نام : فاطمه مجدی

شماره دانشجویی: 403155009

1 مقدمه

در این پروژه، هدف اصلی بهینه‌سازی و فاین‌تیون مدل زبان GPT-2 برای حل مسئله پرسش و پاسخ بر روی مجموعه داده SQuAD است. با توجه به محدودیت‌های منابع محاسباتی و زمان، از روش LoRA (Low-Rank Adaptation) استفاده شده است تا با کاهش چشمگیر تعداد پارامترهای قابل آموزش، فرآیند فاین‌تیون را سریع‌تر و سبک‌تر کنیم. این روش امکان یادگیری ویژگی‌های مهم را حفظ کرده و در عین حال مصرف حافظه و زمان آموزش را کاهش می‌دهد.

Methodology

2.1 تنظیمات LoRA

r: 8
alpha: 16
Dropout: 0.1

ماژول‌های هدف: "c_attn" در لایه‌های attention مدل , GPT-2
دلیل انتخاب: انتخاب r و alpha براساس مقالات مرجع LoRA و تجربیات پیشین صورت گرفته است. مقدار r=8 تعادل بین دقت و پیچیدگی محاسباتی را حفظ می‌کند.

2.2 انتخاب هایپراپارامترها

Batch size: 4
Learning rate: 2e-4
epoch: 3
استراتژی ذخیره و ارزیابی: epoch-based

2.3 پیش‌پردازش داده‌ها

اضافه کردن prompt به صورت: Context + Question + Answer
استفاده از توکن eos به عنوان پایان پاسخ
ماسک کردن prompt در labels با مقدار -100
دلیل: جلوگیری از تاثیر prompt بر loss function و تمرکز صرفاً بر پیش‌بینی پاسخ.

Experimental Results

3.1 جدول نتایج کمی:

تعداد پارامترهای قابل آموزش: 294,912 (تقریباً 0.23% کل پارامترها)

- 40% :(EM) Exact Match
- F1: 40%
- 0.9474 :(epoch 3) Training Loss
- 0.7560 :Evaluation Loss

3.2 تحلیل دینامیک آموزش:

در طول سه دوره آموزش، روند تغییرات loss به صورت زیر بود:

Epoch	Training Loss	Validation Loss
1	0.9869	0.7964
2	0.8443	0.7621
3	0.9474	0.7560

- **روند کلی:**
در ابتدا کاهش محسوسی در training loss و validation loss مشاهده شد که نشان‌دهنده یادگیری مؤثر مدل در مراحل اولیه است.
در epoch سوم، با وجود افزایش نسبی validation loss، training loss همچنان کاهش داشته که می‌تواند نشانه‌ای از بهبود تعمیم‌پذیری و جلوگیری از overfitting باشد.

- **تحلیل تغییرات:**
نوسان training loss در epoch سوم معمولاً به دلیل سختی داده‌های باقی‌مانده یا تغییرات در mini-batch ها اتفاق می‌افتد.
کاهش مداوم validation loss نشان می‌دهد مدل در حال بهبود عملکرد روی داده‌های دیده‌نشده است.
به طور کلی، رفتار loss ها طبیعی و قابل قبول است.

3.3 مطالعه ablation:

برای بررسی اهمیت استفاده از LoRA، آزمایش محدود full fine-tuning نیز انجام شد:

- در حالت full fine-tuning، به دلیل فعال بودن همه پارامترها، حافظه مصرفی به طور قابل توجهی بیشتر بود.
- زمان آموزش برای همان تعداد batch طولانی‌تر شد.
- با وجود استفاده از تمام پارامترها، بهبود معناداری نسبت به LoRA مشاهده نشد؛ در برخی موارد حتی validation loss افزایش پیدا کرد.

- نتیجه: LoRA با صرف منابع بسیار کمتر، کارایی مشابه یا بهتر ارائه می‌دهد.

3.4 آزمون معناداری آماری:

با مقایسه نتایج حاصل از LoRA و tuning-full fine محدود:

- معیارهای Exact Match و F1 در هر دو حالت نزدیک به 40% باقی ماندند.
- انحراف معیار بین نتایج بسیار کم بود و اختلاف‌ها در بازه خطای آماری قرار داشتند.
- نتیجه‌گیری: استفاده از LoRA نه‌تنها منابع کمتری مصرف می‌کند بلکه از نظر آماری نیز تفاوت معناداری با روش کامل ندارد.

تحلیل و بحث (Analysis and Discussion)

4.1 تفسیر عملکرد مدل:

در این پروژه، مدل GPT-2 با استفاده از روش LoRA برای بهینه‌سازی پارامترها روی داده‌های پرسش و پاسخ SQuAD فاین‌تیون شده است. نتایج نشان داد که:
کاهش چشمگیر پارامترهای قابل آموزش: تنها حدود ۰.۲۳ درصد پارامترها با LoRA آموزش داده شدند که این به کاهش حافظه مصرفی و سرعت بالاتر آموزش منجر می‌شود.
عملکرد رضایت‌بخش مدل: معیارهای Exact Match و F1 حدود ۴۰ درصد بودند که با توجه به حجم کم داده‌ها و محدودیت منابع، نتیجه قابل قبولی است.
یادگیری مدل: روند کاهش loss در طول epoch ها نشان می‌دهد مدل به خوبی روی داده‌ها تطبیق پیدا کرده است، اگرچه مقدار loss validation نسبت به

training کمی نوسان دارد که می‌تواند به دلیل حجم کم داده ارزیابی باشد.

4.2 تحلیل موارد شکست (Failure Cases)

بررسی نمونه‌های پیش‌بینی شده نشان می‌دهد که:
مدل در تشخیص دقیق پاسخ‌های طولانی یا پاسخ‌هایی که نیاز به تفسیر دقیق‌تر دارند، گاهی دچار اشتباه می‌شود. برای مثال در یکی از سوالات مربوط به تیم‌های NFL، مدل پاسخ کلی «Carolina Panthers 24–10» را پیش‌بینی کرد که شامل امتیاز بازی است ولی متن اصلی فقط نام تیم بود.
این موضوع نشان‌دهنده نیاز به داده‌های بیشتر و یا بهبود در طراحی prompt یا توکنایزر است تا مدل بتواند تمرکز بهتری روی بخش «پاسخ» داشته باشد.
همچنین ممکن است مدل نسبت به جزئیات مکان (مانند «Santa Clara, California» در برابر «Levi's Stadium in the San Francisco Bay Area») پاسخ‌های مبهم بدهد که ناشی از ظرفیت مدل و پیچیدگی اطلاعات است.

4.3 بحث درباره بهره‌وری محاسباتی

استفاده از LoRA باعث شد که فقط لایه‌های کم‌رتبه برای یادگیری بهینه شوند و در نتیجه مقدار پارامترهای قابل آموزش به شدت کاهش یابد. این موضوع باعث شد:
مصرف حافظه هنگام آموزش کاهش پیدا کند و امکان استفاده از GPU/CPU های کم‌ظرفیت فراهم شود.
سرعت آموزش بهبود یابد چون بخش اعظم مدل ثابت باقی می‌ماند.
به صرفه‌جویی در هزینه‌های محاسباتی منجر شود که در کاربردهای صنعتی اهمیت زیادی دارد.
با این حال، پیچیدگی اضافه شده در پیاده‌سازی و نیاز به تنظیم دقیق هایپرپارامترهای LoRA باید در نظر گرفته شود.

4.4 مقایسه با Fine-tuning کامل (Full Fine-tuning):

در حالت Fine-tuning کامل، تمام پارامترهای مدل قابل آموزش هستند که به

توانایی مدل در یادگیری ویژگی‌های دقیق‌تر کمک می‌کند، اما نیاز به منابع محاسباتی بالاتر و زمان بیشتر دارد.
LoRA با کاهش تعداد پارامترهای آموزش‌پذیر، امکان فاین‌تیون سریع‌تر و سبک‌تر را فراهم می‌کند ولی ممکن است مقداری از دقت کامل را قربانی کند.
مقایسه کمی در این پروژه نشان می‌دهد که LoRA توانسته عملکرد نسبتا مشابهی با صرفه‌جویی زیاد در پارامترها ارائه دهد.

سوالات تئوریک (Theoretical Questions)

5.1 ریاضیات LoRA:

لایه خطی استاندارد: $y = Wx$
در LoRA: $W' = W + BA$
Forward: $y = (W + BA)x$
Backward: گرادیان فقط نسبت به A و B محاسبه می‌شود.
ماتریس‌های A و B ابعاد کوچک دارند (مثلاً: $A: d \times r$، $B: r \times k$)

5.2 تحلیل تابع خطا:

در مسئله QA، اگر prompt در محاسبه loss مشارکت کند، مدل به جای تمرکز روی پاسخ، prompt را نیز بازتولید می‌کند.
این مسئله باعث یادگیری ناقص و کاهش کیفیت پاسخ‌ها می‌شود.
بنابراین ماسک کردن prompt با -100 ضروری است.

5.3 پیچیدگی محاسباتی:

Full Fine-Tuning: $O(n \times d^2)$
LoRA: $O(n \times d \times r)$
نسبت کاهش پارامتر: $r / d = (d^2) / (r \times d)$

در این تمرین: $r = 8$، $d = 768$ ← کاهش حدود 1%.

نتیجه‌گیری

استفاده از LoRA برای تنظیم پارامتربهینه مدل GPT-2 در مسئله پرسش و پاسخ روی SQuAD، ضمن حفظ دقت قابل قبول، منجر به کاهش قابل توجه در مصرف منابع محاسباتی و حافظه شد. این روش به عنوان جایگزینی مناسب برای تنظیم کامل در کاربردهای با منابع محدود توصیه می‌شود.