دانشگاه صنعتی امیرکبیر
دانشکده مهندسی کامپیوتر
پروژه درس رایانش عصبی و یادگیری عمیق

پروژه هشتم

هدف: آشنایی با ساختارهای کدگذار/کدگشا[1] و یادگیری انتقالی.

کد: پیاده سازی این پروژه را به زبان پایتون انجام دهید؛ در این فعالیت مجاز به استفاده از tensorflow یا pytourch می‌باشید.

گزارش: ملاک اصلی انجام فعالیت، گزارش آن است و ارسال کد بدون گزارش فاقد ارزش است. برای این فعالیت یک فایل گزارش در قالب pdf تهیه کنید و در آن برای هر وال، تصاویر ورودی، تصاویر خروجی و توضیحات کامل و جامعی تهیه کنید.

تذکر: مطابق قوانین دانشگاه هر نوع کپی برداری و اشتراک کار دانشجویان غیر مجاز بوده و شدیدا برخورد خواهد شد. استفاده از کدها و توضیحات اینترنت به منظور یادگیری بلامانع است، اما کپی کردن غیرمجاز است.

راهنمایی: در صورت نیاز میتوانید سوالات خود را در خصوص پروژه از تدریسیارهای درس، از طریق ایمیل زیر یا در گروه تلگرامی بپرسید.(لینک گروه تلگرامی در سایت کورسز در دسترس بوده و قبلا به همه‌ی دانشجویان ایمیل شده است)

Email: ann.ceit.aut@gmail.com

توجه: برای آموزش شبکه های عمیق می‌توانید از منابع و بسترهای سخت افزاری برخط رایگان نظیر Google Colab یا Kaggle استفاده نمایید.

تاخیر مجاز: در طول ترم، مجموعا مجاز به حداکثر ده روز تاخیر برای ارسال تمرینات هستید(بدون کسر نمره). این تاخیر را می‌توانید بر حسب نیاز بین تمرینات مختلف تقسیم کنید؛ اما مجموع تاخیرات تمام تمرینات شما نباید بیشتر از ده روز شود. پس از استفاده از این تاخیر مجاز، هر روز تاخیر باعث کسر ۱۰٪ نمره‌ی آن تمرین خواهد شد.

ارسال: فایل های کد و گزارش خود را در قالب یک فایل فشرده با فرمت StudentID_HW08.zip تا تاریخ ۱۴۰۲/۰۴/۱۹ ارسال نمایید.

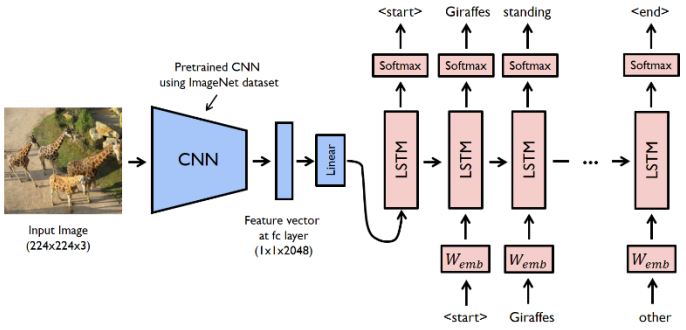
در این پروژه با عملکرد شبکه های کدگذار-کدگشا در کاربرد تولید توضیح (متن) برای تصاویر[2] آشنا خواهید شد. توصیف عکس توسط مدل های زبانی میتواند در مقاصد گوناگونی نظیر ایجاد دستیار برای نابینایان کاربرد عملیاتی و حیاتی داشته باشد. در این پروژه هدف این است که شما با آموزش یک شبکه‌ی عمیق، برای تصویر ورودی یک توضیح به زبان انگلیسی تولید نمایید که بیان کننده‌ی ویژگی‌ها و جزئیات صحنه‌ی عکس باشد. بدین منظور مجموعه داده‌ی Flickr8K [3] در اختیار شما قرار گرفته است که مشتمل بر هشت هزار تصویر ورودی بوده و بطور میانگین برای هر عکس پنج توصیف انگلیسی با طول‌های مختلف در نظر گرفته شده است. نمونه‌ای از تصاویر مجموعه داده در شکل (۱) قابل

شکل ۱ - dog is running through the grass

[1] Encoder-Decoder
[2] Image Captioning
[3] Images: http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description/Flickr8k_Dataset.zip
Text Data: http://nlp.cs.illinois.edu/HockenmaierGroup/Framing_Image_Description/Flickr8k_text.zip

مشاهده است. تقسیم مجموعه داده به دسته های آموزش، آزمون و اعتبار سنج را با نسبت‌های ۷۵:۱۵:۱۰ انجام دهید.

شما در کلاس درس با ساختار کلی تولید متن برای تصاویر آشنا شدید که اساسا ورودی آن یک بردار ویژگی و خروجی آن یک توالی[۴] می‌باشد و در دسته‌ی مدل های یک-به-چند[۵] قرار می‌گیرد. شماتیک جزئی‌تری از این ساختار را در شکل (۲) ملاحظه می‌کنید. در ابتدا از تصویر ورودی بایستی بردار ویژگی استخراج شود و اطلاعات عکس و صحنه کدگذاری شود(Encoding)؛ این بخش می‌تواند توسط هر شبکه‌ی پیچشی عمیق[۶] مرسوم انجام گیرد. حال بردار ویژگی تولید شده را میتوان به کمک یک مدل زبانی[۷] مانند Transformer یا GPT و یا معماری های بازگشتی نظیر LSTM یا GRU دریافت و بردار ویژگی متناظر با متن را تولید نمود و کلمات را بدست آورد تا جمله بصورت کامل حاصل شود(Decoding).



شکل ۲ – شماتیکی از مدل کدگذار-کدگشا برای تولید متن برای تصاویر

در شبکه‌های این چنینی پیش‌پردازش‌های توامی برای قسمت متن و تصویر مورد نیاز است. به عنوان مثال، برای قسمت کدگذاری لازم است ابعاد تصاویر یکسان شود(اگر نباشد) یا برای قسمت کدگشا نیاز است که کلمات همگی به حروف کوچک[۸] تبدیل شوند، علامت های غیر ضروری حذف شود، نشانه‌ها[۹] استخراج شوند، تعبیه[۱۰] عددی و بردار بازنمایی برای آنان در نظر گرفته شود، نشانه‌ی شروع و پایان اضافه شود و... .

۱. مقصود از تعبیه‌ی کلمه[۱۱] چیست و چگونه انجام می‌شود؟ روش مناسبی که خودتان انتخاب و در پیاده سازی استفاده خواهید نمود را کامل توضیح دهید. چگونه میتوان یک متن که حاوی تعدادی کلمه است را تعبیه و آن متن را بازنمایی کرد و مجدد از بازنمایی به کلمات رسید؟ آیا خطایی در بازنمایی وجود دارد؟ چه رویکردی برای این امر اتخاذ می‌شود؟ در این پروژه، متن‌های هر تصویر میتواند طول متفاوتی داشته باشد و این در حالی است که همه‌ی تنظیمات شبکه و ابعاد آن بایستی ثابت باشد؛ چگونه به این چالش پاسخ می‌دهید؟ پاسخ خود را توجیه کنید. (۱۱ امتیاز)

۲. آیا مسئله‌ی تولید متن برای تصویر یک مسئله‌ی دسته‌بندی[۱۲] است یا رگرسیون[۱۳]؟ چرا؟ تابع هزینه چگونه کار می‌کند؟ توضیح دهید. (۴ امتیاز)

۳. طبق توضیحاتی که در قسمت ابتدایی پروژه داده شد، ملاحظه نمودید که برای هر تصویر چندین توضیح متنی در مجموعه داده جمع‌آوری شده است. چگونه می‌توان از همه‌ی متن‌های توضیح یک تصویر در حین آموزش استفاده نمود؟ (۵ امتیاز)

---

[4] Sequence
[5] One-To-Many
[6] Convolution Neural Networks (CNNs)
[7] Language Model
[8] Lower-Case
[9] Tokenize
[10] Embedding
[11] Word Embedding
[12] Classification
[13] Regression

آموزش کامل قسمت کدگذار معمولا زمان‌بر و چالشی بوده و نیازمند صرف منابع و تنظیم شبکه می‌باشد. برای اجتناب از این امر معمولا از یادگیری انتقالی[۱۴] استفاده می‌شود. هدف کلی یادگیری انتقالی این است که شبکه‌ای قبلا یک بار آموزش دیده را در جاهای مختلف از آن استفاده کرد؛ رویکرد های متعددی می‌توان برای یادگیری انتقالی اشاره نمود که یکی از آنان مبتنی بر انتقال قسمت استخراج ویژگی شبکه می باشد؛ بدین صورت که شبکه‌ی مفروض که دارای پیچیدگی و قدرت بالایی بوده و معمولا با هدف دسته‌بندی طراحی شده است، با داده‌های خیلی زیادی بطور دقیق آموزش داده می شود. پس از آموزش، سر پیش‌بینی[۱۵] از شبکه حذف و وزن لایه‌های استخراج ویژگی قفل[۱۶] می‌شود؛ حال در هر فعالیت یا شبکه‌ی دیگری که نیاز به استخراج ویژگی باشد، آن قسمت از شبکه از پیش آموزش دیده شده[۱۷] آورده و استفاده می‌شود. البته گاه بنابر افزایش سازگاری قسمت استخراج ویژگی با سایر اجزای شبکه‌ی جدید، مکانیزمی اندیشه می‌شود که وزن‌ها مجددا تغییر و تنظیم دقیق[۱۸] شوند.

۴. گام‌های انتقال انتقال یادگیری را برای یک شبکه‌ی عمیق پیچشی بیان کنید و مکانیزم فوق را بصورت دقیق بررسی کنید. (۸ امتیاز)

۵. پیش‌پردازش های مورد نیاز برای تصاویر و متن را انجام دهید و سپس یک شبکه‌ی عمیق پیچشی از پیش آموزش دیده شده را به دلخواه انتخاب و فرآیند انتقال قسمت استخراج ویژگی آن را انجام دهید تا قسمت Encoder شبکه‌تان حاصل شود. (بر جسب منابع پردازشی تان، پیشنهادی می‌شود از یکی از شبکه‌های ResNet18، ResNet50، InceptionV3 یا Xception استفاده نمایید. همچنین مجاز به تغییر ابعاد تصاویر مجموعه داده هستید) (۳۰ امتیاز)

۶. در این بخش شما بایستی قسمت Decoder شبکه‌تان را به کمک قسمت Encoder آموزش دهید. برای طراحی Decoder می‌توانید از یک شبکه عمیق بازگشتی دلخواه همانند شکل (۲) استفاده نمایید؛ همچنین می‌توانید با بهره از یکی از مدل‌هایی که در آخرین مباحث درس با آن آشنا شده اید، این مهم را انجام دهید(برای مثال قسمت Decoder شبکه‌ی Transformers). ضمن توضیح فرآیند آموزش، ابعاد ورودی و خروجی این شبکه چگونه و با استناد به چه عوامل یا شرایطی تعیین می‌شود؟ نمودار خطا را به ازای دسته‌های آموزش و اعتبار سنجی رسم نمایید. معماری Decoder طراحی شده‌تان را در گزارش خود توضیح دهید. (۳۵ امتیاز)

۷. از مجموعه‌ی داده‌ی آزمون ۲۵ تصویر بصورت تصادفی انتخاب و برای آنان متن تولید نمایید و در گزارش خود بیاورید. (۷ امتیاز)

موفق باشید

---

[14] Transfer Learning
[15] Prediction head
[16] Lock/Freeze
[17] Pre-Trained
[18] Fine-Tune