Information Retrieval with Contrastive Learning

李吉昌

r08922a27
r08922a27@ntu.edu.tw
Writing Report, Contrastive Model Presentation

黃品硯

r09922a04
r09922a04@ntu.edu.tw
Writing Report, Preprocessing Data QA Model

余友竹

r09922104
r09922104@ntu.edu.tw
Writing Report, Preprocessing Data Model Prediction

Abstract

傳統 Vector Space Method 的方法萃取的是 word-level 的關係，然而在 Question-Answering 的任務中，我們的檢索目標為提供模型推論答案的依據，除了詞共同出現的事件，我們更希望檢索的結果具備更高層次的抽象化資訊，Contrastive Learning 目標為學習一表徵能夠鑑別不同指定抽樣物件的能力，和檢索任務類似，我們期望模型得出的表徵能夠分辨出對於 query 問題的語句 relevant 和 non-relevant 的內容。在這次嘗試題目裡，我們假設出自同一篇文章的語句間具備部分共同的結構性資訊，並且認為這樣的相似性可以類推至 query 問題以及提供推斷內容中，我們在接下來將設計不同的 pair scheme 來訓練 contrastive model，並且在實驗的內容中，我們針對這次任務做了一系列檢索表現的分析，除了包含針對 ground-truth 的檢索探討，我們也利用檢索的內容作為問答模型的輸入，檢驗 query 問題和檢索的內容之間語意上的相似程度，並針對結果討論導致結果的各種可能性。
此外，我們提供這次 final project 的程式碼專案以供複現結果，連結如下：`https://github.com/PM25/Information-Retrieval-with-Contrastive-Learning`

1. INTRODUCTION

在資訊檢索的領域中，Vector Space Model(VSM) 為主流方法之一，以 Term Frequency - Inverse Document Frequency (TF-IDF) 改進延伸的方法在不同的資料上得到了穩定且良好的表現結果，然而，詞彙與詞彙之間的關係以及 corpus 中未知詞彙和 smoothing 的問題仍然是一大挑戰，並且在隨著資料量級的成長，TF-IDF 本身的計算複雜度也會隨之上升，在不同資料量級上的 scalability 受限於方法計算上的限制。在人為設計的特徵抽取中，鮮少能夠得到更抽象且語意上的資訊，我們希望模型接受到的特徵除了帶有詞本身的資訊外，也能夠考慮整體語句結構性的資訊在，進而能夠達到語意上的理解，解決需要推論的進階問題。在著名的 QA benchmarck Fact Extraction and VERification (FEVER)[6] 中，題目會從一串維基百科中相關主題的文章語句(evidences)當中歸納出一 是非問題(claim)，實際預測時只會給定 claim，而 evidences 必須從維基百科中的內容查找，因此，共可分為「檢索」以及從檢索的內容中「推斷」答案，其問題和查找的內容必須具備語意上的相似性，甚至是上下文關係才能使問答模型作出正確的決策，我們會希望第一階段的檢索中，傳統 VSM 的檢索我們可以利用 term 本身出現頻率和不同文章出現的頻率找到 word-level 上相似的文章，但撇除 word-level 自身的內容，我們期望的事情是找到「語意相近」的相關內容，該內容不一定需要字詞之間的高重疊率，而是語意理解上的相似，因此我們發想時想到了一個問題：「在檢索任務中，我們能不能萃取出整句話包含的語意資訊達到檢索目的呢？」

判斷詞彙語意相似性最傳統的作法是 distributed word embedding 的方法，glove 和 fasttext 為傳統學習模型中較具有代表性的方法，假設一句話的不同詞之間具有結構性的資訊，利用 CBOW 和 skip-gram 的訓練方式能夠使模型透過結構上的資訊還原出被 masking 的詞彙，而近年來 transformer based method 更是宰制了各式自然語言處理任務的 benchmark，基於以大量資料預訓練的 language model 得出的表徵具有各語言任務可共享資訊的假設下，fine-tune contextual embedding model 於各式語言任務上，在當代深度學習於自然語言處理的場景中成為了主流，除了不需要標籤資訊外，transformer 可平行化且可觀測全局的性質也是其中的優勢，其輸出具有考慮整句話結構性資訊，可做為一下游任務的輸入特徵。
Contrastive Learning 在自監督學習上，在預訓練領域得到很大的突破。主要任務為透過自訂義的規則，決定兩物件為相似關係的 positive pair，或非相似關係的 negative pair，使模型學習區分正反對表徵的能力。在這一篇報告中，我們預期訓練一 contrastive model 區分我們將語句視為一物件，利用 transformer 作為時序性特徵產生器作為我們 contrastive model 的輸入，以不同的 augmentation 方式來決定不同的 pair scheme，將整句語句壓縮成一拿來「比較」的抽象語句表徵，該表徵濃縮了整句話的結構性資訊，在一篇文章中，我們假設兩句非全部 stop-word 的語句共享了結構性的資訊，而 claim 和 evidences 之間存在類似的關聯性，則我們期望透過 augmentation 的方式能夠 query 的輸入，進而達到自監督學習「語意檢索」的目的，此外，透過 Contrastive Learning 得到的表徵有以下好處： (1) 降低特徵維度，得到 term 之間和語意結構上的潛在相似性。(2)能夠進行 self-supervised learning，不需要有 ground-truth labels，也能夠在文章中找出語意的相似性。(3) 文獻指出，contrastive loss 可以解釋為 mutual information 的一個下界[8]，在理論上透過 Contrastive Learning 得到的表徵對原資料具一定程度代表性。
在第一章前半段，我們敘述了目前 TF-IDF 現存方法上的侷限，因此我們期望能夠在語句上得到額外語意上的資訊，透過 Contrastive Learning，我們利用 augmentation 模擬 query(claim) sentence 的方式達到自監督學習「語意檢索」的目的；在第二章我們將概略介紹 BERT 和 Roberta 兩個知名 contextual embedding model 和 Contrastive Learning 相關的作法與各式變形以及在 text 上的應用；第三章的部份將詳細描述整個訓練以及預測流程，考量到效率問題以及資料的穩定下採用了特定的前處理方式，並且在 contrastive model 中，我們也會說明我們的 loss 和模型架構設計；第四章會詳細記錄了訓練上的設定以及我們評審方法的方式及其動機；我們在最後第五章總結我們這次嘗試的貢獻以及未來展望。

2. RELATED WORK

近年來，transformer 相關模型不斷被提出，BERT[2] 作為第一個 work 問世後，現今 transformer 相關方法已主宰各自然語言任務，RoBERTa[5] 提供了更大架構更多資料的嘗試，並且改進了原 BERT[2] 預訓練方式，在各下游任務中達到更好的表現。使用 BERT 模型進行下游任務時通常需要針對資料集做 fine-tuning 的步驟，通常需要進行資料標記才能得到更好的表現。在我們的任務中，我們將 BERT[2] 作為一 contextual feature extractor，進一步訓練 contrastive model，判斷語句間的相關性，我們的方法可以 unsupervised 的學習，進一步延伸了 BERT 模型的使用方式。在 Sec. 4.3 會需要一問答模型去評審檢索效果的好壞，我們選擇 RoBERTa[5] 作為其基礎架構。

Contrastive Learning 在計算機視覺以及語音等多媒體領域也得到廣泛的發展，在自監督學習的領域上作為預訓練模型能夠得到 data efficiency、performance improvement 的效果，透過「比較」不同 sample 的方式得以學到一理想表徵，其表徵盡可能保有原資料的重要資訊，在 InfoNCE[8] 中，作者證明 contrastive loss 為 mutual information 的一個下界。在 SimCLR[1] 和 CMC[7] 的論文中提出了更多實驗性發現，我們也從中知道 batch size 和 augmentation 的策略會是影響成效的重要因素。在 MoCo[4] 的論文中提出了 momentum encoder 的作法，為 contrastive learning 提供一增加 negative samples 的方式，進而達到增加 batch size 的效果，提升預訓練的成績。除此之外，在語者驗證領域，[9] 利用 contrastive learning 的作法來得到具有驗證能力的表徵。在文字處理的領域，[3] 嘗試結合 transformer 和 Contrastive Learning，在不同的自然語言任務上得到不錯的表現。 啓發於 [9, 3]，我們認為語句檢索的任務亦可視為是一種驗證的延伸，而文字上的資料投入 contrastive loss 被預期能夠自監督式達到檢索效果。

3. METHODOLOGY

我們的架構主要分成三個部分，第一部分是對資料庫裡的文件進行預處理，並事先訓練好整個資料集的 TF-IDF vectorizer。第二部分是將預處理完成的文件，以句子為單位，透過我們制定的規則，抽取出與該句語意相近的句子，訓練 contrastive model 學會判斷句子之間的相似程度。第三部分是根據我們訓練出的模型，使用 VSM 的方式，將句子轉換成 feature vector，透過計算相似度，決定出合適的 evidences，並根據需求，進行不同的分析。其中第三部分的feature extraction, evaluation 涉及到實驗細節，我們會在 Sec. 4 做詳細的介紹。

3.1 Data Preprocessing

我們首先會對整個 corpus (collection) 進行解析、預處理，其中包含三個步驟。第一個步驟是將整個documents set 句子做斷句，並消除特殊符號。第二個步驟是透過 FEVER 官方提供的 TF-IDF builder 建構出 inverted file。第三個步驟則是將 documents set 中處理好的句子 全部透過 TF-IDF 模型事先得出 TF-IDF 的 feature vectors，以方便後續計算相似度。

3.2 Contrastive Model

我們假設出自同一篇文章的兩句話共享了一定程度資訊，有一定的關聯性，我們的目標在於嘗試該關聯性是否能夠推展至 claim-evidence 之間的關係，期望透過嘗試不同 augmentation 策略來模擬 claim-evidence 的關係，我們的 augmentation 策略分為兩種，其一為 uniform sampling (Uniform-CL)，將同一篇文章內的隨機兩個句子作為 positive pair，不同文章的兩句話為 negative pair；其二為每句話取 TF-IDF 的 cosine similarity 前 10% 的另一句話作為 positive pair，不同文章的兩句話為 negative pair (TFIDF-CL)。我們的 contrastive loss 參考 SimCLR[1] 的框架，在同一個 batch 中不同 sample 皆來抽樣自不同的文章，因此可作為彼此的 negative pair，我們的整體流程圖如 Fig. 1 所示。

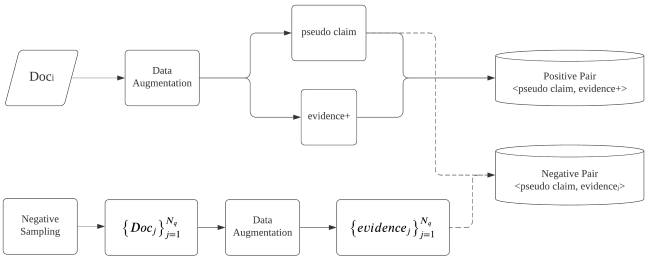

Fig. 1. flowchart

除此之外，因為 contrastive model 參數量過大，計算成本非常高，我們採取混合原 baseline 的 TF-IDF 檢索方法初步降低需要檢索的文章數量，我們先利用 TF-IDF 檢索出前 100 個和 claim 相似內容的語句，其後 contrastive model 將其 100 個候選文章做排序，取前 15 個相似 evidence 來做評審。

4. EXPERIMENTS

我們將初步在 Sec.4.1 介紹採用的訓練、開發資料集以及提供訓練過程的詳細設定；接下來我們會比較傳統方法 Elastic Search (TF-IDF) 和基於 Sec.3.2 中兩種不同 augmentation 策略決定的 pair scheme 訓練出來的 contrastive model 共三種方法的檢索表現，在 Sec.4.2，我們設計了兩種評分，間接地說明不同方法決定出來的表徵在檢索上的表現，其一為在評量三種方法找到用 claim 作為 query 檢索到的 evidence 為真正給定 evidence 的比例，亦即 recall；其二為利用測試資料集所提供的 groundtruth，計算 claim 與 evidences 透過三種不同的方法轉換成 embedding feature vectors 後的 cosine similarity。這個估計分數可以視為是衡量我們所轉換的 representations 是否真的能夠表示 claim 跟 evidences 之間結構上的語意關係。除了將檢索結果比較自訂相似的程度，我們衡量的最終目標為將檢索結果作為 claim 提問的答案內容，我們在 Sec.4.3 比較三種方法作為 QA 模型輸入後推論答案的表現，若答對的程度越高，代表檢索結果中包含越多語意相關性。

4.1 Implementation details

在 contrastive model 的訓練中，基於不同的 augmentation 策略，我們使用相同的訓練設定，模型架構採三層 Bidirectional Long Short-Term Memory (BLSTM) 接一層 Linear Layer 以線性縮放，優化器採用 Adam，其學習率為 0.0025，betas 參數為 (0.9, 0.999)，batch size 為 256，總共訓練 5000 個 steps，temperature 為 0.05，hidden size 為 256，output size 為 128，模型的輸入為 based-bert-uncased 固定參數的輸出，因此 contrastive model 的 input size 為 768。

我們採用的資料集是 Fact Extraction and VERification (FEVER) [6] 是一個2018年公佈的競賽，其中包含185,445個從 Wikipedia 中抽取數句話所生成的 claims，每個 claims 可以被區分成支持 (Supported)、否決 (Refuted)，以及不相關 (NotEnoughInfo)。對於前兩種類別，一個 claim 會對應到數個 evidences 用以證明其屬於哪個類別，其 evidences 為 QA model 判別類別的依據，亦是我們檢索的結果，其 claim 的產生方式為人工依據一系列被蒐集好的相關文章的段落歸納出一個問題以供分類，因此實際上 claim 對應到的 evidence 是有 ground-truth 的，訓練和開發資料集皆有附上其資料，在接下

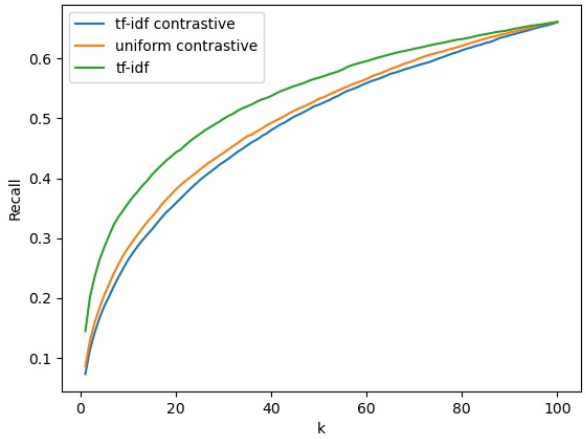

Fig. 2. 檢索數量與 Recall 間的關聯

來的章節，我們會利用該資料作 Intrinsic comparison，以及 Extrinsic comparison 的分析。

4.2 Intrinsic Comparison

Intrinsic Comparison 旨在衡量我們的模型萃取出的 embedding features 是否足以衡量不同語句間的語意相似度，在 Sec. *4.2.1* 我們選擇觀測不同檢索方法在不同檢索數量下，對原 ground-truth evidences 的 recall；在 Sec. *4.2.2* 我們已知原開發集的 ground-truth evidences，得以比較不同 claim 和 ground-truth evidences 的 cosine similarity 。

4.2.1 recall. 如 Fig. 2 所示，橫軸為前不同方法檢索出來的數量，縱軸為對應的 recall 數值，可以發現在數量越大時有明顯的邊際效應。

4.2.2 cosine similarity. 在各式的自監督學習的檢索方法中，是否萃取出有表徵表徵來衡量 claim 和 evidence 的關係是一個評估表現的重要指標。因此我們透過不同檢索模型轉換出來的不同表徵對應的 claims 與其 ground-truth evidences 相似性 (cosine similarity) 作為衡量依據。若相關性越高，表示學習出的表徵能夠很好的分辨具有推論關係的兩個句子之間的相關性，其結果顯示於 Tab. 1。從結果來看，我們使用 TF-IDF 策略所訓練出來的 Contrastive Model 在 ground-truth 上有最高的相似性。而 uniform sampling 的策略則是有最低的相似性。這解釋了三件事：(1) 使用不同的 augmentation 策略確實會對結果產生巨大的影響。(2) 傳統 TF-IDF 模型在比較 cosine similarity 時，會忽略沒有出現的字詞所帶來的語意關係。(3) 使用 TF-IDF 策略的 Contrastive Model 在這個指標上，似乎成功的抓取了一些潛在的語意關係，足以說明我們的模型成功之處。

4.3 Extrinsic Comparison

在這次嘗試中，我們需要驗證的是我們的檢索結果是否具備「語意」上的推論關係。在 FEVER Dataset 中，推論是否為支持/否決才是最重要的任務，假使一問答模型能夠基於檢索

Methods	TF-IDF	Uniform-CL	TFIDF-CL
Cos Similarity	0.022	-0.0081	0.4275

Table 1.

	Precison	Recall	F1
Ground Truth	0.95	0.95	0.95
TF-IDF	0.84	0.83	0.83
w/o Evidence	0.78	0.60	0.68
Uniform Sample	0.90	0.11	0.20
Uniform-CL	0.79	0.79	0.79
TFIDF-CL	0.80	0.80	0.80

Table 2.

出來的 evidences 得到更好的表現，則代表其 evidences 具備了足以提供推論的語意。 我們的問答模型選擇 RoBERTa[5] 作為預訓練模型，將 claim 和 evidences 用 separating token 相隔併入作為輸入字串，預測其標籤類別，優化器選擇 AdamW，學習率設為 0.00001，epochs 設為 3，batch size 設為 8，此外，為求訓練穩定，在訓練 5000 個 steps 之後才 finetune 預訓練模型。

在 Tab. 2 除了比較 Sec. 3 介紹的三種檢索方法，我們也額外嘗試使用空字串或是隨機抽樣不相關的語句作為 evidence，可以發現部分 claim 具有潛在答案的資訊，儘管沒有 evidence 仍然可以回答出正確的結果，而隨機抽樣反而會導致表現大幅度下降，由此可知 evidence 會嚴重影響表現，由上述結果可以知道實際上 contrastive model 的表現並不如預期，使用 baseline 做初步篩選後仍然得到較差的結果，我們認為主要因素是「兩句出自同一文章的語句擁有相似的結構性資訊」的假設未必成立，我們的 augmentation 製造出來的 pair 之間的關聯性和 claim-evidence 之間的關係存在一定差異，從 uniform 的 pair scheme 訓練出來的模型在 Tab. *4.2.2* 的實驗結果來看，其 contrastive model 預測到的 claim 和 ground-truth 之間的相似度是非常低的，並沒有抓到 claim-evidence 之間的關係，但是 TF-IDF 的 pair scheme 相對來說，在比較真正 claim-evidence 有很大的進步，代表 claim-evidence 之間存在 word-level 的關聯性，但我們這一次的嘗試在實驗上沒有跡象顯示我們 contrastive model 有得到除了 word-level 上更多的語意資訊，而找到一 augmentation 使得 pair 能夠更貼近 pair scheme 是我們這次未達到的挑戰。

5. CONCLUSION

在這一次的嘗試中，我們的目標設定在自監督上檢索出 word-level 之外，更抽象的語意資訊，進而改進問答模型的表現，我們分別設置了 uniform 和 TF-IDF 兩種 augmentation 的方法，企圖使模型學到 claim-evidence 之間的關聯性，並且自訂出 Recall 和 claim-evidence-cosine-similarity 兩種評分方式來間接觀察檢索的表現以及趨勢，並且最後投入問答模型，比較各個檢索方法的結果，從結果來看，我們的模型得到的表徵確實一定程度上反映了句子之間的語意關係。雖然我們的模型在 QA 部分的表現不甚理想，但我們認為這樣的模型有很大的可能特化於 QA 任務上做調整，以達到更好的表現。就我們所知，我們的 work 是第一個利用 Contrastive Learning 嘗試「自監督語意檢索」的實驗，除此之外，我們也訂立了一系列評量標準並闡述了使用的動機，在未來可能的嘗試中，我們認為 augmentation 是影響整體表現的關鍵，錯誤的優化方向會使得訓練和測試的目標產生偏差，我們在 TF-IDF 的 augmentation 方法中觀測到 cosine similarity 的改進，表示模型能透過適當的 augmentation 方式學習到 claim-evidence 的相關性的，在延伸方向上，我們認為可以投入更複雜的語句分析方式，例如語言剖析樹，利用語構的資訊讓 pair scheme 更貼近理想的情況。 這次的 project 讓我們更進一步驗證了 contrative model 在 documents retrieval 上的可塑性，我們認為這樣的應用是一值得延伸和探索的研究方向。

6. REFERENCES

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proc. ICML*, 2020.

[2] J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, 2019.

[3] John Michael Giorgi, Osvald Nitski, Gary D Bader, and Bo Wang. Declutr: Deep contrastive learning for unsupervised textual representations. *ArXiv*, abs/2006.03659, 2020.

[4] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proc. CVPR*, 2020.

[5] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692, 2019.

[6] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *NAACL-HLT*, 2018.

[7] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Proc. ECCV*, 2020.

[8] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding, 2019.

[9] Wei Xia, Chunlei Zhang, Chao Weng, Meng Yu, and Dong Yu. Self-supervised text-independent speaker verification using prototypical momentum contrastive learning. In *Proc. ICASSP*, 2021.