به نام خدا

درس مبانی یادگیری عمیق

پاسخنامه تمرین سری ششم

استاد درس : دکتر مرضیه داوودآبادی

دستیاران : مهسا موفق بهروزی، الناز رضایی،
مرتضی حاجی آبادی

دانشگاه علم و صنعت ایران، دانشکده مهندسی کامپیوتر

نیمسال اول تحصیلی ۱۴۰۲ - ۱۴۰۳

پاسخ های مناسب برای هر سوال، لزوما یکتا نیستند

۱. به نظر شما به ازای ورودی های زیر، هر کدام از شبکه های مبتنی بر لایه های هم‌گشتی[1] و مبتنی بر توجه[2] جهت مسئله ای که زیر آن نوشته شده است، چگونه عمل می‌کنند؟توضیح مختصری برای هر مورد ارائه دهید(۱۵ نمره).

(آ) مسئله گربه بودن یا نبودن

(ب) مسئله انسان بودن یا نبودن

شکل ۱: مقایسه شبکه های مبتنی بر لایه‌های هم گشتی و مبتنی بر توجه

(آ) هر دو مدل به درستی عکس(تصویر چپ) را به عنوان یک گربه طبقه بندی می کنند، اما مدل CNN نمی تواند طرح(سمت راست) را به عنوان گربه طبقه بندی کند، در حالی که مدل مبتنی بر توجه موفق است. زیرا CNN ها به بافت حساس تر هستند تا شکل، در حالی که $Transformer$ ها می توانند از هر دو استفاده کنند. CNN ها از فیلترهای کانولوشنال برای استخراج ویژگی های محلی از تصاویر استفاده می کنند، در حالی که $Transformer$ ها از

[1] Convolutional

[2] Attention

مکانیسم های توجه برای گرفتن وابستگی های جهانی و محلی استفاده می کنند. CNN ها برای طبقه بندی بیشتر به بافت متکی هستند تا شکل، در حالی که $Transformer$ ها می توانند از اطلاعات بافت و شکل استفاده کنند. بنابراین، CNN ها ممکن است در طبقه بندی طرح به عنوان گربه مشکل داشته باشند، در حالی که $Transformer$ ممکن است قادر به انجام این کار باشد.

(ب) توسط CNN هر دوی آنها به‌عنوان چهره طبقه‌بندی می‌شوند، زیرا یک CNN فقط بررسی می‌کند که آیا ویژگی‌های خاصی در تصویر ورودی وجود دارد یا خیر و به ترتیب آنها نسبت به یکدیگر اهمیتی نمی‌دهد.

شبکه‌های مبتنی بر توجه، که از وابستگی‌های جهانی و اطلاعات زمینه ای استفاده می‌کنند، احتمالاً تصویر تحریف‌شده را به‌عنوان چهره‌ای غیرانسانی تشخیص می‌دهند. با استفاده از مکانیسم توجه، آنها می توانند ناهماهنگی ها و ناهنجاری های موجود در تصویر را شناسایی کنند، مانند ویژگی های نامناسب و مخدوش صورت. با توجه به این بی‌نظمی‌ها، شبکه می‌تواند اهمیت کمتری برای مناطق تحریف شده قائل شود و طبقه‌بندی دقیق‌تری انجام دهد و به درستی تشخیص دهد که ورودی چهره انسان را نشان نمی‌دهد.

۲. الف) مفاهیم FP، TF،TP و FN را توضیح دهید(۱۰ نمره).

TP مخفف $True\ Positive$ است و تعداد نمونه های مثبتی که به درستی دسته بندی شده اند را نشان می دهد. TN مخفف $True\ Negative$ است و تعداد نمونه های منفی که به درستی دسته بندی شده اند را نشان می دهد. FP یا $False\ Positive$ تعداد نمونه های منفی را که به اشتباه دسته بندی شده اند نشان می دهد FN که معادل $False\ Negative$ است تعداد نمونه های منفی را که به اشتباه دسته بندی شده اند بیان می کند.

ب) فرض کنید پروژه ای برای تشخیص مجرمان هک اسنپ فود به شما داده شده است. با توجه به اهمیت اشتباه تشخیص ندادن افراد بی‌گناه به عنوان مجرم و همچنین امنیت مردم، چه معیار ارزیابی را پیشنهاد می‌دهید؟ توضیح دهید(۱۰ نمره).

برای اینکه مدل خوب باشد باید مطمئن باشیم که فردی که دستگیر میکنیم مجرم است ($Precision$) و همچنین میخواهیم تا حد امکان مجرمان را دستگیر کنیم($Recall$). $F-score$ این مبادله را مدیریت می‌کند.

۳. الف) به نظر شما تخمین چرخش چگونه میتواند برای وظیفه‌ی طبقه‌بندی مفید باشد(۱۰ نمره)؟

از طریق وظیفه پیش بینی چرخش ها، شبکه مجبور می شود یک $representation$ کلی از تصویر ورودی را بیاموزد. همچنین با تمرکز بر چرخش، مدل ممکن است ویژگی هایی را بیاموزد که نسبت به جهت اشیاء ثابت هستند، که می تواند برای وظیفه ی طبقه بندی مفید باشند. لینک مفید

ب) توضیح دهید که $one - hot\ vectors$ چیست و مشکل استفاده از آن‌ها چیست(۱۰ نمره)؟

بردارهای $one\ hot$ بردارهای باینری هستند که برای نمایش داده های $categorical$ استفاده می‌شوند، که در آن همه عناصر ۰ هستند، به جز یکی که ۱ است. موقعیت ۱ در بردار، مربوط به $category$ خاصی است که توسط آن بردار نشان داده شده است. بردارهای $one\ hot$ همه دسته ها را با فاصله مساوی از یکدیگر در نظر می‌گیرند و نمی توانند هیچگونه رابطه معنایی بین دسته ها را نشان دهند؛ همچنین منجر به ناکارآمدی در استفاده از حافظه و منابع محاسباتی می‌شوند.

ج) توضیح دهید که $Word2Vec$ چگونه با تعریف الگوریتم‌های $self - supervised$ مطابقت دارد(۱۰ نمره).

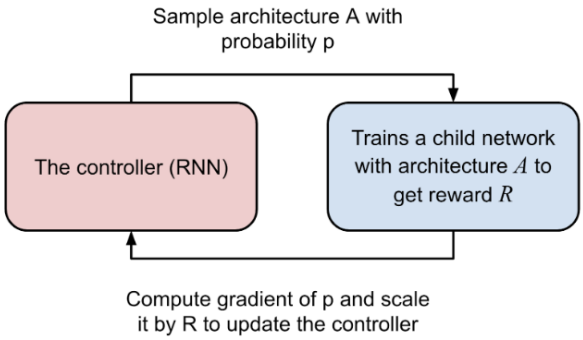
$Word2Vec$ برای آموزش نیاز به برچسب‌های انسانی ندارد و با یادگیری پیش بینی یک کلمه هدف ($target$) از $context$ آن یا پیش بینی کلمات اطراف یک کلمه $target$ روابط معنایی و نحوی پیچیده ای را بین کلمات کشف می کند. این رویکرد خود نظارتی به $Word2Vec$ اجازه می‌دهد تا بازنمایی‌های معنی دار کلمات را ($word\ embeddings$) یاد بگیرد که متن و معنی یک کلمه را به تصویر می‌کشد. سپس می‌توان از این $word\ embedding$ ها به عنوان ویژگی‌های ورودی برای انواع وظایف پردازش زبان طبیعی استفاده کرد.

۴. الف) در جستجوی ساختار شبکه یا ابر پارامتر های شبکه به صورت خودکار از رویکردهای مختلفی استفاده می‌شود. یکی از این رویکرد ها یادگیری تقویتی است. نحوه عملکرد آن را به اختصار توضیح دهید(۱۰ نمره). لینک کمکی

یادگیری تقویتی: در یادگیری تقویتی، یک عامل با ارتباط مستقیم با محیط عمل می‌کند.

در جستجوی ساختار شبکه یا ابر پارامترها، عامل با انجام عمل‌های مختلف در محیط و دریافت پاداش بر اساس عملکرد، به طور تدریجی یاد می‌گیرد که کدام ساختار یا ابر پارامترها بهترین نتایج را تولید می‌کنند. عامل با استفاده از الگوریتم‌هایی مانند $REINFORCE$، سعی می‌کند بهبود یابد و به ساختار یا ابر پارامترهای بهتری هدایت می‌شود کع تابع ضرر آن به صورت زیر است:

$$\nabla_\theta J(\theta) = \sum_{t=1}^{T} \mathbb{E}\left[\nabla_\theta \log P\left(a_t \mid a_{1:(t-1)}; \theta\right) R\right]$$



این روش می‌تواند مزیت‌هایی مانند قابلیت کشف ساختارهای پیچیده‌تر و انعطاف پذیری در مقابل محیط‌های پویا را داشته باشد. ب) در مسأله تشخیص اشیا، ابر پارامترهایی مانند تصویر ورودی و همچنین پارامتر های معماری عصبی همچون تعداد لایه ها، در کارایی مدل تاثیر زیادی دارند. امکان استفاده از رویکرد جستجوی ذکر شده در بخش الف را با ذکر دلیل برای اندازه تصویر ورودی و تعداد لایه ها بررسی کنید(۱۰ نمره).

- اندازه تصویر ورودی : برای جستجوی اندازه ورودی به مسئله تشخیص اشیا. الگوریتم‌های NAS می‌توانند اندازه‌های مختلف را به عنوان انتخاب‌های معماری مورد بررسی قرار داده و یاد بگیرند که اندازه‌ای را انتخاب کنند که دقت اعتبارسنجی را به حداکثر برساند. رویکرد RL می‌تواند با استفاده از سیگنال پاداشی که دقت را نشان می‌دهد، اندازه ورودی بهینه را از طریق آزمون و خطا یاد بگیرد.
- تعداد لایه ها : با استفاده از الگوریتم‌های NAS، می‌توان ساختارهای مختلف لایه‌های کانولوشن را مدل کرده و تطبیق داده شوند. RL می‌تواند با استفاده از سیگنال پاداشی که دقت را نشان می‌دهد، لایه‌های کانولوشن بهینه را از طریق آزمون و خطا یاد بگیرد.

۵. اگر هنگام آموزش یک شبکه GAN استاندارد مقدار تابع ضرر مولد و ممیز در پایان $epoch$ اول و ۱۰۰ تقریباً یکسان باشند، چرا کیفیت تصاویر تولید شده در $epoch$ اول و ۱۰۰ لزوماً مشابه نیستند(۱۵ نمره)؟

نباید انتظار داشته باشیم که آنها یکسان باشند زیرا $loss$ نسبت به مدلها با کیفیت مختلف در طول زمان است. به عبارت دیگر $loss$ مولد در $epoch$ ۱ و ۱۰۰ نسبت به یک ممیز ممکن است بهبود یافته باشد و همین موضوع برای $loss$ ممیز نیز صدق می‌کند. مقدار تابع ضرر ($loss$) تنها یک نمایش عددی از عملکرد مدل در هر $epoch$ است و نشان دهنده کیفیت تصاویر تولیدی نیست. این توابع

ضرر ممکن است بهبود یابند اما کیفیت تصاویر تولید شده نتواند به همان اندازه بهبود یابد.