專題報告

10/16

b1228006張鴻毅
b1228022梁釗豪
b1228029蘇琪文
b1228032張紋菱

背景:

由於之前的 train set有混合了test set
因此我們之前的GPT2和各種模型的pre-train和fine-tuning版本
都再做一次(訓練和)評分

評估:

prepare_data.py→ 切分資料集，統一訓練與測試集
generate.py → 把模型的「輸入/輸出」存起來
evaluate.py → 拿這些輸入輸出，透過 兩種方法 來衡量品質：

1. ROUGE → 檢查字面重疊（比較嚴格，偏向檢查摘要/關鍵字
2. BERTScore → 檢查語意相似（比較靈活，偏向語意理解）

GPT2模型比較
(pre-train)

- **distilgpt2**
- **gpt2-medium**
- **gpt2-large**
- **gpt2-xl**

GPT2預模型ROUGE 分數：

項目	distilgpt2	gpt2-medium	gpt2-large	gpt2-xl	字面重疊
rouge1	0.3185	0.1966	0.205	0.1988	單字重疊率
rouge2	0.2779	0.1639	0.1816	0.1746	雙字重疊率
rougeL	0.3171	0.1965	0.2042	0.1981	最長公共子序列
rougeLsum	0.3182	0.198	0.2051	0.1988	多句摘要

GPT2預模型BERTScore：

項目	distilgpt2	gpt2-medium	gpt2-large	gpt2-xl	語意相似度
Precision	0.8279	0.7988	0.8038	0.8009	生成的詞向量與參考相似度(語意正確?)
Recall	0.9326	0.9216	0.9208	0.9204	參考的詞向量與生成相似度(參考被涵蓋?)
F1	0.8766	0.8556	0.858	0.8562	Precision,Recall 調和平均(整體語意)

GPT2模型比較
(fine-tuning)

- train_gpt2.py
  (distilgpt2)
- GPT2_5%
- gpt2_ok.py
- -

GPT2訓練模型ROUGE 分數：

項目	train_gpt2.py(distilgpt2)	GPT2_5%	gpt2_ok.py	字面重疊
rouge1	0.1678	0.8134	跑	單字重疊率
rouge2	0.1467	0.687	不	雙字重疊率
rougeL	0.1678	0.8124	到	最長公共子序列
rougeLsum	0.168	0.8158	T_T	多句摘要

GPT2訓練模型BERTScore：

項目	train_gpt2.py(distil gpt2)	GPT2_5%	gpt2_ok.py	語意相似度
Precision	0.7824	0.9515	跑	生成的詞向量與參考相似度(語意正確?)
Recall	0.9248	0.9767	不	參考的詞向量與生成相似度(參考被涵蓋?)
F1	0.8475	0.9635	到	Precision,Recall 調和平均(整體語意)

GPT2模型比較(all)

ROUGE 分數：

	Pre-train				Fine-tuning		
項目	distilgpt2	gpt2-medium	gpt2-large	gpt2-xl	train_gpt2.py (distilgpt2)	GPT2_5%	gpt2_ok.py
rouge1	0.3185	0.1966	0.205	0.1988	0.1678	0.8134	跑
rouge2	0.2779	0.1639	0.1816	0.1746	0.1467	0.687	不
rougeL	0.3171	0.1965	0.2042	0.1981	0.1678	0.8124	到
rougeLsum	0.3182	0.198	0.2051	0.1988	0.168	0.8158	T_T

BERTScore：

	Pre-train				Fine-tuning		
項目	distilgpt2	gpt2-medium	gpt2-large	gpt2-xl	train_gpt2.py (distilgpt2)	GPT2_5%	gpt2_ok.py
Precision	0.8279	0.7988	0.8038	0.8009	0.7824	0.9515	跑
Recall	0.9326	0.9216	0.9208	0.9204	0.9248	0.9767	不
F1	0.8766	0.8556	0.858	0.8562	0.8475	0.9635	到

各種模型比較
(pre-train)

- GPT2
- TinyLlama-1.1B
- Qwen3-0.6B
- Opt-350m(facebook)

BERTScore：

項目	GPT2(distilgpt2)	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m	語意相似度
Precision	0.8279	0.8073	0.7935	0.8323	生成的詞向量與參考相似度(語意正確?)
Recall	0.9326	0.9235	0.8649	0.9336	參考的詞向量與生成相似度(參考被涵蓋?)
F1	0.8766	0.8613	0.8271	0.8797	Precision,Recall 調和平均(整體語意)

ROUGE 分數：

項目	GPT2(distilgpt2)	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m	字面重疊
rouge1	0.3185	0.2109	0.1649	0.2761	單字重疊率
rouge2	0.2779	0.1844	0.1224	0.2419	雙字重疊率
rougeL	0.3171	0.2119	0.1636	0.274	最長公共子序列
rougeLsum	0.3182	0.2123	0.163	0.2741	多句摘要

各種模型比較
(fine-tuning)

- GPT2
- TinyLlama-1.1B
- Qwen3-0.6B
- Opt-350m(facebook)

BERTScore：

項目	GPT2_5%	TinyLlama-1.1B	Qwen3-0.6B	語意相似度
Precision	0.9515	0.7929	0.7867	生成的詞向量與參考相似度(語意正確?)
Recall	0.9767	0.9227	0.9087	參考的詞向量與生成相似度(參考被涵蓋?)
F1	0.9635	0.8527	0.8431	Precision,Recall 調和平均(整體語意)

ROUGE 分數：

項目	GPT2_5%	TinyLlama-1.1B	Qwen3-0.6B	字面重疊
rouge1	0.8134	0.1972	0.1485	單字重疊率
rouge2	0.687	0.1701	0.13	雙字重疊率
rougeL	0.8124	0.1963	0.1489	最長公共子序列
rougeLsum	0.8158	0.197	0.1479	多句摘要

各種模型比較
(all)

- GPT2
- TinyLlama-1.1B
- Qwen3-0.6B
- Opt-350m(facebook)

BERTScore：

	Pre-train				Fine-tuning		
項目	GPT2 (distilgpt2)	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m	GPT2_5%	TinyLlama-1.1B	Qwen3-0.6B
Precision	0.8279	0.8073	0.7935	0.8323	0.9515	0.7929	0.7867
Recall	0.9326	0.9235	0.8649	0.9336	0.9767	0.9227	0.9087
F1	0.8766	0.8613	0.8271	0.8797	0.9635	0.8527	0.8431

ROUGE 分數：

	Pre-train				Fine-tuning		
項目	GPT2 (distilgpt2)	TinyLlama-1.1B	Qwen3-0.6B	Opt-350m	GPT2_5%	TinyLlama-1.1B	Qwen3-0.6B
rouge1	0.3185	0.2109	0.1649	0.2761	0.8134	0.1972	0.1485
rouge2	0.2779	0.1844	0.1224	0.2419	0.687	0.1701	0.13
rougeL	0.3171	0.2119	0.1636	0.274	0.8124	0.1963	0.1489
rougeLsum	0.3182	0.2123	0.163	0.2741	0.8158	0.197	0.1479