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)

Shap-CAM: این روش اهمیت هر پیکسل را با تخمین سهم حاشیه‌ای آن در خروجی مدل (اطمینان[1] از کلاس) تعیین می‌کند. برای این کار، مفهوم مقدار شاپلی[2] از نظریه بازی‌های مشارکتی[3] معرفی شده است. مقدار شاپلی برای یک پیکسل با در نظر گرفتن سهم آن در تمام زیرمجموعه‌های ممکن محاسبه می‌شود. این فرآیند به طور ذاتی روابط و تعاملات بین پیکسل‌های مختلف را لحاظ می‌کند. این روش یک نسبت‌دهی در سطح ویژگی ارائه می‌دهد که سهم کلی ویژگی‌های ورودی را مشخص می‌کند.

Grad-CAM: این یک روش مبتنی بر گرادیان است که اهمیت هر کانال از نقشه‌های فعال‌سازی لایه نهایی کانولوشن را با استفاده از گرادیان اطمینان کلاس نسبت به این نقشه‌ها محاسبه می‌کند. این گرادیان‌ها نشان می‌دهند که هر مکانی چقدر به پیش‌بینی

[1] Confidence

[2] Shapley

[3] Cooperative

کلاس کمک می‌کند. نقشه نهایی برجستگی، یک جمع وزنی از نقشه‌های فعال‌سازی است. **Grad-CAM** توضیحات مکانی ارائه می‌دهد و مناطقی از فعال‌سازی بالا را که به نواحی مهم در ورودی مربوط هستند، برجسته می‌کند. این روش سهم را براساس فعال‌سازی می‌سنجد، اما مانند مقدار شاپلی به طور صریح تعامل بین ویژگی‌ها یا پیکسل‌ها را در نظر نمی‌گیرد. **Grad-CAM** فاقد دقت و جزئیات اهمیت فردی ویژگی‌هاست که روش‌های مبتنی بر شاپلی فراهم می‌کنند.

- وابستگی به ساختار مدل و نیاز به گرادیان:

Shap-CAM: یک روش توضیح بصری پساحسابی[۴] است. ویژگی کلیدی این روش این است که وابستگی به گرادیان‌ها را حذف می‌کند. این روش یک رویکرد بدون گرادیان است که اهمیت پیکسل را از طریق مقدار شاپلی به‌دست می‌آورد، نه از طریق گرادیان‌ها. این به این معناست که عملکرد آن به پایداری یا قابلیت اعتماد گرادیان‌ها وابسته نیست.
Grad-CAM: نیز یک روش پساحسابی است. این روش اساساً یک رویکرد مبتنی بر گرادیان است. این روش به گرادیان‌هایی که از طریق پس‌انتشار محاسبه می‌شوند، برای ترکیب نقشه‌های ویژگی کانولوشنی تصحیح‌شده[۵]، متکی است. **Grad-CAM** گرادیان‌ها را به‌عنوان میزان اهمیت هر کانال نسبت به احتمال کلاس در نظر می‌گیرد. این به این معناست که برای دسترسی به گرادیان‌های خروجی مدل نسبت به نقشه‌های ویژگی لایه کانولوشنی نیاز دارد و به آن‌ها وابسته است.

- دقت در شناسایی نواحی مهم تصویر:

Shap-CAM: عملکرد بصری بهتر و عدالت بیشتری در تفسیر فرآیند تصمیم‌گیری نسبت به روش‌های موجود نشان می‌دهد. به دلیل برتری مقدار شاپلی و توجه آن به روابط بین پیکسل‌ها، **Shap-CAM** توضیحات منطقی‌تر و دقیق‌تری از سهم هر پیکسل ارائه می‌دهد. از نظر کیفی، **Shap-CAM** نقشه‌های برجستگی هموارتر و با نویز تصادفی کمتری نسبت به روش‌های مبتنی بر گرادیان تولید می‌کند. از نظر کمی نیز در وظایف شناسایی (کاهش بیشتر Average Drop و افزایش بیشتر Average Increase) و وظایف مکان‌یابی (نسبت بالاتر) عملکرد بهتری نسبت به روش‌های قبلی مانند **Grad-CAM** و Grad-CAM++ دارد. این نتایج نشان می‌دهد که **Shap-CAM** می‌تواند فرآیند تصمیم‌گیری مدل را با دقت بیشتری آشکار کند و نقشه‌های برجستگی تولیدشده نویز کمتری دارند.
Grad-CAM: شفافیت مدل‌های CNN را با نمایش نواحی ورودی مهم برای پیش‌بینی

[4] post-hoc
[5] Rectified

بهبود بخشید. با این حال، روش‌های مبتنی بر گرادیان مانند **Grad-CAM** به دلیل فرضیه‌های اثبات‌نشده در مورد وزن‌های نقشه‌های فعال‌سازی، نمی‌توانند اطلاعات اصیل[۶] را به درستی نمایش دهند. این روش‌ها فاقد پایه نظری مستحکم بوده و به عنوان روش‌هایی توصیف شده‌اند که به اندازه کافی مقاوم[۷] و قابل اعتماد نیستند.

- حساسیت به تغییرات کوچک در ورودی:

Shap-CAM: مقالات اشاره می‌کنند که روش‌های مبتنی بر گرادیان مانند **Grad-CAM** می‌توانند به راحتی توسط دستکاری‌های مخرب مدل که گرادیان‌ها را بدون ایجاد تغییرات محسوس در تصویر تغییر می‌دهند، فریب داده شوند. در مقابل، **Shap-CAM** می‌تواند اثر این مشکل را کاهش دهد، زیرا اهمیت پیکسل‌ها را با دقت بیشتری تخمین می‌زند؛ بخشی از این دقت به دلیل پایداری مفهوم مقدار شاپلی و توجه آن به روابط بین پیکسل‌هاست.
Grad-CAM: روش‌های مبتنی بر گرادیان برای تولید نقشه‌های کلاس-فعال‌سازی به عنوان روش‌هایی توصیف شده‌اند که به اندازه کافی مقاوم و قابل اعتماد نیستند. این روش‌ها می‌توانند به راحتی با تغییر گرادیان‌ها بدون تغییرات محسوس در تصاویر اصلی فریب داده شوند. این موضوع نشان‌دهنده حساسیت این روش‌ها به تغییرات کوچک ورودی است که می‌تواند گرادیان‌ها را به طور قابل توجهی تحت تأثیر قرار دهد.

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؟

ما لزوماً انتظار نداریم که **Shap-CAM** و **Grad-CAM** رفتار مشابهی داشته باشند، به‌ویژه زمانی که حساسیت مدل ناشی از تغییرات کوچک ورودی و تأثیر آن‌ها بر گرادیان‌ها باشد. **Grad-CAM** یک روش مبتنی بر گرادیان است که اهمیت ویژگی‌ها را با استفاده از گرادیان‌های اعتماد به کلاس نسبت به نقشه‌های فعال‌سازی تعیین می‌کند. مقالات اشاره می‌کنند که روش‌های مبتنی بر گرادیان مقاومت کافی ندارند و می‌توانند با دستکاری گرادیان‌ها بدون تغییر محسوس تصویر فریب داده شوند. بنابراین، اگر تغییرات کوچک ورودی گرادیان‌ها را به‌شدت تحت تأثیر قرار دهد، توضیحات **Grad-CAM** نیز ناپایدار

[6] Authentic
[7] Robust

خواهند بود. **Shap-CAM**، در مقابل، روشی بدون استفاده از گرادیان است که اهمیت پیکسل‌ها را بر اساس مقدار شاپلی تخمین می‌زند و تعاملات بین پیکسل‌ها را در نظر می‌گیرد. چون به گرادیان‌ها وابسته نیست، حساسیت آن به پایداری گرادیان‌های مدل مرتبط نیست و توضیحات پایدارتر و دقیق‌تری ارائه می‌دهد. در نتیجه، در شرایطی که حساسیت مدل به رفتار گرادیان‌ها مربوط باشد، **Shap-CAM** می‌تواند نسبت به **Grad-CAM** توضیحاتی پایدارتر و قابل اعتمادتر ارائه کند.

- کدام روش می‌تواند پایداری بیشتری داشته باشد؟ توضیح دهید.

بر اساس مقالات، انتظار می‌رود که **Shap-CAM** در مواجهه با حساسیت مدل نسبت به تغییرات کوچک ورودی، پایدارتـر عمل کند. شواهدی که مقالات برای پایداری بیشتر **Shap-CAM** ارائه داده‌اند شامل موارد زیر است: روش‌های مبتنی بر گرادیان، از جمله **Grad-CAM**، به عنوان «غیرمقاوم و غیرقابل اعتماد» توصیف شده‌اند. این روش‌ها «به راحتی می‌توانند با دستکاری گرادیان‌ها بدون تغییر محسوس تصویر فریب داده شوند»، که نشان‌دهنده ناپایداری در برابر تغییرات ورودی است که بر گرادیان‌ها اثر می‌گذارند. در مقابل، **Shap-CAM** با استفاده از مقدار شاپلی، وابستگی به گرادیان‌ها را حذف می‌کند. مقدار شاپلی، مفهومی از نظریه بازی‌های مشارکتی با پایه نظری قوی و ویژگی‌هایی چون کارایی [8] است که تعامل بین ویژگی‌ها را نیز در نظر می‌گیرد و می‌تواند توضیحات منطقی‌تر و دقیق‌تری ارائه کند. مقالات بیان می‌کنند که **Shap-CAM** می‌تواند «اثر این مشکل» (دستکاری گرادیان‌ها) را کاهش دهد و اهمیت پیکسل‌ها را دقیق‌تر تخمین بزند. به صورت کیفی، **Shap-CAM** نقشه‌های برجستگی [9] صاف‌تر و با نویز کمتری نسبت به روش‌های مبتنی بر گرادیان تولید می‌کند که می‌تواند نشانه‌ای از پایداری بیشتر باشد.

۲.  به سوالات زیر در مورد شبکه های عصبی پیچشی [10] پاسخ دهید(۱۰ نمره)

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

در پاسخ این قسمت از فایل آقای حسین‌زاده استفاده شده است

[8] Efficiency
[9] saliency maps
[10] Convolutional

در شبکه‌های عصبی پیچشی، از فیلترهایی استفاده می‌شود که در سراسر تصویر حرکت می‌کنند (عملیات کانولوشن). این فیلترها دارای مجموعه ای از وزن ها (پارامترها) هستند که در کل تصویر یکسان باقی می‌مانند. اگر می‌خواهیم یک ویژگی را تشخیص بدهیم، می‌توانیم از همان آشکارساز در گوشه پایین سمت چپ تصویر و در سمت راست بالای تصویر استفاده کنیم. در شبکه‌های Fully Connected، هر نورون به همه‌ی ورودی‌ها وصل است و وزن مخصوص به خودش را دارد. ولی در CNN به خاطر به اشتراک‌گذاری، تعداد وزن‌ها خیلی کمتر می‌شود. در واقع این ویژگی باعث می‌شود سرعت یادگیری بیشتر شود چون تعداد وزن‌هایی که باید به روزرسانی شوند کمتر است. هم‌چنین مدل سبک‌تر شده و حافظه کمتری مصرف می‌شود. به‌علاوه مدل باید سعی کند ویژگی‌هایی را یاد بگیرد که در کل تصویر کاربرد دارد بنابراین پایداری وgeneralization هم افزایش میابد.

(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:

در پاسخ این قسمت از فایل آقای مرادی استفاده شده است

بله، زیرا CNNها در تشخیص اشیا و دسته بندی تصاویر، به دلیل داشتن فیلترهای متفاوت برای یافتن الگوهایی ویژه (مانند رنگ بدن، شکل گوش و ...)، توانایی بالایی دارند و برای شناسایی حیوانات در تصاویر طبیعی بسیار مناسب‌اند.

- استخراج متن از درون صوت:

در پاسخ این قسمت از فایل آقای مرادی استفاده شده است

بله، زیرا شبکه‌های عصبی کانولوشنی این قابلیت را دارند که با استفاده از فیلترها و عملیات ریاضی کانولوشن، الگوهای محلی موجود در سیگنال صوتی را که مربوط به کلمات خاصی هستند، شناسایی کنند. به این ترتیب، می‌توانند کلمه به کلمه اطلاعات را استخراج کرده و در نهایت جمله‌ی کامل را بسازند. اما نکته‌ی منفی این روش این است که برای پوشش طیف وسیعی از کلمات و تلفظ‌ها، باید تعداد زیادی فیلتر آموزش ببینند تا مدل بتواند تفاوت‌های ظریف بین کلمات را تشخیص دهد.

- شناسایی عمل انجام شده درون ویدیو:

شبکه‌های CNN برای انجام این وظیفه مناسب نیستند. زیرا تشخیص نوع حرکت در طول چندین فریم، یک مسئله‌ی نیازمند حافظه است که با روش‌هایی مانند RNN یا Transformer که دارای قابلیت به یاد سپاری در طول یک دنباله هستند قابلیت حل دارد. شبکه‌های CNN تنها می‌توانند نقش استخراج ویژگی از روی فریم‌های یک ویدیو

برای پردازش در مراحل بعد را بازی کنند. (شبکه‌های CNN سه‌بعدی در تئوری برای این وظیفه مناسب هستند و می‌توانند آن را حل کنند اما به دلیل سنگین بودن این نوع مدل‌ها، نیاز شدید به دیتای زیادی که خاص منظوره برچسب خورده‌اند و نیز هزینه‌ی زیاد در هنگام استنتاج به دلیل پردازش پنجره‌ی ثابتی از فریم‌های متوالی از منظر عملی مناسب و بهینه نبوده و نیز دارای دقت پایینی هستند)

- داوری انجام حرکت میل‌زنی در مسابقات زورخانه‌ای:

برای این وظیفه مناسب هستند. باتوجه به نوع داوری مسابقات میل‌زنی که براساس زاویه‌ی دست، بدن، میل و هم‌چنین برخورد میل با سر و ... انجام می‌شود شبکه‌های عصبی CNN می‌توانند مناسب باشند. برای انجام آن کافی است با روش‌های pose estimation نقاط مشخصه بدن (landmark) در لحظه‌ای خاص را استخراج کنیم و سپس از روی این نقاط یافتن زاوایای دست، آرنج، بازو و ... به‌راحتی قابل استخراج است.

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

در پاسخ این قسمت از فایل خانم میاهی‌نیا استفاده شده است

در واقع، لایه‌ی کانولوشن دو بعدی هم یک عملیات خطی هست، خیلی شبیه به لایه‌ی FC، ولی با اشتراک‌گذاری وزن‌ها و ساختار مکانی. ما می‌توانیم کانولوشن رو به شکل زیر بازنویسی کنیم:

$$Z = W * X + b$$

درحالی که W یک کرنل کانولوشنی، b بایاس و X ورودی می‌باشد. از طرفی فرمول batch normalization:

$$y = \gamma \times \frac{Z - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta$$

معادله تلفیق آن را می‌توان به صورت زیر نوشت:

$$W_{\text{fused}} = \gamma \times \frac{W}{\sqrt{\sigma^2 + \epsilon}}$$

$$b_{\text{fused}} = \gamma \times \frac{b - \mu}{\sqrt{\sigma^2 + \epsilon}} + \beta$$

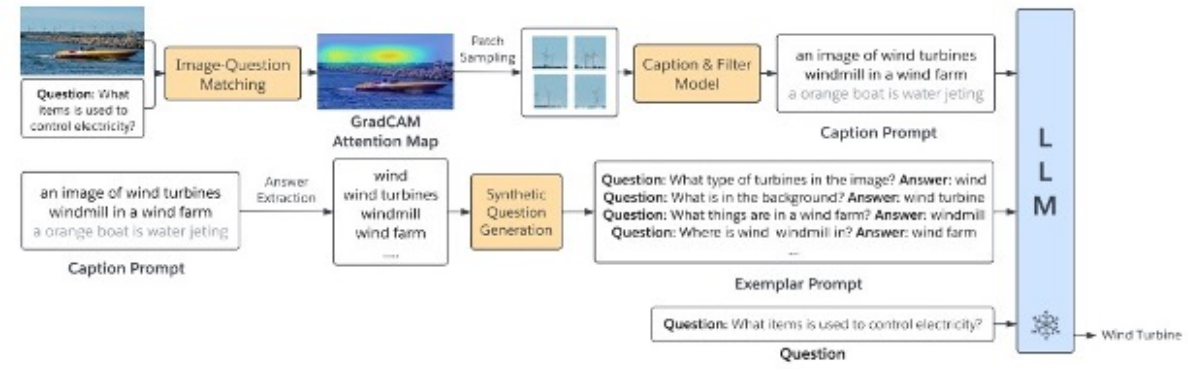
$$Y = W_{\text{fused}} \times X + b_{\text{fused}}$$

ترکیب کردن لایه‌ها مثل Conv2D + BatchNorm باعث می‌شود این عملیات‌ها به جای اجرا

به‌صورت جداگانه، به‌صورت یک عملیات واحد روی سخت افزار اجرا بشوند. یعنی، کاهش تعداد دفعات دسترسی به حافظه، کاهش زمان اجرای کل شبکه و مناسب برای اجرا روی GPU های محدود. این عملیات‌های ترکیب‌شده هیچ اثری روی دقت مدل ندارند زیرا عملیات‌های Conv و BatchNorm خطی هستند و می‌توانند به صورت ریاضیاتی ترکیب بشوند و فقط ساختار گراف تغییر می‌کند، نه منطق مدل یا وزن‌های آموزش دیده.

(د) یکی از کاربردهای مدل‌های چندوجهی[11] مانند ChatGPT وظیفه‌ی VQA[12] (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.

این سوال از روی ایده‌ی مقاله‌ی Img2LLM طرح شده است. همانطور که در شکل زیر که نمایی کلی از متد پیشنهادی مدل مقاله‌ی ذکر شده است، با استفاده از GradCAM بر روی attention map های مدل question-image matching می‌توانند به ناحیه‌ی حدودی از تصویر که سوال ورودی به آن مربوط است برسند. البته لازم به ذکر است که این مقاله برای بهبود عملکرد مدل نهایی ایده‌ی خوب دیگری داشته است که در اسکوپ این تمرین نیست و خواننده‌های علاقه‌مند را به مطالعه‌ی مقاله‌ی سایت شده ارجاع می‌دهیم.



[11] Multi-Modal

[12] Visual question answering

دلیل اینکه Attention map های مدل QIM می‌توانند به ناحیه‌ای از تصویر توجه کنند که سوال بیشتر مربوط به آن ناحیه است، یادگیری فضای بازنمایی مشترک میان متن و تصویر از نوع آموزش چندوجهی آن است. توجه: پاسخ این سوال یکتا نیست و پاسخ فوق تنها نمونه‌ای از آن‌ها و پیشنهاد ما می‌باشد. برای نمره‌دهی این سوال حتما به میزان منطقی و عملیاتی بودن پاسخ‌های دانشجویان اهمیت داده خواهد شد.

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)`

ابتدا تعداد پارامترهای هر لایه را محاسبه می‌کنیم. سپس به سراغ receptive field لایه‌ها می‌رویم.

فرمول محاسبه پارامترها:

لایه‌ی کانولوشن:

$$\text{num\_filters} \times (\text{kernel\_size} \times \text{kernel\_size} \times \text{num\_in\_channels} + 1)$$

لایه batch norm:

برای هر چنل، ۲ پارامتر یادگیری دارد پس

$$\text{num\_channel} \times 2$$

لایه‌ی FC:

$$\text{output\_size} \times (\text{input\_size} + 1)$$

تعداد پارامترها:

(آ) Conv1: $32 \times (1 + 3 \times 7 \times 7) = 4736$

(ب) bn1: $32 \times 2 = 64$

(ج) Conv2: $64 \times (1 + 32 \times 5 \times 5) = 51264$

(د) bn2: $64 \times 2 = 128$

(ه) Conv3: $128 \times (1 + 64 \times 3 \times 3) = 73856$

(و) bn3: $128 \times 2 = 256$

(ز) Conv4: $128 \times (1 + 128 \times 3 \times 3) = 147584$

(ح) bn4: $128 \times 2 = 256$

(ط) Conv5: $256 \times (1 + 128 \times 3 \times 3) = 295168$

(ی) bn5: $256 \times 2 = 512$

(ک) fc1: $1024 \times (43264 + 1) = 44303360$

(ل) fc2: $1024 \times (1024 + 1) = 1049600$

(م) fc3: $10 \times (1024 + 1) = 10250$

مجموعا حدودا ۹.۴۵ میلیون پارامتر

میدان‌دید لایه‌ها (منبع: Stanford CS231n):

میدان‌دید هر لایه از طریق فرمول زیر محسابه می‌شود:

$$R_i = R_{i-1} + (k_i - 1) \cdot d_i \cdot j_{i-1}$$

$$j_i = j_{i-1} \cdot s_i$$

که R_{i-1} میدان‌دید لایه قبلی، j_{i-1} به‌عنوان پرش از لایه قبلی به لایه فعلی، k_i اندازه کرنل و d_i نیز stride همان لایه است.

$$d_i = \text{dilataion}, R_0 = j_0 = 1$$

(آ) Conv1: $7 \times 7, k_1 = 7, d_1 = 1, j_1 = 1$

(ب) Conv2: $11 \times 11, k_2 = 5, d_2 = 1, j_2 = 2$

(ج) AvgPool1: $13 \times 13, k_3 = 2, d_3 = 1, j_3 = 4$

(د) Conv3: $29 \times 29, k_4 = 3, d_4 = 2, j_4 = 4$

(ه) Conv4: $37 \times 37, k_5 = 3, d_5 = 1, j_5 = 4$

(و) AvgPool2: $41 \times 41, k_6 = 2, d_6 = 1, j_6 = 8$

(ز) Conv5: $57 \times 57, k_7 = 3, d_7 = 1, j_7 = 8$

(ح) AvgPool3: $65 \times 65, k_8 = 2, d_8 = 1, j_8 = 16$

مقدار جمع و ضرب:

(آ) Conv1: $256 \times 256 \times 32 \times 7^2 \times 3 \times 2$

(ب) Bn1: $256 \times 256 \times 32 \times 2$

(ج) Conv2: $126 \times 126 \times 64 \times 5^2 \times 32 \times 2$

(د) Bn2: $126 \times 126 \times 64 \times 2$

(ه) AvgPool1: $63 \times 63 \times 64 \times 4$

(و) Conv3: $59 \times 59 \times 128 \times 5^2 \times 64 \times 2$

(ز) Bn3: $59 \times 59 \times 128 \times 2$

(ح) Conv4: $57 \times 57 \times 128 \times 3^2 \times 128 \times 2$

(ط) Bn4: $57 \times 57 \times 128 \times 2$

(ی) AvgPool2: $28 \times 28 \times 128 \times 4 \times 2$

(ک) Conv5: $26 \times 26 \times 256 \times 3^2 \times 128 \times 2$

(ل) Bn5: $26 \times 26 \times 256 \times 2$

(م) AvgPool3: $13 \times 13 \times 256 \times 4$

(ن) Fc1: $43264 \times 1024 \times 2$

(س) Fc2: $1024 \times 1024 \times 2$

(ع) Fc3: $1024 \times 10 \times 2$

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).

به نوتبوک Deep4032_HW2_TM_Ans.ipynb مراجعه کنید

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)

شکل ۱: نمونه ای از تصویر عدد ۴

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 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

به نوتبوک Deep4032_HW2_CNN_Ans.ipynb مراجعه کنید

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.

در پاسخ این قسمت از فایل خانم میاهی‌نیا استفاده شده است

به نوتبوک Resnext.ipynb مراجعه کنید

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).

در پاسخ این قسمت از فایل آقای جاوید استفاده شده است

به نوتبوک q7.ipynb مراجعه کنید

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).

به نوتبوک Deep4032_HW2_SS_Ans.ipynb مراجعه کنید