به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

درس شبکه‌های عصبی و یادگیری عمیق

تمرین امتیازی

پرسش ۱	نام دستیار طراح	توحید عبدی
	رایانامه	tohid.abdi@ut.ac.ir
پرسش ۲	نام دستیار طراح	محمدمهدی سلمانی
	رایانامه	m.salmani78@ut.ac.ir
	مهلت ارسال پاسخ	۱۴۰۴.۰۴.۰۱

فهرست

قوانین

قبل از پاسخ دادن به پرسش‌ها، موارد زیر را با دقت مطالعه نمایید:

- از پاسخ‌های خود یک گزارش در قالبی که در صفحه‌ی درس در سامانه‌ی Elearn با نام ***REPORTS_TEMPLATE.docx*** قرار داده شده تهیه نمایید.
- پیشنهاد می‌شود تمرین‌ها را در قالب گروه‌های دو نفره انجام دهید. (بیش از دو نفر مجاز نیست و تحویل تک نفره نیز نمره‌ی اضافی ندارد) توجه نمایید الزامی در یکسان ماندن اعضای گروه تا انتهای ترم وجود ندارد. (یعنی، می‌توانید تمرین اول را با شخص A و تمرین دوم را با شخص B و ... انجام دهید)
- **کیفیت گزارش شما در فرآیند تصحیح از اهمیت ویژه‌ای برخوردار است**؛ بنابراین، لطفا تمامی نکات و فرض‌هایی را که در پیاده‌سازی‌ها و محاسبات خود در نظر می‌گیرید در گزارش ذکر کنید.
- در گزارش خود مطابق با آنچه در قالب نمونه قرار داده شده، برای شکل‌ها زیرنویس و برای جدول‌ها بالانویس در نظر بگیرید.
- الزامی به ارائه توضیح جزئیات کد در گزارش نیست، اما باید نتایج بدست آمده از آن را گزارش و تحلیل کنید.
- **تحلیل نتایج الزامی می‌باشد، حتی اگر در صورت پرسش اشاره‌ای به آن نشده باشد.**
- **دستیاران آموزشی ملزم به اجرا کردن کدهای شما نیستند**؛ بنابراین، هرگونه نتیجه و یا تحلیلی که در صورت پرسش از شما خواسته شده را به طور واضح و کامل در گزارش بیاورید. در صورت عدم رعایت این مورد، بدیهی است که از نمره تمرین کسر می‌شود.
- **کدها حتما باید در قالب نوت‌بوک با پسوند .ipynb تهیه شوند، در پایان کار، تمامی کد اجرا شود و خروجی هر سلول حتما در این فایل ارسالی شما ذخیره شده باشد.** بنابراین برای مثال اگر خروجی سلولی یک نمودار است که در گزارش آورده‌اید، این نمودار باید هم در گزارش هم در نوت‌بوک کدها وجود داشته باشد.
- **در صورت مشاهده‌ی تقلب امتیاز تمامی افراد شرکت‌کننده در آن، 100- لحاظ می‌شود.**
- تنها زبان برنامه نویسی مجاز **Python** است.
- **استفاده از کدهای آماده برای تمرین‌ها به هیچ وجه مجاز نیست. در صورتی که دو گروه از یک منبع مشترک استفاده کنند و کدهای مشابه تحویل دهند، تقلب محسوب می‌شود.**
- نحوه محاسبه تاخیر به این شکل است: پس از پایان رسیدن مهلت ارسال گزارش، حداکثر تا یک هفته امکان ارسال با تاخیر وجود دارد، پس از این یک هفته نمره آن تکلیف برای شما صفر خواهد شد.

- سه روز اول: بدون جریمه
- روز چهارم: ۵ درصد
- روز پنجم: ۱۰ درصد
- روز ششم: ۱۵ درصد
- روز هفتم: ۲۰ درصد

- حداکثر نمره‌ای که برای هر سوال می‌توان اخد کرد ۱۰۰ بوده و اگر مجموع بارم یک سوال بیشتر از ۱۰۰ باشد، در صورت اخد نمره بیشتر از ۱۰۰، اعمال نخواهد شد.
  - برای مثال: اگر نمره اخذ شده از سوال ۱ برابر ۱۰۵ و نمره سوال ۲ برابر ۹۵ باشد، نمره نهایی تمرین ۹۷.۵ خواهد بود و نه ۱۰۰.
- لطفا گزارش، کدها و سایر ضمایم را به در یک پوشه با نام زیر قرار داده و آن را فشرده سازید، سپس در سامانه‌ی Elearn بارگذاری نمایید:

HW[Number] _[Lastname]_[StudentNumber]_[Lastname]_[StudentNumber].zip

(مثال: HW1_Ahmadi_810199101_Bagheri_810199102.zip)

- برای گروه‌های دو نفره، بارگذاری تمرین از جانب یکی از اعضا کافی است ولی پیشنهاد می‌شود هر دو نفر بارگذاری نمایند.

پرسش ۱. تحلیل عملکرد شبکه‌های عصبی تحت حملات متخاصم

این تمرین برای تلفیق مفاهیم تئوری و عملی یادگیری عمیق طراحی شده و شامل بخش‌های تحلیل تاثیر نویز، انتقال یادگیری، حملات متخاصم و همچنین ارزیابی دیداری مدل‌ها با ابزارهای توضیح‌پذیری است. بخش تئوری نیز برای درک عمیق‌تر مفاهیم طراحی شده است.

نکته مهم: در مواردی که مقادیر هایپرپارامترها یا تنظیمات دقیق ذکر نشده‌اند، انتخاب آن‌ها بر عهده خود دانشجو است. در بخش‌هایی که مسئله نیاز به تفسیر دارد، استفاده از خلاقیت و ابتکار شخصی، امتیاز بیشتری نسبت به پرسیدن سوالات مستقیم خواهد داشت.

۱-۱. آموزش مدل ResNet روی تصاویر نویزی (۲۵ نمره)

در این بخش باید یک مدل ResNet (مانند ResNet18 یا ResNet34) را بدون استفاده از وزن‌های پیش‌آموزش‌دیده (pretrained=False) روی دیتاست CIFAR-100 آموزش دهید. سپس بررسی کنید که نویز گوسی چه تاثیری روی عملکرد مدل دارد.

مراحل:

۱. دیتاست CIFAR-100 را بارگذاری کنید.

۲. نویز گوسی با میانگین ۰ و واریانس ۰.۰۵ را به داده‌های آموزشی اضافه کنید.

۳. مدلResNet را با داده‌های نویزی و سپس بدون نویز، هر دو برای ۲۰ ایپاک آموزش دهید.

٤. نمودارهای دقت و خطای اعتبارسنجی را برای هر دو حالت رسم کنید.

٥. عملکرد مدل‌ها در دو حالت را مقایسه کرده و اثر نویز را تحلیل کنید.

۱-۲. انتقال یادگیری با ViT روی Flowers-102 (۲۵ نمره)

در این بخش با استفاده از یک مدل پیش‌آموزش‌دیده Vision Transformer (ViT) ، عملیات انتقال یادگیری روی دیتاست Flowers-102 انجام می‌شود.

مراحل:

۱. مدل vit_base_patch16_224 را با وزن‌های پیش‌آموزش‌دیده بارگذاری کنید و لایه‌ی خروجی مدل را برای دسته‌بندی ۱۰۲ کلاس گل تغییر دهید.

۲. مدل را ۵ ایپاک روی Flowers-102 تنظیم مجدد (fine-tune) کنید.

۳. اکنون همین مدل را بدون وزن‌های پیش‌آموزش دیده بارگذاری کنید و ۱۰ ایپاک آموزش دهید.

٤. نمودارهای دقت و خطای اعتبارسنجی را برای هر دو حالت رسم کنید.

٥. با بررسی جزئیات دیتاست flowers-102 و دیتاستی که مدل با وزن‌های پیش‌آموزش‌دیده بر روی آن fine-tune شده است، عملکرد دو مدل را تحلیل کنید.

۱-۳. حملات متخاصم و دفاع (۳۰ نمره)

در این بخش مدل‌های آموزش‌دیده از بخش‌های قبلی را در برابر حملات متخاصم FGSM و PGD بررسی و مقاوم‌سازی می‌کنید.

مراحل:

۱. حمله FGSM با پارامتر $\epsilon=0.1$ و حمله PGD با ۷ تکرار و گام $\alpha=0.02$ پیاده‌سازی کنید.

۲. تصاویر متخاصم را برای ۴ مدل بخش‌های قبلی تولید کرده و عملکرد آن‌ها را روی این تصاویر ارزیابی نمایید. در صورت نگرفتن نتیجه مطلوب، پارامترهای حملات را تغییر دهید.

۳. مدل‌ها را مجددا با استفاده از adversarial training برای ۱۰ ایپاک آموزش دهید.

۴. دقت مدل‌ها را در سه حالت بدون حمله، تحت حمله و پس از دفاع مقایسه کنید.

۱-۴. سوالات تئوری (۲۰ نمره)

در این بخش باید به سؤالات مفهومی به صورت تشریحی و دقیق پاسخ دهید. استفاده از فرمول و نمودارهای توضیحی توصیه می‌شود.

پرسش‌ها:

۱. با وجود اینکه مدل ResNet دارای میلیون‌ها پارامتر است، چرا در بسیاری از موارد با داده‌های نویزی همچنان عملکرد مناسبی دارد؟

۲. مدل‌های Vision Transformer (ViT) در صورت استفاده از وزن‌های پیش‌آموزش‌دیده بهتر از حالت بدون pretrain عمل می‌کنند. این اختلاف را با استفاده از مفاهیم مینیمم تیز (sharp) و تخت (flat) تحلیل کنید.

۳. adversarial training را به عنوان شکلی از data augmentation تحلیل کنید.

۴. با تحلیل ساختار ResNet و ViT و استفاده از مفاهیم dropout و ensemble ، تفاوت این دو مدل در برابر حملات متخاصم را توجیه کنید.

۱-۵. بخش اختیاری (۵ نمره)

برای مدل‌های ResNet و ViT آموزش‌دیده در بخش‌های ۱-۱ تا ۱-۳، نمونه‌هایی از کلاس‌های مختلف انتخاب کرده و به کمک Grad-CAM نشان دهید مدل در کدام نواحی تصویر تمرکز کرده است.

.

پرسش ۲ – تولید توضیحات متنی برای تصاویر

تولید توضیحات متنی برای تصاویر (Image Captioning) یکی از مسائل مهم در تقاطع حوزه‌های بینایی ماشین و پردازش زبان طبیعی محسوب می‌شود. این فرایند معمولاً شامل استخراج ویژگی‌های بصری از تصویر و تبدیل آن‌ها به جملات توصیفی به زبان طبیعی می‌باشد. برای این منظور، مدل‌های مختلفی به کار گرفته می‌شوند؛ از ترکیب شبکه‌های کانولوشنی (CNN) و مدل‌های زبانی کلاسیک مانند RNN گرفته تا معماری‌های پیشرفته‌تر مبتنی بر ترنسفورمرها و بینایی ماشینی مانند ViT. این تکنولوژی در کاربردهایی مانند دسترسی به محتوای تصویری برای افراد نابینا، جستجوی تصاویر و تجزیه و تحلیل محتوا در شبکه‌های اجتماعی مفید است.

۲-۱. آماده سازی داده (۱۵ نمره)

برای این تمرین، از نسخه ترجمه‌شده مجموعه داده COCO Captions استفاده می‌کنیم که شامل ۴۰,۰۰۰ جفت تصویر-کپشن است. (لزومی ندارد از کل داده برای تمرین استفاده شود)

مراحل:

۱. دریافت داده‌ها:

- فایل داده‌ی coco-flickr-fa-40k.zip را از این لینک دانلود نمایید.

۲. پیش‌پردازش تصاویر:

- تصاویر را به ابعاد یکسان تغییر اندازه دهید و مقادیر پیکسل‌ها را به محدوده [0,1] نرمال‌سازی کنید.

۳. پیش‌پردازش کپشن‌ها:

- با استفاده از کتابخانه hazm، داده‌های متنی فارسی را نرمال‌سازی کنید.
- همچنین می‌توانید علائم نگارشی، نمادهای خاص و اعداد غیرضروری را نیز حذف کنید.

۴. ایجاد واژگان (Vocabulary):

- کپشن‌ها را به توکن‌های مجزا بشکنید و یک دیکشنری از کلمات بسازید. به هر کلمه یک عدد صحیح یکتا اختصاص دهید.
- از توکن‌های ویژه مانند <pad>، <sos>، <eos> و <unk> به شکل مناسب استفاده کنید.

۵. آماده‌سازی برای آموزش:

- داده‌ها را به نسبت مناسب به دسته‌های train، validation و test تقسیم کنید.

- طول کپشن‌ها را با استفاده از توکن <pad> یکسان کنید. (توجه کنید که برای این توکن در مرحله آموزش، خطا یا گرادیان نباید محاسبه شود.)

۶. بررسی داده‌ها:

- چند نمونه تصویر به همراه کپشن‌های آن‌ها را به شکل مناسب نمایش دهید.
- تحلیل آماری شامل هیستوگرام طول کپشن‌ها، ۱۰ کلمه پرتکرار و تعداد کلمات یکتا در واژگان را گزارش کنید.

۲-۲. پیاده‌سازی معماری CNN+RNN با مکانیزم توجه (۲۵ نمره)

در این بخش، هدف پیاده‌سازی یک مدل Image Captioning با استفاده از ترکیب شبکه کانولوشنی (CNN) برای استخراج ویژگی‌های بصری و شبکه بازگشتی (RNN) با مکانیزم توجه برای تولید کپشن است.[۱]

۱. **پیاده سازی Encoder:**

- یک مدل CNN از پیش‌آموزش‌دیده مانند EfficientNet-B7 را انتخاب کنید.
- در این بخش به ماتریسی از بردارها نیاز داریم بطوریکه هر بردار نشان‌دهنده ویژگی یک منطقه از تصویر باشد. برای این منظور لایه‌های طبقه‌بندی و pooling آخر را حذف کنید تا ماتریس ویژگی‌ها استخراج شود.

۲. **پیاده سازی مکانیزم توجه:**

- بررسی کنید چه مکانیزم‌های توجهی در مقاله "Show, Attend and Tell" پیاده‌سازی شده است.
- مکانیزم Soft Attention را با روش Additive Attention پیاده‌سازی کنید:
  - حالت مخفی (hidden state) فعلی LSTM را با ویژگی‌های تصویر مقایسه کنید.
  - وزن‌های توجه (attention weights) را محاسبه کنید.
  - بردار زمینه (context vector) را به‌عنوان میانگین وزن‌دار ویژگی‌های تصویر تولید کنید.

۳. **پیاده سازی Decoder:**

- از یک لایه Embedding، برای تبدیل کلمات به بردارهای متراکم استفاده کنید.

اختیاری: میتوانید از Embeddingهای از پیش‌آموزش‌دیده مانند fastText برای فارسی استفاده کنید.

[۱] **مقاله مرجع:** Show, Attend and Tell: Neural Image Caption Generation with Visual Attention

- از یک LSTM تک‌لایه (یا دو لایه) برای تولید توالی کلمات استفاده کنید.
- بردار زمینه تولیدشده توسط مکانیزم توجه را در هر گام زمانی به LSTM وارد کنید.
- از یک لایه Linear و تابع SoftMax برای پیش‌بینی کلمه بعدی استفاده کنید.
- تکنیک‌های regularization مانند dropout در LSTM و weight decay در بهینه‌ساز را اعمال کنید.
- تعداد کل پارامترهای مدل و تعداد پارامترهای قابل آموزش را گزارش کنید.

۲-۳. آموزش و ارزیابی (۲۵ نمره)

۱. **آموزش مدل:**

- از تابع خطا و بهینه‌ساز مناسب استفاده کنید و هایپرپارامترهای خود را گزارش نمایید.
- از روش teacher forcing برای آموزش استفاده کنید.
- در پایان هر دوره یک نمونه تصویر و کپشن تولیدشده را نمایش دهید.
- نمودار خطای آموزش و اعتبارسنجی را گزارش نمایید.

۲. **ارزیابی مدل:**

- چند عکس تصادفی را از مجموعه داده تست انتخاب، و روی مدل اجرا کنید و کپشن‌ها را با دو روش Greedy Search و Beam Search (عرض پرتو ۳) تولید نمایید.
- مقاله از چه معیارهایی برای ارزیابی استفاده کرده است؟
- خروجی‌ها را با معیارهای BLEU-1 و BLEU-4 ارزیابی کنید.
- برای هریک از این تصاویر، کپشن واقعی (متن مرجع) و کپشن‌های تولیدشده را نمایش دهید.

۲-۴. تحلیل و بهبود مدل (۳۵ نمره)

۱. **تحلیل خطاهای مدل:**

- ۵ نمونه تصویر انتخاب کنید که کپشن‌های تولیدشده کیفیت پایینی دارند.
- وزن‌های توجه را به‌صورت نقشه حرارتی (heatmap) روی تصویر اصلی رسم کنید. برای هر کلمه تولیدی، یک نقشه حرارتی نشان دهید که مناطق مورد تمرکز مدل را مشخص کند.
- تحلیل کنید چرا مدل در این موارد اشتباه کرده است. مثلاً، آیا به مناطق نادرست تصویر توجه کرده؟ آیا کلمات نامناسب انتخاب شده‌اند؟ و یا اینکه مشکل به دیتاست یا معماری ربط دارد؟

۲. پیاده‌سازی Scheduled Sampling[1]:

در Teacher Forcing، مدل همیشه از کپشن‌های مرجع برای آموزش استفاده می‌کند. در Scheduled Sampling، با احتمال مشخصی از کلمات تولیدشده توسط مدل (به جای مرجع) استفاده می‌شود.

- یک مکانیزم احتمالاتی پیاده‌سازی کنید که در هر گام آموزشی، با احتمال p از کلمه مرجع و با احتمال $1 - p$ از کلمه تولیدشده توسط مدل استفاده کند.
- احتمال p را به‌صورت خطی یا نمایی کاهش دهید (مثلاً از ۱ به ۰.۵ در طول آموزش).
- این مکانیزم را به کد آموزش مدل بخش سوم اضافه کنید.
- آموزش را تکرار کرده و نمودار خطا را با حالت قبل مقایسه کنید.

۳. آزمایش با مکانیزم‌های مختلف توجه:

- در این بخش، هدف پیاده‌سازی مکانیزم توجه Scaled Dot-Product Attention بجای روش استفاده‌شده در بخش ۲-۲ است. این مکانیزم توجه، که در معماری Transformer معرفی شده، از ضرب داخلی نرمال‌شده بین بردارهای Query و Key برای محاسبه وزن‌های توجه استفاده می‌کند:
  - حالت مخفی فعلی LSTM را به‌عنوان Query و ویژگی‌های تصویر را به‌عنوان Key و Value در نظر بگیرید.
  - ضرب داخلی بین Query و Key را محاسبه کرده و با فاکتور مقیاس $\sqrt{d_k}$ (بعد بردار ویژگی) نرمال کنید.
  - از SoftMax برای محاسبه وزن‌های توجه و سپس محاسبه بردار زمینه به‌صورت میانگین وزن‌دار استفاده کنید.
- بخش ۲-۳ را برای این حالت مجدد تکرار کرده و نتایج را با مدل اولیه مقایسه کنید.

[1] **مقاله مرجع:** Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks