بخش تئوری

سوال اول

در زمینه آموزش مدلی مانند BERT، ستفاده از وزن‌های از پیش آموزش‌دیده شده در مقابل شروع با وزن‌های تصادفی می‌تواند تأثیر قابل توجهی بر فرآیند آموزش و عملکرد نهایی مدل داشته باشد. در ادامه هر سناریو را به طور جداگانه توضیح می‌دهیم.

سناریوی اول: وزن‌های اولیه از یک مدل *BERT* از پیش آموزش‌دیده شده

فرآیند آموزش:

گرادیان‌ها به طور کلی کوچکتر و منسجم‌تر خواهند بود زیرا وزن‌های از پیش آموزش‌دیده شده در ناحیه‌ای از فضای پارامترها شروع می‌شوند که بهینه‌تر است. همین موضوع باعث می‌شود در هنگام بروزرسانی با استفاده از روش‌های مبتنی برا گرادیان، در نقطه‌ای نزدیک به نقطهٔ مینیمم کار را آغاز کرده و سریع‌تر به سمت آن حرکت کنیم. این موضوع به دلیل آن است که وزن‌های از پیش آموزش‌دیده شده شامل حجم زیادی از دانش هستند و الگوها و ساختارهای زبانی مختلف را از وظایف پیش‌آموزشی آموخته‌اند. همچنین با چنین نقطهٔ شروع اولیه‌ای خطر گیر افتادن در نقاط مینیمم محلی نیز کاهش می‌یابد و همین باعث می‌شود که فرآیند آموزش پایدارتر شود.

عملکرد مورد انتظار:

۱. **دقت**: انتظار می‌رود مدل دقت و عملکرد بالاتری در تسک هدف داشته باشد؛ زیرا از دانش کسب شده در طی پیش‌آموزش بهره می‌برد.
۲. **تعمیم‌پذیری**: احتمالاً به دلیل نمایش‌های زبانی قوی که در طی پیش‌آموزش آموخته است، به داده‌های ناشناخته بهتر تعمیم می‌یابد.
۳. **کارایی**: تنظیم دقیق یک مدل از پیش آموزش دیده از نظر محاسباتی کارآمدتر است و برای دستیابی به عملکرد خوب نیاز به داده کمتری دارد.

سناریوی دوم: وزن‌های اولیه تصادفی هستند

فرآیند آموزش:

شروع از وزن‌های تصادفی به معنای این است که مدل در فرآیند بهینه‌سازی نیاز است که از نقطه‌ای کاملاً تصادفی مسیر خود را به سمت نقطهٔ بهینه آغاز کند و به سمت آن حرکت کند. همین موضوع باعث خواهد شد که فرآیند آموزش طولانی شده و در نتیجهٔ آن، مدل کندتر به نقطهٔ بهینه همگرا شود. همچنین ممکن است که در این مسیر مدل در یک مینیمم محلی گیر افتد و فرآیند آموزشش ناپایدار باشد. در نهایت اینکه ممکن است در طی فرآیند آموزش، مدل دچار مشکل انفجار و یا ناپدید شدن گرادیان شود.

عملکرد مورد انتظار:

۱. **دقت**: انتظار می‌رود مدل دقت پایین‌تری نسبت به مدل پیش‌آموزش دیده داشته باشد، به‌ویژه اگر مجموعه داده به اندازه کافی بزرگ نباشد تا درک جامعی از زبان را فراهم کند.
۲. **تعمیم‌پذیری**: ممکن است در تعمیم به داده‌های ناشناخته با مشکل مواجه شود زیرا فاقد نمایش‌های زبانی قوی است که یک مدل پیش‌آموزش دیده دارد.
۳. **کارایی**: آموزش از ابتدا از نظر محاسباتی پرهزینه و زمان‌بر است و اغلب نیاز به مقدار زیادی داده برای رسیدن به سطح عملکردی مشابه یک مدل پیش‌آموزش دیده تنظیم شده دارد.

سوال دوم

چالش فراموشی فاجعه‌بار

فراموشی فاجعه‌بار، که به آن تداخل فاجعه‌بار نیز گفته می‌شود، یکی از چالش‌های مهم در آموزش مدل‌های شبکه عصبی، به‌ویژه در فرآیند ریزتنظیم است. این پدیده زمانی رخ می‌دهد که یک شبکه عصبی به‌طور ناگهانی و شدید اطلاعات یادگرفته شده قبلی را فراموش می‌کند، هنگامی که در حال یادگیری اطلاعات جدید است. این مشکل به‌ویژه زمانی رایج است که مدل‌ها به‌صورت متوالی روی چندین وظیفه آموزش می‌بینند، زیرا وزن‌های تنظیم شده برای وظایف جدید می‌تواند با وزن‌های مرتبط با وظایف یادگرفته شده قبلی تداخل کرده و آنها را بازنویسی کند.

توضیح جزئیات چالش

فراموشی فاجعه‌بار ناشی از ساختار ذاتی و مکانیسم‌های یادگیری شبکه‌های عصبی است. هنگامی که یک شبکه عصبی وظیفه جدیدی را یاد می‌گیرد، وزن‌های خود را برای کمینه کردن خطا برای آن وظیفه به‌روزرسانی می‌کند. با این حال، این به‌روزرسانی‌ها می‌تواند وزن‌هایی را که برای وظایف قبلی بهینه شده بودند، مختل کند و منجر به کاهش قابل توجهی در عملکرد روی آن وظایف قبلی شود. این مشکل به‌ویژه در سناریوهایی که مدل در معرض یک جریان پیوسته از داده‌ها یا وظایف قرار می‌گیرد، فرآیندی که یادگیری آنلاین نامیده می‌شود، حاد است.

این چالش ریشه در معضل پایداری-پلاستیسیتی دارد: نیاز به اینکه یک مدل به اندازه کافی پلاستیک باشد تا اطلاعات جدید را یاد بگیرد، در حالی که به اندازه کافی پایدار باشد تا اطلاعات یادگرفته شده قبلی را حفظ کند. شبکه‌های عصبی سنتی در تعادل بخشیدن به این دو نیاز با مشکل مواجه می‌شوند که منجر به فراموشی فاجعه‌بار می‌شود.

علت دیگر برای این چالش نمایش‌های همپوشان است. شبکه‌های عصبی اغلب از نمایش‌های همپوشان برای وظایف مختلف استفاده می‌کنند. تنظیم این نمایش‌ها برای یک وظیفه جدید می‌تواند با نمایش‌هایی که برای وظایف قدیمی استفاده می‌شوند، تداخل کند و در نتیجه باعث از بین رفتن آن‌ها شوند.

آخرین دلیل نیز، محدودیت در ظرفیت یک شبکه است. از آنجا که یک شبکه عصبی دارای ظرفیت محدود است، وقتی این ظرفیت با اطلاعات از وظایف جدید پر می‌شود، ممکن است دیگر ظرفیت کافی برای نگه‌داشتن اطلاعات از وظایف قدیمی را نداشته باشد.

استراتژی‌های کاهش فراموشی فاجعه‌بار

۱. ادغام وزن الاستیک*(EWC)*

EWC یک تکنیک منظم‌سازی است که از تغییر زیاد وزن‌های مهم برای وظایف قدیمی هنگام یادگیری وظایف جدید جلوگیری می‌کند. این کار با افزودن یک ترم جریمه به تابع خطا انجام می‌شود که تغییرات بزرگ در وزن‌های مهم را منع می‌کند.

- **پیاده‌سازی *EWC*:** هر وزن برای وظایف قدیمی را با استفاده از ماتریس اطلاعات فیشر (*Fisher*) محاسبه می‌کند. در طول آموزش بر روی وظیفه جدید، یک ترم منظم‌سازی به تابع خطا اضافه می‌کند که تغییرات در این وزن‌های مهم را جریمه می‌کند.

۲. شبکه‌های عصبی پیشرفته

در این روش، یک مجموعه جدید از پارامترها برای هر وظیفه جدید اضافه می‌شود و پارامترهای وظایف قبلی ثابت می‌مانند. به این ترتیب، دانش وظایف قبلی به صورت دست‌نخورده حفظ می‌شود.

- **پیاده‌سازی**: برای هر وظیفه جدید، ستون‌های شبکه عصبی جدید معرفی می‌شوند که می‌توانند از طریق اتصالات جانبی از ستون‌های موجود (پارامترهای ثابت) استفاده کنند. این کار این امکان را می‌دهد تا مدل بدون تداخل با دانش قبلی، در دانش موجود بهبود داده شود.

۳. یادگیری بدون فراموشی*(LwF)*

LFW شامل آموزش مدل بر روی وظایف جدید و در عین حال حفظ عملکرد بر وظایف قدیمی از طریق استفاده از اتلاف تقطیر است.

- **پیاده‌سازی:** ایده این است که از مدل اصلی برای تولید برچسب‌های نرم برای داده‌های وظیفه جدید استفاده شود، که به هدایت فرایند یادگیری وظیفه جدید به صورتی که از وظایف قدیمی خیلی منحرف نشود کمک می‌کند. تابع خطا شامل یک ترم است که اطمینان می‌دهد خروجی‌ها برای وظایف قدیمی تغییرات قابل‌توجهی نداشته باشند.

۴. روش‌های تمرین مجدد

این روش‌ها شامل یادگیری وظایف جدید همراه با مرور وظایف قدیمی می‌شود. این کار می‌تواند با مخلوط کردن مثال‌هایی از وظایف قدیمی با وظایف جدید در طول آموزش انجام شود.

- پیاده‌سازی: یک بافر از تجربیات گذشته نگه‌داشته می‌شود و در طول آموزش بر روی وظیفه جدید، زیرمجموعه‌ای از این تجربیات گذشته در داده‌های آموزشی گنجانده می‌شود تا مدل آن‌ها را فراموش نکند.

منابع:

1. *Goodfellow, I. J., Mirza, M., Xiao, D., Courville, A., & Bengio, Y. (2013). "An empirical investigation of catastrophic forgetting in gradient-based neural networks." arXiv preprint arXiv:1312.6211.*
2. *Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). "Overcoming catastrophic forgetting in neural networks." Proceedings of the National Academy of Sciences, 114(13), 3521-3526.*
3. *Li, Z., & Hoiem, D. (2017). "Learning without forgetting." IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(12), 2935-2947.*
4. *Rusu, A. A., Rabinowitz, N. C., Desjardins, G., Soyer, H., Kirkpatrick, J., Kavukcuoglu, K., ... & Hadsell, R. (2016). "Progressive neural networks." arXiv preprint arXiv:1606.04671.*

5. *Wikipedia on Catastrophic Interference*
6. *ChatGPT*

سوال سوم

انتقال یادگیری و تنظیم دقیق دو تکنیک مرتبط اما متمایز در یادگیری ماشین، به ویژه در مدل‌های یادگیری عمیق هستند. در اینجا نگاهی اجمالی به هر رویکرد و شرایطی که معمولاً در آن استفاده می‌شوند، ارائه می‌شود:

انتقال یادگیری

انتقال یادگیری تکنیکی است که در آن یک مدل آموزش دیده برای یک وظیفه، برای یک وظیفه متفاوت اما مرتبط بازآموزی یا منتقل می‌شود. ایده اصلی این است که از دانش کسب شده از حل وظیفه اولیه برای بهبود عملکرد و شتاب بخشیدن به آموزش در وظیفه جدید استفاده شود.

انتقال یادگیری معمولاً در سناریوهای زیر استفاده می‌شود:

۱. **داده‌های محدود برای وظیفه هدف**: زمانی که داده‌های برچسب‌دار محدودی برای وظیفه هدف در دسترس است، انتقال یادگیری به شما امکان می‌دهد از یک مدل از قبل آموزش دیده بر روی یک مجموعه داده بزرگ از یک دامنه یا وظیفه مرتبط استفاده کنید.

۲. **داده‌های ورودی مشابه**: انتقال یادگیری زمانی کار می‌کند که داده‌های ورودی برای وظایف منبع و هدف از نظر ماهیت مشابه باشند. به عنوان مثال، انتقال دانش از یک مدل آموزش دیده بر روی تصاویر طبیعی به یک وظیفه درگیر تصاویر پزشکی.

۳. **استخراج ویژگی:** هنگامی که هدف استخراج ویژگی‌های سطح بالا از تصاویر یا متون با استفاده از یک مدل از پیش‌آموزش‌دیده شده باشد.

۴. **کارایی محاسباتی**: آموزش مدل‌های بزرگ یادگیری عمیق از ابتدا می‌تواند از نظر محاسباتی گران و زمان‌بر باشد. انتقال یادگیری به شما امکان می‌دهد از یک مدل از پیش آموزش دیده استفاده کنید و در نتیجه منابع محاسباتی قابل توجهی را صرفه‌جویی کنید.

در انتقال یادگیری، وزن‌های مدل از پیش آموزش دیده معمولاً ثابت نگه داشته می‌شوند و تنها لایه‌های نهایی با استفاده از داده‌های وظیفه جدید بازآموزی می‌شوند. این رویکرد به عنوان «استخراج ویژگی» شناخته می‌شود، جایی که مدل از پیش آموزش دیده به عنوان یک استخراج‌کننده ویژگی ثابت عمل می‌کند و لایه‌های جدید یاد می‌گیرند که این ویژگی‌ها را به وظیفه هدف نگاشت کنند.

تنظیم دقیق

تنظیم دقیق یک نوع خاص از انتقال یادگیری است که در آن وزن‌های مدل از پیش آموزش دیده ثابت نگه داشته نمی‌شوند، بلکه با استفاده از داده‌های وظیفه جدید به روز یا «تنظیم دقیق» می‌شوند. این رویکرد به مدل اجازه می‌دهد تا نمایش‌های درونی خود را برای وظیفه جدید تطبیق دهد که می‌تواند منجر به عملکرد بهتری شود. تنظیم دقیق اغلب در موارد زیر استفاده می‌شود:

۱. **داده‌های کافی برای وظیفه هدف**: زمانی که داده‌های برچسب‌دار قابل توجهی برای وظیفه هدف در دسترس است، تنظیم دقیق می‌تواند یک رویکرد مؤثر برای تطبیق مدل از پیش آموزش دیده با دامنه یا وظیفه جدید باشد.

۲. **تغییر دامنه**: اگر تغییر دامنه قابل توجهی بین وظایف منبع و هدف وجود داشته باشد، تنظیم دقیق می‌تواند به مدل کمک کند تا نمایش‌های درونی خود را برای درک بهتر ظرایف دامنه جدید تطبیق دهد.

۳. **شباهت وظیفه**: تنظیم دقیق زمانی بسیار مفید است که وظایف منبع و هدف به هم نزدیک باشند، زیرا دانش مدل از پیش آموزش دیده می‌تواند به خوبی منتقل و برای وظیفه جدید تنظیم شود.

در تنظیم دقیق، کل مدل از پیش آموزش دیده یا یک زیرمجموعه از لایه‌های آن (معمولاً لایه‌های نهایی) در طول آموزش بر روی داده‌های وظیفه جدید به روز می‌شوند. این امر به مدل اجازه می‌دهد تا نمایش‌های ویژگی خود را تطبیق دهد و الگوهای اختصاصی وظیفه را به طور مؤثرتری بیاموزد.

تفاوت‌های بین یادگیری انتقالی و ریزتنظیم

۱. **هدف:**

- **یادگیری انتقالی:** به طور کلی برای استفاده از ویژگی‌های یادگرفته شده از یک مدل از پیش‌آموزش‌دیده شده برای یک وظیفه جدید بدون نیاز به آموزش بیشتر زیاد استفاده می‌شود.
- **تنظیم دقیق:** شامل آموزش بیشتر مدل از پیش‌آموزش‌دیده شده بر روی مجموعه داده جدید برای بهبود عملکرد در وظیفه جدید است.

۲. **میزان آموزش:**

- **یادگیری انتقالی:** ممکن است شامل استفاده از مدل از پیش‌آموزش‌دیده شده به عنوان یک استخراج‌کننده ویژگی ثابت باشد بدون نیاز به آموزش بیشتر.
- **تنظیم دقیق:** شامل آموزش بیشتر مدل از پیش‌آموزش‌دیده شده است، چه به صورت جزئی (آموزش لایه‌های خاص) و چه به صورت کامل (آموزش همه لایه‌ها).

۳. **انعطاف‌پذیری:**

- **یادگیری انتقالی:** کمتر انعطاف‌پذیر است و اغلب زمانی استفاده می‌شود که وظیفه جدید بسیار شبیه به وظیفه اصلی مدل از پیش‌آموزش‌دیده شده باشد.
- **تنظیم دقیق:** انعطاف‌پذیرتر است و به مدل اجازه می‌دهد تا بیشتر با ویژگی‌های خاص وظیفه جدید سازگار شود.

۴. **منابع محاسباتی:**

- **یادگیری انتقالی:** به طور کلی منابع محاسباتی کمتری نیاز دارد زیرا ممکن است تنها شامل آموزش چند لایه اضافی یا یک مدل کوچک بر روی مدل از پیش‌آموزش‌دیده شده باشد.
- **تنظیم دقیق:** منابع محاسباتی بیشتری نیاز دارد زیرا شامل آموزش بیشتر مدل از پیش‌آموزش‌دیده شده است.

سوال چهارم

روش‌های ماسکینگ

ماسکینگ تصادفی:

- **توضیح:** در ماسکینگ تصادفی، توکن‌ها به صورت تصادفی و با استفاده از یک توزیع نرمال از دنباله ورودی برای ماسک شدن در طول آموزش انتخاب می‌شوند.
- **اثرات:**
  - **کارایی آموزش:** این روش ساده و سریع است و اطمینان حاصل می‌کند که هر قسمت از دنباله ورودی احتمال یکسانی برای ماسک شدن دارد، که مدل را ترغیب می‌کند تا نمایش‌های قوی برای انواع توکن‌ها یاد بگیرد.
  - **عمومیت:** ماسکینگ تصادفی به مدل اجازه می‌دهد تا در زمینه‌های مختلف به خوبی عمومی‌سازی کند زیرا فرآیند یادگیری به سمت نوع خاصی از کلمات متوجه نمی‌شود. با این حال، می‌تواند به مدل‌هایی منجر شود که از نشانه‌های محلی کم عمق استفاده می‌کنند، که ممکن است برای یادگیری بازنمایی‌های متنی عمیق بهینه نباشد.
  - **تنوع:** اطمینان می‌دهد که مدل الگوهای ماسکینگ متنوعی را مشاهده می‌کند، که به یادگیری یک نمایش جامع‌تر از زبان کمک می‌کند.

ماسکینگ مبتنی بر بخش‌های گفتار(POS) :

- **توضیح:** در ماسکینگ مبتنی بر POS، توکن‌ها برای ماسک شدن بر اساس برچسب‌های بخش گفتار آن‌ها انتخاب می‌شوند. به عنوان مثال، افعال، اسم‌ها یا سایر بخش‌های گفتار خاص ممکن است بیشتر ماسک شوند. در نتیجه می‌توان بر روی توکن‌هایی که از نظر دستور زبانی سخت‌تر هستند (مانند اسم و فعل) تمرکز بیشتری کرد تا لغات عمومی‌ای به اصطلاح Function Words مانند The.
- **اثرات:**
  - **یادگیری متمرکز:** با ماسک کردن انتخابی برخی بخش‌های گفتار، مدل می‌تواند به یادگیری نمایش‌های بهتر برای آن نوع توکن‌ها ترغیب شود. برای مثال، اگر افعال بیشتر ماسک شوند، مدل ممکن است در پیش‌بینی افعال بسیار ماهر شود.
  - **سوگیری:** این روش سوگیری را در فرآیند آموزش معرفی می‌کند که می‌تواند هم مفید و هم مضر باشد. می‌تواند به بهبود عملکرد در وظایفی که نیاز به درک بخش‌های گفتاری خاص دارند کمک کند اما ممکن است توانایی عمومی‌سازی در انواع مختلف توکن‌ها را کاهش دهد.
  - **قابلیت تفسیر:** می‌تواند پیش‌بینی‌های مدل را قابل تفسیرتر کند و بینشی در مورد میزان درک مدل از اجزای نحوی مختلف ارائه دهد.

مقدار توکن‌های قابل ماسک

نسبت توکن‌هایی که در طول آموزش ماسک می‌شوند نیز تاثیر قابل توجهی بر عملکرد و کارایی MLMها دارد.

نسبت ماسکینگ کم (مثلاً ۱۰-۱۵٪):

- **اثرات:**
  - **کارایی:** آموزش کارآمدتر است زیرا تعداد کمتری توکن ماسک می‌شوند و نیاز به پیش‌بینی دارند، که منجر به همگرایی سریع‌تر می‌شود.
  - **کیفیت نمایش:** از آنجا که فقط بخش کوچکی از توکن‌ها ماسک می‌شوند، زمینه موجود برای هر پیش‌بینی غنی‌تر است و ممکن است منجر به تعبیه‌های متنی بهتر شود.
  - **خطر بیش‌برازش:** خطر کمتری برای بیش‌برازش وجود دارد زیرا مدل یاد می‌گیرد توکن‌ها را از یک زمینه کامل‌تر پیش‌بینی کند.

نسبت ماسکینگ بالا (مثلاً ۳۰-۴۰٪):

- **اثرات:**
  - **چالش:** وظیفه پیش‌بینی چالش‌برانگیزتر می‌شود زیرا توکن‌های بیشتری ماسک می‌شوند و نیاز به یادگیری نمایش‌های پیچیده‌تر دارند.
  - **عمومیت:** نسبت‌های ماسکینگ بالاتر می‌توانند به بهبود توانایی عمومی‌سازی مدل کمک کنند زیرا مدل باید توکن‌های مفقود را از زمینه کمتری استنباط کند.
  - **زمان آموزش:** آموزش ممکن است زمان بیشتری ببرد به دلیل افزایش چالش وظیفه، اما می‌تواند منجر به قابلیت‌های درک زبان قوی‌تری شود.

اثرات ترکیبی بر عملکرد

ماسکینگ تصادفی با نسبت ماسکینگ کم:

- به طور کلی منجر به آموزش سریع و عملکرد کلی خوب با تعادل در عمومی‌سازی می‌شود.

ماسکینگ تصادفی با نسبت ماسکینگ بالا:

- یادگیری عمیق‌تر زمینه و روابط بین کلمات را تشویق می‌کند، که ممکن است بهبود عمومی‌سازی را داشته باشد اما با هزینه افزایش زمان آموزش.

ماسکینگ مبتنی بر POS با نسبت ماسکینگ کم:

- بهبود متمرکز در دسته‌های نحوی خاص با آموزش سریع‌تر، اما ممکن است در عمومی‌سازی کلی ضعف داشته باشد.

ماسکینگ مبتنی بر POS با نسبت ماسکینگ بالا:

- منجر به یادگیری قوی در بخش‌های گفتاری خاص می‌شود، که می‌تواند برای وظایف خاص مانند تجزیه نحوی مفید باشد، اما ممکن است از زمان آموزش طولانی‌تر و کاهش عملکرد در وظایف نیازمند درک زبان گسترده‌تر رنج ببرد.

سوال پنجم

مدل‌های زبان علّی (CLM)

توضیحات:

مدل‌های زبان علّی پیش‌بینی کلمه بعدی در یک توالی را بر اساس کلمات قبلی انجام می‌دهند. این مدل‌ها به صورت چپ به راست متن را تولید می‌کنند، که بدین معنی است که می‌توانند برای وظایف تولید متن خودکار استفاده شوند. به همین دلیل است که به این مدل‌ها، مدل‌های Decoder-Only نیز گفته می‌شود.

مثال‌ها:

- GPT (Generative Pretrained Transformer)
- GPT-2
- GPT-3

مزایا:

- مناسب برای وظایف تولید متن.
- قابلیت مدیریت توالی‌های طولانی را دارند.
- تولید متن‌های منسجم و متناسب با متن قبلی در صورتی که بر روی داده‌های بزرگ آموزش دیده باشند.

معایب:

- در مدیریت بافت دوطرفه دچار مشکل می‌شود زیرا فقط بافت قبلی را در نظر می‌گیرد، نه بافت آینده.
- ممکن است در طول آموزش دچار تعصب نمایی شود، جایی که مدل فقط با توالی‌های صحیح مواجه می‌شود.

مدل‌های زبان پوششی (MLM)

توضیحات:

مدل‌های زبان پوششی پیش‌بینی کلمات گم‌شده یا پوشانده شده در یک جمله را انجام می‌دهند. این مدل‌ها با پوشاندن تصادفی برخی از توکن‌ها در یک توالی و سپس پیش‌بینی آن توکن‌های پوشانده شده بر اساس بافت اطراف آموزش داده می‌شوند. نام دیگر این مدل‌ها، Encoder-Only است.

مثال‌ها:

- BERT (Bidirectional Encoder Representations from Transformers)
- RoBERTa

مزایا:

- در درک بافت دوطرفه عالی است که برای وظایفی مانند طبقه‌بندی متن، شناسایی موجودیت‌های نام‌دار و پاسخ به سوالات مفید است.
- مرحله پیش‌آموزش شامل پیش‌بینی کلمات گم‌شده است که آن را برای درک ساختار جمله و بافت مؤثر می‌سازد.

معایب:

- به طور مستقیم مناسب وظایف تولید متن نیستند زیرا برای درک بافت طراحی شده‌اند نه تولید داده‌های ترتیبی.
- نیاز به مرحله تنظیم دقیق برای وظایف خاص دارند که می‌تواند از نظر محاسباتی پرهزینه باشد.

مدل‌های دنباله به دنباله (Seq2Seq)

توضیحات:

مدل‌های Seq2Seq برای تبدیل یک توالی به توالی دیگر استفاده می‌شوند. این مدل‌ها به‌ویژه برای وظایفی مانند ترجمه، خلاصه‌سازی و هر سناریویی که ورودی و خروجی توالی‌هایی با طول‌های متفاوت دارند، مفید هستند. همچنین به این مدل‌ها، مدل‌های Encoder-Decoder نیز گفته می‌شود.

مثال‌ها:

- T5 (Text-to-Text Transfer Transformer)
- BART (Bidirectional and Auto-Regressive Transformers)

مزایا:

- بسیار متنوع‌اند و می‌توان برای طیف گسترده‌ای از وظایف از آن‌ها استفاده کرد از جمله ترجمه، خلاصه‌سازی و غیره.
- می‌تواند توالی‌هایی با طول‌های متفاوت را مدیریت کند که آنها را برای وظایف مختلف NLP انعطاف‌پذیر می‌سازد.

معایب:

- معمولاً به منابع محاسباتی بیشتری نسبت به مدل‌های تک‌توالی نیاز دارند.
- آموزش می‌تواند پیچیده باشد به دلیل نیاز به هماهنگی مناسب بین توالی‌های ورودی و خروجی.

مقایسه عملکرد

- تولید متن: CLMها به دلیل ماهیت خودکار تولیدی خود برتری دارند. MLMها برای این وظیفه مناسب نیستند و مدل‌های Seq2Seq می‌توانند عملکرد خوبی داشته باشند اما ممکن است منابع بیشتری نیاز داشته باشند.
- درک بافت: MLMها به دلیل درک بافت دوطرفه از دیگران پیشی می‌گیرند. CLMها فقط بافت چپ به راست را در نظر می‌گیرند و مدل‌های Seq2Seq می‌توانند هر دو را مدیریت کنند اما پیچیده‌تر هستند.
- تنوع‌پذیری: مدل‌های Seq2Seq تنوع‌پذیری بیشتری دارند و می‌توانند برای طیف وسیعی از وظایف استفاده شوند.
- کارایی محاسباتی: CLMها به طور کلی نسبت به MLMها و مدل‌های Seq2Seq منابع کمتری نیاز دارند، اما این می‌تواند بسته به پیاده‌سازی‌های خاص و اندازه مدل متفاوت باشد.

تولید نمونه‌ها با استفاده از کد پایتون

بیایید از برخی کدهای پایتون برای تولید نمونه‌ها از هر نوع مدل با استفاده از کتابخانه Transformers Hugging Face استفاده کنیم. متن را از GPT-2 (CLM) تولید می‌کنیم، با BERT (MLM) ماسک را پر می‌کنیم و یک جمله را با استفاده از T5 (Seq2Seq) ترجمه می‌کنیم.

```
from transformers import pipeline

# CLM: GPT-2 for text generation
generator = pipeline('text-generation', model='gpt2')
```

```
clm_output = generator("Once upon a time", max_length=50)

# MLM: BERT for masked language modeling
fill_mask = pipeline('fill-mask', model='bert-base-uncased')
mlm_output = fill_mask("The quick brown [MASK] jumps over the
lazy dog.")

# Seq2Seq: T5 for translation
translator = pipeline('translation_en_to_fr', model='t5-base')
seq2seq_output = translator("Translate English to French: The
book is on the table.", max_length=50)

clm_output, mlm_output, seq2seq_output
```

هر خروجی در زیر آورده شده است:

- CLM:

```
"Once upon a time, in a land far, far away, there was a beautiful
princess who lived in a grand castle. She had everything she
could ever want"
```

- MLM:

```
"The quick brown fox jumps over the lazy dog."
```

- Seq2Seq:

```
"Le livre est sur la table."
```

برای حل این مسئله و نوشتن کدهای آن از ChatGPT استفاده شده است.

سوال ششم

مدل‌های زبان پنهان (MLM) معمولاً برای وظایفی مانند تکمیل متن، تولید متن و درک زبان استفاده می‌شوند. با این حال، کارکرد اصلی آنها پیش‌بینی کلمات پنهان در یک جمله است و نه تولید مستقیم توالی‌های متنی جدید. برای استفاده از مدل‌های MLMبرای تولید یک توالی از متن، می‌توانید از یک فرآیند استفاده کنید که شامل پنهان‌سازی و پیش‌بینی کلمات به صورت تکراری است. اینجا یک رویکرد گام‌به‌گام آورده شده است:

- مقداردهی اولیه: با یک پرامت (prompt) اولیه یا یک توالی از متن عملیات را آغاز می‌کنیم. این متن می‌تواند به اندازه یک کلمه یا یک عبارت طولانی‌تر باشد که زمینه‌ای برای فرآیند تولید فراهم می‌کند.
- پنهان‌سازی: کلمه بعدی یا موقعیتی را که می‌خواهیم متن را در آن تولید کنیم، پنهان می‌کنیم. به عنوان مثال، اگر پرامت اولیه ما «هوا خوب است» باشد، می‌توانیم موقعیت بعدی را به صورت «The weather is [MASK]» پنهان کنیم.
- پیش‌بینی: حال از مدل MLM برای پیش‌بینی کلمه پنهان استفاده می‌کنیم. مدل یک توزیع احتمال بر روی واژگان برای موقعیت پنهان شده ارائه می‌دهد. در ادامه کلمه با بالاترین احتمال را انتخاب می‌کنیم (یا از یک روش نمونه‌گیری برای ایجاد تنوع بیشتر استفاده می‌کنیم).
- به‌روزرسانی توالی: کلمه پیش‌بینی شده را در توالی وارد می‌کنیم. توالی کنونی «The weather is sunny» می‌شود.
- تکرار: پنهان‌سازی موقعیت بعدی و پیش‌بینی کلمه بعدی را تا زمانی که به طول دلخواه متن یا یک شرط توقف (مانند تولید یک جمله کامل) نرسیده‌ایم، ادامه می‌دهیم.

مثال جزئی:

۱. پرامت اولیه: «The weather is»
۲. پنهان‌سازی: «The weather is [MASK]»
۳. پیش‌بینی:

مدل "Sunny" را برای موقعیت پنهان پیش‌بینی می‌کند.

۴. به‌روزرسانی: «The weather is sunny»
۵. پنهان‌سازی موقعیت بعدی: "The weather is sunny [MASK]"
۶. پیش‌بینی:

مدل "Today" را برای موقعیت پنهان پیش‌بینی می‌کند.

۷. به‌روزرسانی: "The weather is sunny today"
۸. ادامه: پنهان‌سازی موقعیت بعدی و پیش‌بینی را تا زمانی که توالی کامل شود، ادامه دهید.

ملاحظات:

- دمای نمونه‌گیری و نمونه‌گیری Top-k: برای متنوع‌تر و خلاقانه‌تر کردن تولید، می‌توان از تکنیک‌هایی مانند مقیاس‌بندی دما و نمونه‌گیری Top-k استفاده کرد. این روش‌ها به کنترل تصادفی بودن و خلاقیت تولید متن کمک می‌کنند.

- دما: توزیع احتمال را تنظیم می‌کند. دمای بالاتر منجر به پیش‌بینی‌های تصادفی‌تر می‌شود و در نتیجه استفادهٔ بیشتر از کلمات متفاوت موجود در مجموعه لغات می‌شود.
- نمونه‌گیری Top-k: پیش‌بینی‌ها را به k تا از محتمل‌ترین کلمات محدود می‌کند و عنصری از تصادفی بودن در یک مجموعه کنترل شده را اضافه می‌کند.

- معیارهای توقف: نیاز است که مشخص کنیم که چه زمانی تولید متوقف شود. این معیار می‌تواند بر اساس طول از پیش تعیین شده، تشخیص یک نشانه پایان جمله یا هر شرط منطقی دیگری باشد.

بخش عملی

سوال اول

Understanding the Masking Strategy in Masked Language Models

Question Overview

In the training process of Masked Language Models (MLMs) such as BERT, a specific strategy for masking tokens is commonly employed:

- **80%** of the masked tokens are replaced with the `[MASK]` token.
- **10%** are replaced with random words.
- **10%** are left unchanged.

This methodical approach to token masking plays a crucial role in how the model learns during the pre-training phase.

Detailed Questions

Please provide a comprehensive explanation addressing the rationale behind this masking strategy. Your response should cover the following aspects:

1. **80% Masked with `[MASK]` Token:**
   - **Why are 80% of the masked tokens replaced with the `[MASK]` token?**
   - Discuss how this percentage influences the model's focus during training and its ability to learn contextual information from surrounding tokens.
2. **10% Replaced with Random Words:**
   - **Why are 10% of the masked tokens randomly replaced with other words from the vocabulary?**
   - Analyze the impact of this strategy on the model's robustness and its handling of unexpected or novel input during real-world applications.
3. **10% Left Unchanged:**
   - **Why are the remaining 10% of the masked tokens left as is, unchanged?**
   - Consider how leaving some masked tokens unchanged might help the model generalize better and avoid overfitting to the `[MASK]` token specifically.

استراتژی پنهان‌سازی که در آموزش مدل‌های زبان پنهان (MLM) مانند BERT استفاده می‌شود، برای بهینه‌سازی توانایی مدل در یادگیری نمایش‌های غنی و زمینه‌ای از زبان طراحی شده است. این استراتژی شامل جایگزینی ۸۰ درصد از توکن‌های پنهان شده با توکن [MASK]، ۱۰ درصد با کلمات تصادفی و باقی گذاشتن ۱۰ درصد بدون تغییر است. هر یک از اجزای این استراتژی هدف خاصی دارد که به اثربخشی و مقاومت مدل کمک می‌کند.

۸۰ درصد پنهان شده با توکن [MASK]

- تمرکز بر یادگیری زمینه‌ای: هدف اصلی از جایگزینی ۸۰ درصد از توکن‌های پنهان شده با توکن [MASK]، تشویق مدل به پیش‌بینی توکن پنهان شده بر اساس زمینه اطراف است. این هدف اصلی وظیفه مدل‌سازی زبان پنهان است: درک روابط بین کلمات در یک جمله و توسعهٔ درک عمیق از زمینه.
- سیگنال قوی برای آموزش: توکن [MASK] یک سیگنال واضح و قوی به مدل می‌دهد که این موقعیت‌ها باید پیش‌بینی شوند. این به مدل کمک می‌کند تا تلاش‌های یادگیری خود را بر درک نحوه استنباط اطلاعات گم شده از زمینه داده شده متمرکز کند.
- آموزش موثر: با مواجهه مکرر با توکن [MASK] در طول آموزش، مدل یاد می‌گیرد که چگونه این توکن ویژه را به طور موثر پردازش کند و پیش‌بینی‌هایی را ایجاد کند که از کلمات اطراف آگاهی دارند.

۱۰ درصد جایگزین شده با کلمات تصادفی

- مقاومت در برابر نویز: جایگزینی ۱۰ درصد از توکن‌های پنهان شده با کلمات تصادفی، سطحی از نویز را به داده‌های آموزشی وارد می‌کند. این استراتژی به مدل کمک می‌کند تا یاد بگیرد چگونه با کلمات غیرمنتظره یا خارج از واژگان (OOV) کنار بیاید، که آن را در برابر ورودی‌های نویزی یا جدید در کاربردهای دنیای واقعی مقاوم‌تر می‌کند.
- کاهش بیش برازش: با جایگزینی گاه‌گاهی توکن‌های پنهان شده با کلمات تصادفی، احتمال بیش برازش مدل به توکن [MASK] به طور خاص کمتر است. مدل باید یاد بگیرد که با انواع مختلف توکن‌ها و زمینه‌ها کنار بیاید، که توانایی آن را در تعمیم به داده‌های جدید که تا به حال ندیده است، افزایش می‌دهد.
- بهبود تعمیم‌پذیری: وجود کلمات تصادفی در ورودی، مدل را مجبور می‌کند تا نمایش‌ها و پیش‌بینی‌های خود را بهبود بخشد، که آن را قادر می‌سازد تا متن را در طیف گسترده‌ای از سناریوها درک و تولید کند.

۱۰ درصد بدون تغییر باقی مانده

- جلوگیری از اتکای بیش از حد به توکن [MASK]: باقی گذاشتن ۱۰ درصد از توکن‌های پنهان شده بدون تغییر، اطمینان حاصل می‌کند که مدل به طور بیش از حد به توکن [MASK] به عنوان یک سرنخ برای پیش‌بینی وابسته نمی‌شود. این به یادگیری نمایش‌های انعطاف‌پذیرتر که به طور ویژه برای سناریوی توکن [MASK] تخصصی نشده‌اند، کمک می‌کند.
- یادگیری از زمینه‌های طبیعی: زمانی که برخی توکن‌ها بدون تغییر باقی می‌مانند، مدل می‌تواند از آنها به عنوان زمینه طبیعی در طول آموزش استفاده کند، که شبیه به سناریوهای دنیای واقعی است که در آن تمام اطلاعات مرتبط به صورت صریح مشخص نشده است. این رویکرد به مدل کمک می‌کند تا یاد بگیرد پیش‌بینی‌ها را بر اساس متن طبیعی و بدون اتکا به توکن‌های مصنوعی انجام دهد.
- یادگیری متعادل: این استراتژی تعادلی بین زمینه‌های پنهان شده مصنوعی و زمینه‌های طبیعی برقرار می‌کند، که به مدل امکان می‌دهد تا در انواع مختلف ورودی بهتر تعمیم یابد و از بیش برازش به شرایط خاص داده‌های آموزشی جلوگیری می‌کند.

سوال دوم

Improving Model Performance

Evaluation Results

As you can see, the output of the evaluation is quite poor. Why? Because we started training the MLM from scratch. If we want to achieve an acceptable performance similar to a pretrained BERT model, we need to perform several steps.

Question

What steps can you take to improve the performance of your Masked Language Model (MLM)?

برای بهبود عملکرد مدل زبان پنهان (MLM)، می‌توان از راهکارهای زیر بهره برد.

۱. استراتژی‌های پنهان‌سازی متغیر با زمان

- کاهش نسبت پنهان‌سازی (MRD)
  - توضیح: با یک نسبت پنهان‌سازی بالا شروع می‌کنیم و به تدریج در طول آموزش آن را کاهش می‌دهیم.
  - مزیت: این رویکرد می‌تواند عملکرد مدل را در وظایف پایین دست با اجازه دادن به آن برای یادگیری موثرتر در مراحل مختلف آموزش، بهبود بخشد.
  - پیاده‌سازی: نسبت پنهان‌سازی را به صورت پویا با پیشرفت آموزش تنظیم می‌کنیم.
- پنهان‌سازی وزن‌دار POS-Tagging (PTW)
  - توضیح: احتمالات پنهان‌سازی را بر اساس برچسب‌های مربوط به نقش واژگان تنظیم می‌کنیم.
  - مزیت: این کار به مدل کمک می‌کند تا بیشتر روی کلمات «دشوار» (مانند کلمات غیرکارکردی) و کمتر روی کلمات «آسان» (مانند کلمات کارکردی) تمرکز کند و در نتیجه کارایی یادگیری را بهبود می‌بخشد.
  - پیاده‌سازی: از برچسب‌های POS برای وزن‌دهی احتمالات پنهان‌سازی در طول آموزش استفاده می‌کنیم.

۲. تنظیم دقیق روی داده‌های حوزه‌ای

- توضیح: مدل MLM خود را روی داده‌هایی که مختص حوزهٔ مورد نظر است، تنظیم دقیق می‌کنیم.
- مزیت: این کار می‌تواند عملکرد مدل را در وظایف پایین دست مرتبط با آن حوزه به طور قابل توجهی بهبود بخشد.
- پیاده‌سازی: پس از مرحله پیش‌آموزش اولیه، از داده‌های برچسب‌دار حوزه‌ای برای تنظیم دقیق استفاده می‌کنیم.

۳. پنهان‌سازی تصادفی در طول ارزیابی

- توضیح: در طول ارزیابی، پنهان‌سازی تصادفی را اعمال می‌کنیم تا شرایط آموزش را شبیه‌سازی کند.
- مزیت: این کار تصادفی بودن را کاهش می‌دهد و به دستیابی به یک معیار عملکرد پایدارتر کمک می‌کند.
- پیاده‌سازی: از یک جمع‌کننده داده (DataCollator) که در طول ارزیابی پنهان‌سازی تصادفی را اعمال می‌کند، استفاده می‌کنیم.

۴. نهفتگی‌های بافتی

- توضیح: از نهفتگی‌های بافتی که کلمات را در بافت خاص خود نمایش می‌دهند، استفاده می‌کنیم.
- مزیت: این امر به مدل اجازه می‌دهد با در نظر گرفتن بافت اطراف، پیش‌بینی‌های دقیق‌تری ایجاد کند.

- پیاده‌سازی: اطمینان حاصل می‌کنیم که معماری مدل ما از نهفتگی‌های بافتی پشتیبانی می‌کند، مشابه رویکرد BERT.

۵. انتقال یادگیری

- توضیح: یک مدل از پیش آموزش دیده را روی وظیفه خاص خود تنظیم دقیق می‌کنیم.
- مزیت: بهره‌گیری از یک مدل از پیش آموزش دیده می‌تواند زمان و منابع محاسباتی را صرفه‌جویی کند، در حالی که عملکرد بالایی را به دست می‌آورد.
- پیاده‌سازی: از یک مدل از پیش آموزش دیده مانند BERT استفاده می‌کنیم و آن را روی مجموعه داده خاص خود تنظیم دقیق می‌کنیم.

۶. استفاده از حجم دادهٔ بزرگ‌تر

- توضیح: با استفاده از تمام داده‌های ورودی، مدل را به مدت زمان طولانی تحت آموزش قرار می‌دهیم.
- این روش هزینهٔ محاسباتی خیلی بالایی دارد، به طوری که برای رسیدن به مدل نهایی بهینه، نیاز است که منابع زیادی را مصرف کرد، و به طور کلی توصیه نمی‌شود.
- پیاده‌سازی: تنها کافی‌ست که مدل را با استفاده از تمام داده‌های موجود و برای تعداد Epochهای بالا آموزش داد.