییرات زیاد میان کلیشه و جست‌وجو و ... . از این رو روش های مبتنی بر شبکه‌های عصبی پیچشی برای این مسئله پیشنهاد شدند که دارای دقت عملکردی بالا در مدت زمان معقولی بودند.

بیشتر کد درون نوتبوک برای شما به‌صورت آماده آورده شده است. هدف از این سوال این است که آن را مطالعه کنید و درون گزارشی توضیح دهید که شبکه‌های عصبی پیچشی درون این کد چگونه به حل این مسئله کمک کرده‌اند (از آوردن جزئیاتی مانند: نحوه‌ی محاسبه‌ی confidence، توابع کمکی، توابع رسم نتایج، NMS و ... بپرهیزید و تنها اشاره کنید شبکه‌های عصبی پیچشی چگونه دقت و سرعت این مسئله را افزایش داده‌اند)

خرجی‌های مورد انتظار درون نوتبوک فراهم شده‌اند(۱۵ نمره).

۵.  در این سوال قرار است برای مجموعه‌ی داده‌ی زیر برای شناسایی اعداد دست‌نویس از روی تصویر ورودی، یک شبکه‌ی عصبی پیچشی با معماری دلخواه توسعه دهید. تصاویر این مجموعه‌داده، تصاویر رنگی ۶۴ در ۶۴ تایی از اعداد انگلیسی ۱ تا ۴ هستند که باید توسط شبکه‌های عصبی پیچشی آن‌ها را شناسایی کنند. این تصاویر برچسب ندارند و از روی اسم هر فایل باید ساخته شود. در شکل ۱ نمونه‌ای از این تصاویر برای شما آورده شده است(۱۵ نمره)

می‌توانید برای این مجموعه‌داده رویه‌های مختلف داده‌افزایی را اعمال کنید.

برای انجام آن به نوتبوک HW2_CNN.ipynb که به‌همراه سوالات پیوست گردیده است رجوع کرده

شکل ۱: نمونه ای از تصویر عدد ۴

و درون آن سعی کنید نواحی خالی را پر کنید.

در این سوال انتظار می‌رود بتوانید مدلی را توسعه دهید که برای مجموعه‌داده‌ی آموزشی و آزمایشی (با نرخ ۸۰ به ۲۰ درصد از کل مجموعه‌داده با random seed = 42 برای جداسازی) به دقت بالای ۹۰ درصد دست یابید. لطفا ابرپارامترهای مورد نیاز را برای احقاق نیازمندی‌های پروژه تنظیم کنید. در معماری مدل مختار هستید و می‌توانید از هر نوع مدلی استفاده کنید.

پیشنهاد می‌شود از callback هایی مانند early stopping و learning rate scheduler برای بهبود روند آموزش مدل استفاده کنید. (در استفاده نکردن از آن‌ها آزاد هستید.)

درون نوتبوک رویه‌ی ساخت برچسب واقعی برای هر تصویر برای شما پیاده‌سازی شده است.

لینک مجموعه داده

۶. مقاله ResNeXt را مطالعه کنید و سعی کنید به دلیل موفقیت خلاقیت به کار رفته در آن خوب فکر کنید. در این سوال میخواهیم یک بلاک مشابه بلاک معرفی شده در مقاله پیاده سازی کرده و به کمک آن یک شبکه کامل بسازیم و سپس آنرا با دیتاست cifar100 آموزش دهیم.

به نوتبوک Resnext.ipynb مراجعه کنید. ابتدا یک کلاس برای بلاک resnext طراحی کنید. سپس یک کلاس برای طراحی کامل شبکه بنویسید.در این بخش نه تنها نیازی نیست به شبکه های معرفی شده در مقاله(مانند resnext29) وفادار باشید، بلکه توصیه میشود در طراحی خلاقیت خود را به کار بگیرید. در ساماندهی شبکه مادامی که به ایده اصلی مقاله پایبند باشید پیاده سازی شما مورد قبول است.

در قسمت بعد پیش پردازش مناسب روی داده ها انجام دهید و دیتا لودرهای مورد نیاز خود را بسازید. در این قسمت تمام دانشی که در کلاس درس دراین مورد به دست آورده اید به کار بگیرید. در قسمت بعد آموزش مدل را شروع کنید. مدیریت نرخ یادگیری، نگهداری بهترین مدل و رگولاریزیشن مناسب از جمله مواردی هستند که باید به آنها توجه کافی داشته باشید. در قسمت پایانی دقت مدل را روی مجموعه دادگان تست اندازه گرفته و گزارش کنید(۱۵ نمره).

توجه : تسلط شما به مقاله و ایده آن در زمان ارایه سنجیده خواهد شد و نیازی به گزارش نویسی برای آن نیست.

۷.  در این تمرین قصد داریم به مصورسازی آنچه یک شبکه پیش آموخته در مورد یک کلاس خاص فکر میکند بپردازیم. در کلاس درس یک روش برای بهینه سازی ورودی به قصد تولید تصویر بیشینه کننده احتمال تعلق به یک کلاس برای شبکه های پیش آموخته معرفی شد. در صورت نیاز به جزییات بیشتر در مورد این روش میتوانید به این لینک مراجعه کنید. در این روش ابتدا یک کلاس(مثلا فلامینگو) را در نظر میگیرید و با شروع از یک تصویر تصادفی و بهینه سازی آن به تصویری میرسید که شبکه آن را متعلق به کلاس مربوطه بداند.

در قسمت دوم سوال قصد داریم با استفاده از حمله fgsm تصویری را که به نظر ما و شبکه متعلق به یک کلاس است را با کمترین تغییر به تصویری تبدیل کنیم که به نظر شبکه متعلق به کلاس مورد نظر نباشد. در صورت نیاز میتوانید این لینک را درمورد حمله fgsm مطالعه کنید.

یک کلاس مشخص برای هر دوقسمت سوال در نظر بگیرید و تصویر به دست آمده از دو بخش را باهم مقایسه کنید(۲۰ نمره).

سوال امتیازی

۸.  در این سوال قصد داریم یک مدل ناحیه‌بند معنایی برای تصاویر X-Ray موجود برای شناسایی نواحی پوسیدگی دندان توسعه دهیم. مسئله‌ی ناحیه‌بندی معنایی یکی از شناخته‌شده‌ترین مسائل بینایی کامپیوتر می‌باشد که در آن هدف دسته‌بندی تمامی پیکسل‌های موجود در تصویر است. شبکه‌های عصبی پیچشی با ساختاری خاص نشان‌داده‌اند که دارای قدرت مناسبی برای حل این مسئله هستند. مدل‌های SS انواعی از شبکه‌های عصبی پیچشی هستند که ابتدا تصویر ورودی را رمزگذاری می‌کنند تا بتوانند ویژگی‌های مناسب از تصویر ورودی را استخراج کنند. سپس در رویه‌ای به نام رمزگشایی از روی ویژگی‌های ساخته شده، سعی می‌کنند به فضای اولیه‌ی تصویر ورودی (با همان ابعاد یکسان) بازگردند و پیکسل‌های تصویر را دسته‌بندی می‌کنند. یکی از انواع موفق شبکه‌های عصبی پیچشی در این حوزه، مدل‌های Unet می‌باشند که ابتدا به‌صورت تدریجی تصویر را رمزگذاری کرده و سپس به‌صورت تدریجی ابعاد ویژگی‌های استخراج شده را به ابعاد تصویر اولیه بازمی‌گردانند تا عملیات دسته‌بندی را انجام دهند و در این میان تعداد مشخصی ارتباط میان رمزگذار و رمزگشا برقرار می‌کند تا بتواند از ویژگی‌های سطح پایین در این مسئله بهره ببرد. در این مسئله برچسب ورودی تصویر هم بعد تصویر ورودی با تعداد چنل ۱ است که مقدار پیکسل‌های آن نشان‌دهنده‌ی کلاس آن پیکسل می‌باشد. از موارد کاربرد این مدل‌ها می‌توان به مسائل پزشکی و یافتن نواحی

دارای تومور، پوسیدگی و ... از روی تصاویر پزشکی اشاره کرد.

در این سوال نوع خاصی از مدل Unet به نام ResUNet در نظرگرفته شده است. این مدل ساختار یکسانی با مدل Unet اصلی داراست با این تفاوت که رمزگذار آن یک مدل پیچشی Residual است (دارای ارتباطات residual درون رمزگذار است.). به نوتبوک HW2_SS.ipynb رجوع کرده و تلاش کنید با پر کردن جاهای خالی این وظیفه را به درستی پیاده‌سازی کنید.

در نظر داشته باشید که از مدل فوق تنها قسمت رمزگذار آن از شما خواسته شده و باقی قسمت‌های مدل برای شما پیاده‌سازی شده‌اند.

مجموعه‌داده‌ی این مسئله درون نوتبوک قرار داده شده است و با اجرای سل مربوط، مجموعه‌داده برای شما دانلود خواهد شد. اما در نظر داشته باشید که برخی از برچسب‌های این مجموعه‌داده به شدت نویزی شده‌اند که می‌تواند روند آموزش مدل شما را با مشکل روبرو کنند. برای مدیریت آن‌ها رویه‌ی متناسبی در نظر بگیرید.

سعی کنید تابع هزینه (loss function) مناسبی برای این مسئله توسعه دهید. این مسئله دسته‌بندی هر پیکسل میان دو کلاس (۰ یعنی پوسیده نیست و ۱ یعنی پوسیده هست) می‌باشد. یکی از توابع هزینه‌ی مناسب برای این کار BCE می‌باشد و برای این نوع مسائل نیز استفاده می‌شود. اما استفاده از آن باید برای این مسئله بهینه شود (استفاده از تابع هزینه‌ی BCE اولیه و بدون تغییر جواب خوبی نخواهد داد). (برای استفاده از سایر توابع هزینه‌ای که در کلاس تدریس نشده‌اند حتما منبع آن را ذکر کنید.)

رسیدن به dice score بالای ۴۰ درصد هم برای تصاویر آموزشی و هم آزمایشی الزامی است(۲۰ نمره).