CS224u lit review

Lara Thompson
Principle Data Scientist @ Salesforce
lara.thompson@salesforce.com

March 4, 2023

1 General problem/task definition

For my final project, I'm interested in the intersection of language models and recommender systems. As alluded to in the course, recommender systems are similar retrieval systems albeit with a different aim: rather than searching for a few key results to answer a query, a recommender system has no query and should be less reliant on a current context (even if some information is given, eg. last book read, we still don't know what the reader is in the mood for next). Regardless, there is much to learn from information retriever systems such as ColBERT (Khattab and Zaharia, 2020). (Malkiel et al., 2020) take a BERT-based approach to pure text content recommendations. Hybrid recommender systems that leverate user-item and user-user graphs mostly use only very simple text encodings.

Sentence-BERT (Reimers and Gurevych, 2019) is a good paper to compare with ColBERT in their usage of pre-trained BERT and arguably is a better starting place for ColBERT's finetuning that BERT itself. "A simple but Tough-to-Beat Baseline for Sentence Embeddings" (Arora et al., 2017) is a very interesting word embeddings aggregation approach that may win if system performance is more important than the last few points of accuracy.

More recently, researchers from Alibaba propose M6-Rec...

2 Concise summaries of the articles

2.1 A Simple but Tough-to-Beat Baseline for Sentence Embeddings (Arora et al., 2017)

In contrast to Sentence-BERT, Arora et. al. from Princeton developed a very simple sentence embedding. They take a weighted average of the word embeddings (downweighting common words) then subtract out the projection to the principal component (as estimated across several sentences; again to remove the common discourse component).

To motivate their approach, they modify a random walk model of corpus generation. Rather than generating only words near a slowly changing discourse vector $c_s(t)$, they allow for the large deviations to common words ('the', 'and', etc) in two ways: first, by separating a common discourse component c_0 and, second, by allowing all words a chance to appear proportionally to their typical frequency $p(w)$. Note that c_s, c_0 and v_w, the word embedding of w all reside in the same embedding space. The original probability of observing a word w is:

$$\Pr[w(t)|c_s(t)] \propto \exp\left(\langle c_s(t), v_w \rangle\right) \tag{1}$$

After the two modifications for common words and common syntax:

$$\Pr[w(t)|c_s(t)] = \alpha p(w) + (1-\alpha)\frac{\exp\left(\langle \tilde{c}_s(t), v_w \rangle\right)}{Z_{\tilde{c}_s}} \tag{2}$$

where $\tilde{c}_s = \beta c_0 + (1-\beta)c_s$. The authors then go on to show how this relates to a $p(w)$-weighted embedding average that then has its component along the principle component $\sim c_0$.

They test their sentence embeddings across the STS benchmarks and compare against other sentence embedding approaches (not Sentence BERT since it came out later!). They find they beat many more complicated approaches in text similarity tasks but the RNN approaches outperform them on sentiment tasks. It seems a bag-of-words approach cannot capture the combined sentiment; in fact, this approach may specifically downweight various negation terms (eg. "not") because they are common.

2.2 Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (Reimers and Gurevych, 2019)

Reimers and Gurevych set out to train better sentence embeddings that the various word embeddings-based approaches for text similarity

tasks that nevertheless scaled better than BERT /RoBERTa which much run on every pair individually.

They use pretrained BERT and RoBERTa in two main configurations: in a siamese network fine-tuned on SNLI (Bowman et al., 2015), Multi-Genre NLI (Williams et al., 2018) and (separately) on STS Argument Facet Similarity (Misra et al., 2016) and STS benchmarks (Cer et al., 2017) in a triplet network using the Wikipedia sections distinction dataset (Dor et al., 2018).

The SNLI and multi NLI datasets are familiar from the course; the AFS dataset similarly has labelled pairs with a 0-5 similarity rating but, as argumentative excerpts from dialogue, the notion of similarity involves to the reasoning of the excerpts. While SBERT/SRoBERTa-NLI typically outperforms BERT on STS benchmarks, on the AFS dataset, the excerpt spanning attention in BERT appears to be important and SBERT-AFS lags by a several points in this task.

The Wikipedia sections distinction dataset involves triplets: with two passages from one section of a Wikipedia article and a third from another section. They train with a triplet network (the three passages go through a single BERT model) and use a triplet loss that pushes the embeddings for similar passages closer and the dissimilar passages further in embedding space. This must be a challenging dataset to learn from: the sentences within a given section are not always semantically related; but it's an ingeneous way to generate a very large dataset. SBERT-WikiSec does quite well.

As a final evaluation, SBERT-NLI is tested in a transfer learning setting (Conneau and Kiela, 2018): even though SBERT is intended to be finetuned to each task (here it is tested on held out tasks), it outperformed many other sentence embeddings.

2.3 ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT (Khattab and Zaharia, 2020)

ColBERT sets out to leverage the full quality of BERT for contextualized information retrieval while reducing query time a hundred fold. Their design lands in the happy middle ground for query-document interactions in the network: late enough that the bulk of the expensive computations can be precomputed and indexed and yet there is an interaction layer that improves over completely separate query/document embeddings.

Like SBERT, the same BERT model is used to encode both query and document ([Q]/[D] markers encodes which a sequence is). The query is padded with mask tokens that effectively augment the query in a differentiable way ("the [MASK] positions are represented by paying attention (in the 12th layer) to all other query tokens" (Khattab)). The output of the BERT model passes through a linear layer to reduce the final embedding dimenionality. The late interaction is the sum of maximum similarity between each query token and the document tokens. The BERT layers are fine-tuned while the linear layer and [Q]/[D] marker's embeddings are trained from scratch using a pairwise softmax cross-entropy loss on triples of $\langle q, d^+, d^- \rangle$ (a query and positive + negative document match).

The form of this late interaction allows for very efficient querying: from faster loading of embeddings into the GPU to a form amenable to optimized large-scale vector-similarity search (specifically, `faiss` (Johnson et al., 2017)).

ColBERT was evaluated on two information retrieval benchmark datasets: MS Marco Ranking (Nguyen et al., 2016) and TREC Complex Retrieval (Dietz et al., 2017). MRR@10 of ColBERT matched the far more costly direct adaptation of BERT to ranking (Nogueira and Cho, 2019): in "re-ranking" mode, it did a little better than BERT$_{base}$; in "end-to-end" mode, it did a little worse than BERT$_{large}$. In the ablation study, the biggest improvements in MRR@10 were for the maximum similarity in the late interaction and the query augmentation.

2.4 RecoBERT: A Catalog Language Model for Text-Based Recommendations (Malkiel et al., 2020)

RecoBERT is an early example from Microsoft of using a pretrained LLM for text-based item recommendations, a pure item-to-item recommender system. Earlier attempts that also incorporated context and/or user-item interactions (eg. (Djuric et al., 2015), (Zheng et al., 2017) or (de Souza Pereira Moreira et al., 2018)) trained text embeddings from scratch with a custom network, making that possibly more costly to develop and less general.

Title, document pairs are input to BERT$_{large}$ as the sequence [CLS][title tokens][SEP][document tokens], 15% of which are masked; the output em-

beddings for the title and document are separately averaged. The cosine distance between the title, document embeddings, C_{TDM} is used in the title-description model (TDM) loss:

$$\mathcal{L}_{TDM} = -\frac{1}{2}\sum_{i=1}^{n}\left(y_i \log\left(C^i_{TDM}\right) + (1-y_i)\log\left(1-C^i_{TDM}\right)\right) \quad (3)$$

The total loss includes a mask language model component (following (Devlin et al., 2019)) for a classification layer maps the BERT [CLS] embedding back to vocabulary space.

For inference, to test the similarity of a candidate title′, document′ to a known title, document they compute: cosine similarity of titles, cosine similarity of documents and the TDM model cosine distance crossed title, document′ and title′, document pairs; the sum of which define a similarity metric that they rank by.

They evaluate RecoBERT on two datasets: a wine review catalogue (zackthoutt, 2017) (with an expert annotated test set[1]), and a fashion catalog (no further details provided, presumably proprietary). RecoBERT is trained separately for each: to train on the wine dataset with 120k examples on a system with one NVIDIA V100 32GB GPU. They compare against other sentence embeddings, pre-trained BERT without finetuning, and BERT finetuned to each domain (on the same reviews that RecoBERT trains on but without the TDM). RecoBERT outperformed the other approaches, often by a large margin, as quantified by either mean reciprocal ratio (MMR) or hit ratio (HR) for various top-k.

2.5 M6-Rec: Generative Pretrained Language Models are Open-Ended Recommender Systems (Cui et al., 2022)

3 Compare and contrast

Unfortunately, the SBERT paper doesn't include "a simple but tough-to-beat baseline for sentence embeddings" (Arora et al., 2017) in their comparisons. Strangely, the two papers state fairly different performance for the same embeddings on the same tasks (eg. simple average GloVe embeddings (Pennington et al., 2014) on the STS benchmarks); either the GloVe embeddings used were trained on different corpuses or they're reporting different summaries of results across the subtasks ("a simple..." stated mean across STS tasks and used the glove.840B.300d embeddings; the SBERT paper doesn't say). Since both state a 5-20 point improvement over unweighted GloVe embedding averages, it would have been very interesting to see how SBERT performs against a better baseline.

Like ColBERT, RecoBERT must to finetuned to each domain but much of the heavy computation can be precomputed and indexed, although the authors did not consider this. RecoBERT uses title, document pairs and train with a cosine loss to discriminate positive vs negative pairs; compare that with the triplets trained on in both SBERT and ColBERT.

4 Future work

Many studies suggest that a contrastive loss with triplets performs better; triplet loss works with the embeddings directly rather than forcing them first through bottleneck layers (Schroff et al., 2015). RecoBERT should see improvements switching to triplet training from their current (pairwise) constrastive loss.

Furthermore, the efficacy of triplet training relies on finding "good" positive/negative pairs (distinguishable but barely so), though gradient clipping must ease learning from too-hard pairs; in image and speech domains, much attention is given to choosing better pairs (Moindrot, 2018). Adapting these techniques to NLP may improve both SBERT and ColBERT.

[1] See the author's github page for this paper, https://github.com/r-papso/recobert

References

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2017. A simple but tough-to-beat baseline for sentence embeddings. *5th International Conference on Learning Representations, ICLR 2017*.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642. https://nlp.stanford.edu/projects/snli/.

Daniel Cer, Mona Diab, Eneko Agirre, Iigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 1–14. https://github.com/brmson/dataset-sts.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. https://github.com/facebookresearch/SentEval.

Zeyu Cui, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. M6-rec: Generative pretrained language models are open-ended recommender systems.

Gabriel de Souza Pereira Moreira, Felipe Ferreira, and Adilson Marques da Cunha. 2018. News session-based recommendations using deep neural networks. In *Proceedings of the 3rd Workshop on Deep Learning for Recommender Systems*, DLRS 2018, pages 15–23, New York, NY, USA. Association for Computing Machinery.

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

Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. Trec complex answer retrieval overview. *Proceedings of Text REtrieval Conference (TREC)*. http://trec-car.cs.unh.edu/.

Nemanja Djuric, Hao Wu, Vladan Radosavljevic, Mihajlo Grbovic, and Narayan Bhamidipati. 2015. Hierarchical neural language models for joint representation of streaming documents and their content. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15, page 248–255, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Liat Ein Dor, Yosi Mass, Alon Halfon, Elad Venezian, Ilya Shnayderman, Ranit Aharonov, and Noam Slonim. 2018. Learning thematic similarity metric from article sections using triplet networks. *Proceedings of the 56th Annual Meeting of the As- sociation for Computational Linguistics*, 2:49–54. https://research.ibm.com/haifa/dept/vst/debating_data.shtml.

Jeff Johnson, Matthijs Douze, and Herve Jegou. 2017. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*. https://github.com/facebookresearch/faiss.

Omar Khattab. Question about query augmentation.

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 39–48.

Itzik Malkiel, Oren Barkan, Avi Caciularu, Noam Razin, Ori Katz, and Noam Koenigstein. 2020. Recobert: A catalog language model for text-based recommendations. pages 1704–1714.

Amita Misra, Brian Ecker, and Marilyn A. Walker. 2016. Measuring the similarity of sentential arguments in dialogue. *Proceedings of the SIGDIAL 2016 Conference, The 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 276–287. https://nlds.soe.ucsc.edu/node/44.

Olivier Moindrot. 2018. Triplet loss and online triplet mining in tensorflow.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with bert. *arXiv preprint arXiv:1901.04085*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1:1112–1122. https://cims.nyu.edu/~sbowman/multinli/.

zackthoutt. 2017. Wine reviews. https://www.kaggle.com/datasets/zynicide/wine-reviews.

Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint deep modeling of users and items using reviews for recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, pages 425–434, New York, NY, USA. Association for Computing Machinery.