به نام خدا

دانشگاه صنعتی شریف
دانشکده‌ی مهندسی برق

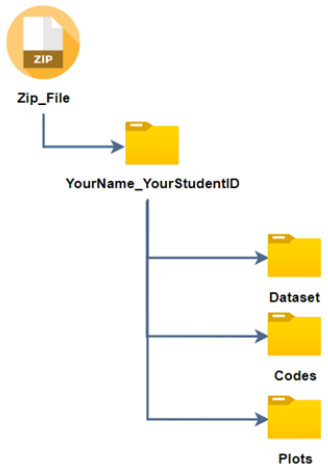
یادگیری عمیق

پاییز ۱۴۰۲
استاد: دکتر فاطمی‌زاده

پروژه نهایی | GAN-BERT | مهلت ارسال: ۱۲ بهمن

لطفاً به هنگام انجام پروژه و آماده‌سازی نتایج به موارد زیر توجه نمایید.

- گزارش پروژه باید به صورت کامل و با تمام جزئیات نوشته شود. در گزارش خود، بخش‌ها و زیربخش‌های مربوط به هر بخش را به صورت جداگانه بیاورید.
- کدها خوانا و مرتب نوشته شده و تا حد امکان کامنت‌گذاری شوند.
- کدها بدون ایراد اجرا شده و خروجی‌های مطلوب را تولید نمایند. بدیهی است در صورتی که کد دارای ایراد بوده و اجرا نشود، نمره‌ی آن بخش به دانشجو تعلق نمی‌گیرد.
- در انجام پروژه مشورت مجاز است ولی بدیهی است در صورت مشاهده هرگونه تشابه غیر معمول بین کدها و یا نتایج، طرفین نمره صفر از پروژه دریافت خواهند کرد.
- لطفاً هرگونه ابهام و یا سؤال را در کوئرای درس مطرح نمایید تا سایر دانشجویانی که سوالی مشابه دارند نیز به پاسخ‌ها دسترسی داشته باشند.
- در تمامی فایلها، برای آنکه نتایج بدست آمده قابلیت باز تولید داشته باشند، حتماً seed ها با مقدار ۴۲ تنظیم نمایید.
- در پایان تمامی مستندات لازم را در یک فایل zip قرار دهید. نام فایل باید به صورت YourName-YourStudentID باشد. در داخل فایل zip باید یک پوشه به همین نام وجود داشته باشد. توجه: لازم نیست تا دیتاست مورد استفاده را به همراه فایل‌های خود آپلود نمایید، بلکه ضروری است تا آدرس دهی‌ها برای دسترسی به دیتاست مطابق ساختار زیر باشد.



شکل ۱: ساختار مطلوب فایل‌های ارسالی

مقدمه

یادگیری نیمه‌نظارتی و محدودیت دادگان آموزشی: حوزه یادگیری عمیق در سال‌های اخیر پیشرفت‌های چشمگیری را شاهد بوده است. با این حال، یکی از چالش‌های اصلی در این حوزه، نیاز به داده‌های بسیار زیاد برای آموزش مدل‌های یادگیری عمیق است. این داده‌ها می‌توانند از منابع مختلفی نظیر تصاویر، متن، صدا و غیره باشند. تهیه و پردازش این حجم از داده‌ها هم زمان‌بر و هم هزینه‌بر است. برای مثال، برای جمع‌آوری داده‌های تصویری ممکن است نیاز به تجهیزات خاص، مجوزهای قانونی و یا حتی انجام عملیات پیچیده‌ای مانند برچسب‌زنی دستی به داده‌ها باشد. به همین دلیل یکی از شاخه‌های تحقیقاتی فعال در حوزه یادگیری عمیق تمرکز بر روی روش‌هایی است که با استفاده از دادگان آموزشی کمتر، بتوان به نتایج قابل قبولی دست یافت. این روش‌ها شامل تکنیک‌هایی نظیر یادگیری بدون نظارت [۱]، یادگیری نیمه‌نظارتی [۲] و یادگیری تقویتی [۳] هستند. یادگیری نیمه‌نظارتی، یک روش در حوزه یادگیری ماشین است که از ترکیبی از داده‌های برچسب‌دار و بدون برچسب استفاده می‌کند. این روش دارای اهمیت بالایی بوده زیرا معمولاً داده‌های برچسب‌دار کمیاب هستند در حالی که داده‌های بدون برچسب به راحتی در دسترس هستند.

مدل‌های ترنسفورمری و حوزه NLP: ترنسفورمرها، که برای اولین بار در مقاله "Attention is All You Need" در سال ۲۰۱۷ معرفی شدند، یک انقلاب در حوزه پردازش زبان‌های طبیعی (NLP) ایجاد کردند. این مدل‌ها با استفاده از مکانیزم توجه (Attention Mechanism) ، قادر به درک برداری از ارتباطات پیچیده بین کلمات در یک جمله یا متن هستند. ترنسفورمرها باعث شدند که ما بتوانیم مدل‌هایی را آموزش دهیم که قادر به تولید متن، ترجمه ماشینی، خلاصه‌سازی متن و دیگر وظایف (NLP) با دقت بسیار بالا هستند. برخی از مدل‌های معروف که بر پایه ترنسفورمرها ساخته شده‌اند عبارتند از: GPT-3 ، GPT-2 ، BERT و T5.

یکی از دلایل اصلی این انقلاب این است که ترنسفورمرها بر محدودیت‌های مدل‌های قبلی، مانند RNNs و LSTM، فائق آمده‌اند. به عنوان مثال، با استفاده از این مدل‌ها، می‌توان متن‌های بسیار طولانی‌تر را پردازش کرده و ارتباطات بین کلماتی که فاصله زیادی از هم دارند را درک کرد. با این حال، این مدل‌ها نیز محدودیت‌های مخصوص به خود را دارند. به عنوان مثال، آن‌ها نیاز به داده‌های آموزشی بسیار زیاد دارند و همچنین ممکن است در برخی موارد به خاطر اندازه بزرگ مدل، به چالش‌هایی در زمینه حافظه و محاسباتی برخورد کنند. این مدل‌ها معمولا در ابتدا بر روی یک دسته بسیار بزرگ از دادگان آموزشی پیش‌آموزش [۴] می‌شوند و سپس برای یک وظیفه خاص تنظیم دقیق می‌شوند تا بتوانند به بالاترین دقت ممکن برسند. [۵]

طبقه‌بندی متن‌های واقعی از ساختگی: استفاده از مدل‌های بزرگ زبان (LLMs) برای تولید محتوا در کانال‌های مختلف مانند اخبار، رسانه‌های اجتماعی، انجمن‌های پاسخ به سوالات، و حتی موارد علمی روز به روز در حال افزایش است. مدل‌های پیشرفته‌ای مانند ChatGPT و GPT-4، می‌توانند پاسخ‌های بسیار روانی را به انواع متفاوتی از پرسش‌های کاربر ایجاد کنند. این پیشرفت چشمگیر در حوزه مدل‌های بزرگ زبانی، آن‌ها را برای جایگزینی نیروی انسانی در بسیاری از سناریوها جذاب می‌کند. با این حال، این امر منجر به نگرانی‌هایی در مورد سوء استفاده احتمالی از آن‌ها، مانند انتشار اطلاعات غلط و ایجاد اختلال در سیستم آموزشی شده است. از آنجایی که تشخیص متون تولید شده توسط این دسته از مدل‌ها برای انسان‌ها کار سختی به حساب می‌آید، نیاز به توسعه سیستم‌های خودکار برای شناسایی متن تولید شده توسط ماشین وجود دارد.

در این پروژه ما به دنبال آموزش یک مدل یادگیری عمیق برای حل مسئله طبقه‌بندی متن هستیم تا تشخیص دهیم که متن ورودی توسط یک انسان و یا توسط ماشین نگارش شده است. اما همانطور که پیشتر اشاره شد، میخواهیم از تکنیک یادگیری نیمه‌نظارتی استفاده کرده و شرایطی را شبیه‌سازی نماییم که تنها برچسب‌های مربوط به بخشی از دادگان را در اختیار داریم. سپس تلاش خواهیم کرد تا با استفاده از ایده شبکه‌های مولد تخاصمی [۶] دقت خروجی‌های بدست آمده از مدل را افزایش دهیم. شما در ابتدا به پیاده‌سازی یک مدل پایه برای انجام وظیفه طبقه‌بندی خواهید پرداخت. سپس مدل خود را با استفاده از تکنیک یادگیری نیمه‌نظارتی آموزش داده و به بررسی عملکرد مدل می‌پردازید. در ادامه، با

---

[۱] Unsupervised Learning
[۲] Semi-Supervised Learning
[۳] Reinforcement Learning
[۴] Pre-Training
[۵] Fine-Tunning
[۶] Generative Adversarial Networks (GANs)

طراحی یک مدل GAN و استفاده از آن در مدل معماری مدل پایه [۷] ، تلاش می‌کنید تا دقت مدل اولیه خود را افزایش دهید. مقاله GAN-BERT [۱] که مرجع اصلی در این پروژه می‌باشد، از طریق این لینک قابل دسترس است.
پیشنهاد می‌شود پیش از شروع به انجام پروژه، این مقاله را به دقت مطالعه نمایید.

معرفی دیتاست

با رشد مدل‌های زبانی بزرگ، متن‌های تولیدی توسط آنها بسیار با کیفیت شده‌اند به گونه‌ای که تشخیص آنها برای انسان‌ها دشوار و عملکردی تقریبا مشابه تصادفی برای تشخیص متون دارند. این مدل‌ها مزایای بسیار زیادی دارند اما در بعضی مواقع مشکل‌زا نیز هستند. در نتیجه نیاز به یک سیستم برای تشخیص متون تولیدی توسط مدل‌های زبانی و انسان احساس می‌شود. مجموعه دادگان شامل ۶ کلاس Human، ChatGPT، Cohere، Davinci، BloomZ، Dolly، و هدف طبقه‌بندی دادگان بر اساس متن ورودی به هر یک از این کلاس‌ها می‌باشد.
در این پروژه می‌بایست درصد کوچکی از دادگان موجود را به عنوان دادگان برچسب دار و بقیه دادگان را بدون برچسب در نظر گرفته و شبکه مورد نظر را آموزش دهید. برای آزمایشات خود می‌توانید در چندین مرحله از ۱۰۰ نمونه داده به ازای هر کلاس شروع کرده و به تدریج تعداد نمونه‌ها را در آزمایشات بعدی اضافه کنید.
ما Subtask B از SemEval-2024 Task 8 را در نظر می‌گیریم و دیتاست از طریق این لینک قابل دسترس است.

معرفی GAN-BERT

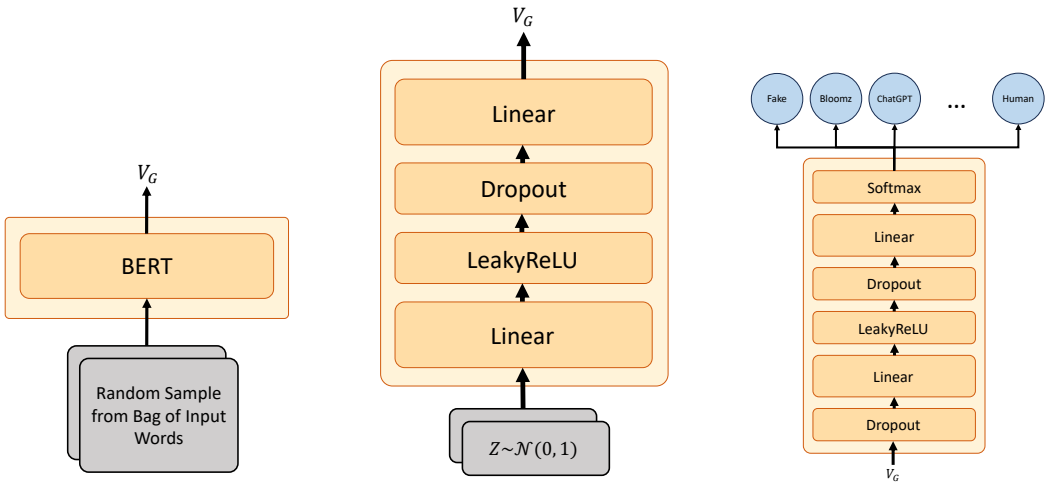
ساختار GAN-BERT مبتنی بر شبکه‌های مولد تخاصمی (GAN) بنا شده است و در طراحی آن از مدل BERT استفاده می‌شود که برای مسائل پردازش زبان‌های طبیعی (NLP) مناسب است. در این پروژه می‌خواهیم که از این ساختار برای مسئله طبقه‌بندی متن به $K = ۶$ کلاس استفاده کنیم. این ساختار GAN از یک شبکه تمییزدهنده D و یک شبکه مولد G تشکیل می‌شود. شبکه تمییزدهنده D آموزش داده می‌شود تا متن ورودی را به $K + ۱$ کلاس طبقه‌بندی کند که این شامل K کلاس برای نمونه‌های واقعی موجود در دیتاست و همچنین یک کلاس برای نمونه‌های جعلی است. از طرفی شبکه مولد G با این هدف آموزش داده می‌شود که نمونه‌های جعلی تولیدشده توسط آن توزیع آماری مشابهی با نمونه‌های واقعی موجود در دیتاست داشته باشند.
در ادامه جزئیات بیشتر این ساختار را معرفی می‌کنیم و توضیح می‌دهیم که چگونه از مدل BERT برای مسئله طبقه‌بندی استفاده می‌شود. فرض کنید که X_i یک نمونه واقعی از دیتاست باشد. ممکن است که برای این نمونه یک برچسب y_i در دسترس باشد که تعلق آن به K کلاس را مشخص می‌سازد. ما فرض می‌کنیم که تنها بخشی از نمونه‌های دیتاست آموزش برچسب‌دار هستند. پس از انجام پردازش‌هایی نظیر Padding و Truncation که توسط Tokenizer انجام می‌شوند، از مدل BERT برای محاسبه نمایش برداری نمونه X_i استفاده می‌کنیم. بطور مشخص، تنها یک بردار $v_B \in \mathbb{R}^{۷۶۸}$ را از مدل BERT به عنوان خروجی در نظر می‌گیریم که حالت مخفی متناظر با توکن CLS است.
نمایش برداری v_B به شبکه تمییزدهنده D داده می‌شود تا طبقه‌بندی انجام شود. همانطور که در شکل ۲آ نشان داده شده است، این شبکه معماری feed-forward دارد و شامل یک لایه مخفی خطی با اندازه ورودی و خروجی ۷۶۸ است. در طراحی آن از Dropout و تابع فعالسازی LeakyReLU استفاده شده است. در خروجی شبکه یک لایه خطی دیگر داریم که اندازه خروجی آن $K + ۱$ است. خروجی نهایی این شبکه را با $\hat{y}_i$ نشان می‌دهیم که برای آن بیشترین احتمال در خروجی SoftMax حاصل شده است.
پس از مشاهده یک ورودی نویزی، شبکه مولد G بردار $v_G \in \mathbb{R}^{۷۶۸}$ را در خروجی تولید می‌کند و سعی دارد که نمایش برداری v_B از مدل BERT به ازای نمونه واقعی X_i را تقلید کند. بردارهای v_B و v_G به شبکه تمییزدهنده D ورودی داده می‌شوند تا به $K + ۱$ کلاس طبقه‌بندی شوند.
در ادامه به معرفی معماری شبکه مولد می‌پردازیم. دو معماری $G_۱$ و $G_۲$ را برای شبکه مولد معرفی می‌کنیم.

- شبکه مولد $G_۱$ که در شکل ۲ب معرفی شده است، یک معماری feed-forward دارد و از دو لایه خطی تشکیل شده است. در طراحی آن از Dropout و تابع فعالسازی LeakyReLU استفاده شده است. اولین لایه خطی آن با ابعاد ورودی ۱۰۰ و خروجی ۷۶۸ و دومین لایه خطی آن با ابعاد ورودی و خروجی ۷۶۸ هستند. بردار نویزی $z \in \mathbb{R}^{۱۰۰}$ که مؤلفه‌های آن با توزیع گوسی استاندارد تولید شده اند به شبکه داده می‌شود تا بردار $v_G \in \mathbb{R}^{۷۶۸}$ محاسبه شود. به ازای مدل مولد $G_۱$ ساختار GAN-BERT در شکل ۳ خلاصه شده است.

[۷] Baseline

- همانطور که در شکل ۲ ج نشان داده شده است، شبکه مولد $G_۲$ از یک مدل BERT از قبل آموزش‌دیده استفاده می‌کند. لازم به ذکر است که این مدل BERT که در معماری شبکه $G_۲$ قرار دارد، مستقل از مدل BERT قبلاً معرفی شده است که نمونه‌های واقعی دیتاست را پردازش می‌کرد. ورودی نویزی در شبکه $G_۲$ به شکل متفاوتی در نظر گرفته می‌شود. ابتدا یک Bag of Words ساخته می‌شود که شامل کلمات حاضر در نمونه‌های آموزشی است. با توجه به فراوانی کلمات در نمونه‌های آموزشی، نمونه تصادفی $\tilde{X}$ تولید می‌شود. کلمات $\tilde{X}$ به صورت مستقل و با توزیع یکسان انتخاب می‌شوند. نمایش برداری $v_G \in \mathbb{R}^{۷۶۸}$ توسط مدل BERT محاسبه می‌شود.

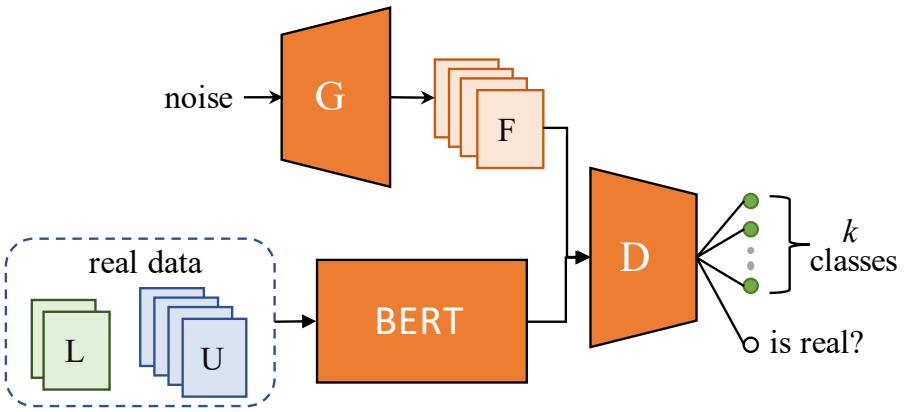
در ادامه نحوه آموزش شبکه‌ها به صورت نیمه‌نظارتی را توضیح می‌دهیم. تابع هزینه شبکه تمییزدهنده D به صورت $\mathcal{L}_D = \mathcal{L}_{D,\text{sup.}} + \mathcal{L}_{D,\text{unsup.}}$ قابل بیان است که عبارت $\mathcal{L}_{D,\text{sup.}}$ به هزینه ناشی از طبقه‌بندی نمونه‌های واقعی به K کلاس دیتاست اشاره دارد. این درحالی است که عبارت $\mathcal{L}_{D,\text{unsup.}}$ به هزینه GAN در طبقه‌بندی اشتباه نمونه‌های واقعی به عنوان جعلی و همچنین طبقه‌بندی اشتباه نمونه‌های جعلی به عنوان واقعی اشاره دارد. برای نمونه‌های آموزشی بدون برچسب، صرفاً عبارت $\mathcal{L}_{D,\text{unsup.}}$ محاسبه می‌شود. از طرفی، تابع هزینه شبکه مولد G به صورت $\mathcal{L}_G = \mathcal{L}_{G,\text{feat.}} + \mathcal{L}_{G,\text{unsup.}}$ تعریف می‌شود. عبارت $\mathcal{L}_{G,\text{unsup.}}$ به هزینه GAN در تشخیص درست نمونه‌های جعلی توسط شبکه D اشاره دارد و تابع هزینه انطباق ویژگی [۸] به صورت $\mathcal{L}_{G,\text{feat.}} = \frac{۱}{۷۶۸}\|\mathbb{E}_{x\sim p_{\text{data}}} f(x) - \mathbb{E}_{x\sim p_G} f(x)\|_۲^۲$ تعریف می‌شود که در آن $f(x)$ فعالیت لایه‌ی مخفی در شبکه تمییزدهنده D است.
لازم به ذکر است که پارامترهای مدل BERT که نمونه‌های واقعی دیتاست را پردازش می‌کرد و همچنین پارامترهای شبکه D برای حداقل‌سازی تابع هزینه $\mathcal{L}_D$ بروزرسانی می‌شوند. از طرفی، پارامترهای شبکه مولد G برای حداقل‌سازی تابع هزینه $\mathcal{L}_G$ بروزرسانی می‌شوند. بعد از آموزش شبکه‌ها، شبکه مولد G کنار گذاشته می‌شود.



(آ) شبکه D (ب) شبکه $G_۱$ (ج) شبکه $G_۲$

شکل ۲: معماری شبکه‌ها

[۸] feature matching



شکل ۳: ساختار GAN-BERT که در آن معماری G_1 فرض شده است. شبکه مولد یک مجموعه از نمونه‌های جعلی را بصورت تصادفی تولید می‌کند. دیتاست شامل نمونه‌های برچسب‌دار L و نمونه‌های بدون برچسب U است. نمایش برداری این نمونه‌های واقعی توسط مدل BERT محاسبه می‌شود. این مقادیر و مقادیری که شبکه مولد تولید کرده است، به شبکه تمییزدهنده D ورودی داده می‌شوند. شکل از مقاله [۱] آورده شده است.

خواسته‌های پروژه

۱. ابتدا مدل BERT را پیاده‌سازی کنید. شما می‌توانید از مدل‌های Pre-trained در سایت Hugging Face استفاده کنید. مدل BERT را Fine-tune کنید و دقت طبقه‌بندی را گزارش کنید. در تمامی خواسته‌های پروژه نمودارهایی شبیه مقاله GAN-BERT [۱] ارائه دهید که در آن تعداد نمونه‌های برچسب‌دار در دیتاست متغیر در نظر گرفته می‌شود. مثلاً می‌توانید ۱، ۵، ۱۰ و ۵۰ درصد از داده‌های آموزشی را بصورت برچسب‌دار در نظر بگیرید. چه نتیجه‌ای می‌گیرید؟

۲. (امتیازی) شما می‌تواند مدل خود را اصلاح کنید تا فرآیند آموزش سریع‌تر و قوی‌تر شود. برای مثال می‌توانید از Adapter ها [۲] استفاده کنید تا تعداد پارامترهای آموزش‌پذیر را کاهش دهید.

۳. حالا ساختار GAN-BERT را پیاده‌سازی کنید. دو معماری G_1 و G_2 که معرفی گردید را پیاده‌سازی کنید. دقت طبقه‌بندی را با بخش اول مقایسه کنید.

۴. (امتیازی) با نوآوری به دقت‌های بالاتر از ۸۰ درصد در طبقه‌بندی نمونه‌های ارزیابی برسید. مثلاً می‌توانید معماری شبکه مولد را بهبود دهید. نمره این بخش بصورت رقابتی بین دانشجویان در نظر گرفته می‌شود.

References

[1] D. Croce, G. Castellucci, and R. Basili, “GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2114–2119, Association for Computational Linguistics, Jul 2020.

[2] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, “Parameter-efficient transfer learning for NLP,” in International Conference on Machine Learning, pp. 2790–2799, PMLR, 2019.