Bachelorarbeit Verteidigung

30.11.2021
Philipp Haustein

BHT
Berliner Hochschule für Technik

Studiere Zukunft

Inhalt

1. Thema
2. Hintergrund
3. Methodik
4. Auswertung
5. Zusammenfassung
6. Ausblick

Thema

Transformer Based Medication Prediction from Clinical Notes

Thema

Wozu?

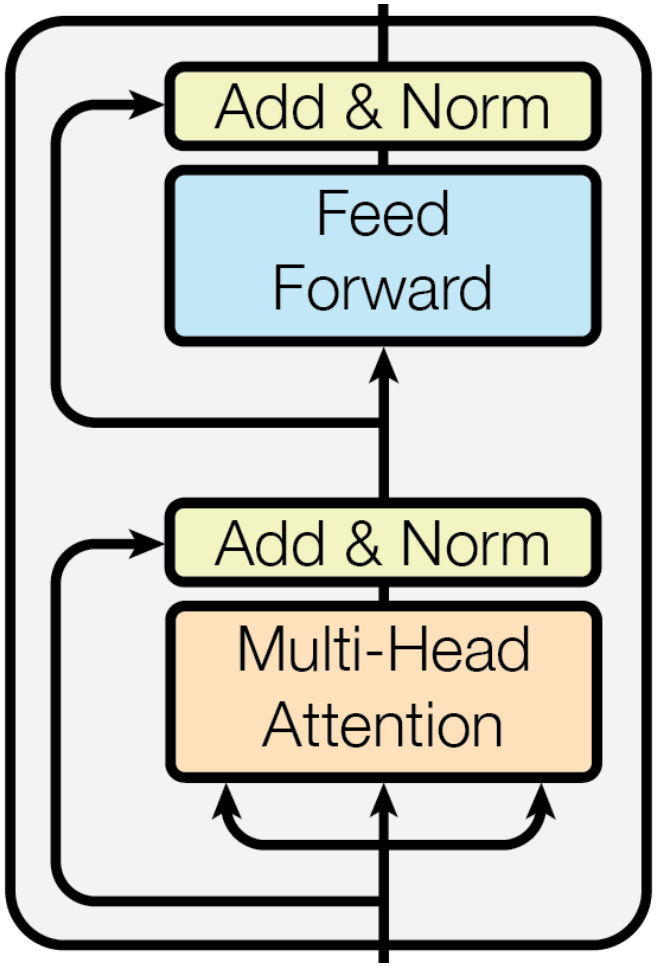
- Ärtz*innen haben wenig Zeit
- viele Medikamente und Krankheiten
- komplexe Entscheidungen + wenig Zeit = Schwierig
- Lösung: Automatisierte Vorschläge

Hintergrund

1. BERT
2. Bi-Encoder
3. Clinical Outcome Prediction
4. Medication Prediction

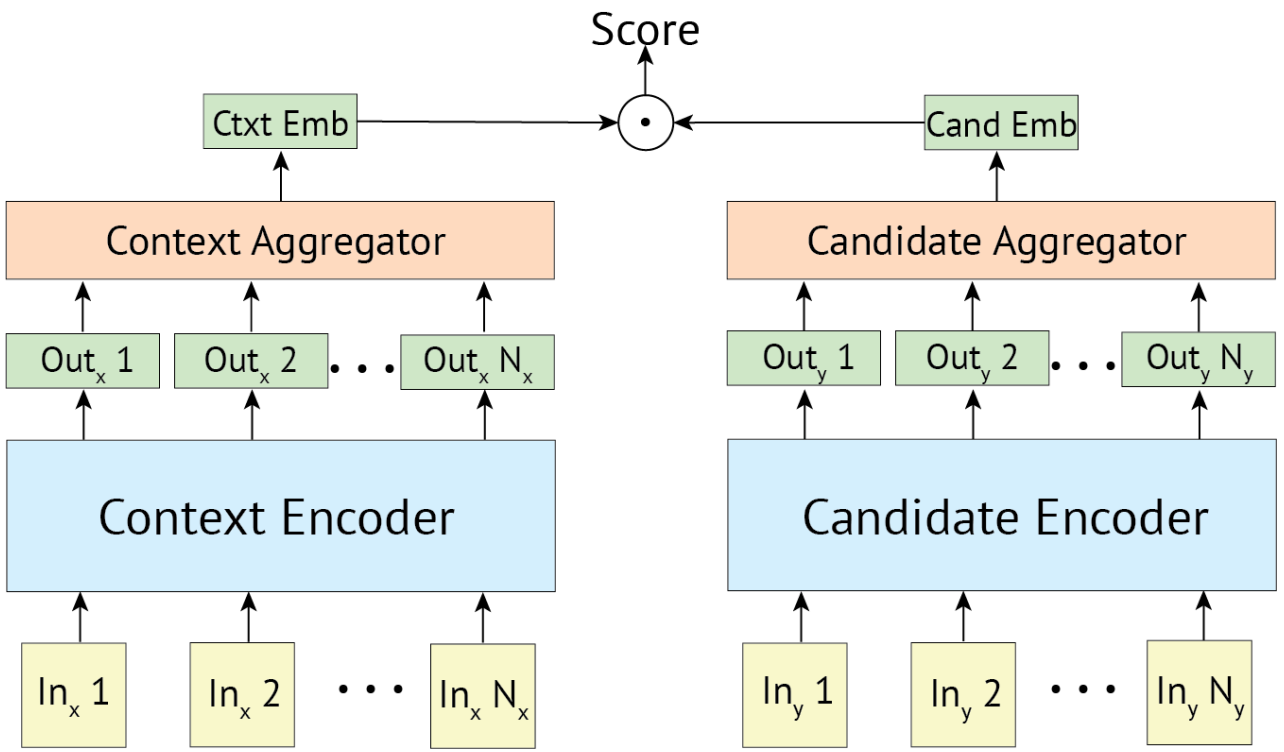
BERT

- Devlin et al. (2019)
- Nutzt Transformer-Encoder-Layers (Vaswani et al. (2017))
- Pretraining
- Finetuning
- BioBERT (Lee et al. (2019))



Bi-Encoder

- Humeau et al. (2020)
- Zwei Encoder
- Embeddings annähern durch Training



Clinical Outcome Prediction

- Verwand mit Medication Prediction
- Vorhersagen von klinischer Behandlung mithilfe von Patientenrepräsentation
- Textuelle Patientenrepräsentation (van Aken et al (2021), Zhang et al. (2020))

Medication Prediction

- Wenige veröffentlichungen
- Vorhersage von Haupt- und Nebenwirkungen (Jang et al. (2018), Wu et al. (2019))
- Auf Basis von Krebs DNA (Baptista et al. (2020), Kim et al. (2020))
- Auf Basis ICD-Code-Sequenz (Choi et al. (2016), Liu et al. (2020), Song et al. (2021))
- Keine textuelle Patientenrepresentation

Methodik

- Definition des Tasks
- Daten
- Ansätze

Definition des Tasks

- Vorhersage von Medikamenten
- Auf Basis von textueller Patientenrepräsentation
- Fokus auf Wirkstof

Daten

- MIMIC-III (Johnson et al. (2016))
- Admission Notes (van Aken et al. (2021))
- Wikipedia
- Trigram Fuzzy Matching

Ansätze

- BERT-Basierter Classifier
- BERT-Basierter Bi-Encoder

Classifier

- BERT-Classifier (Devlin et al. (2019))
- BioBERT
- Drug Classifier (Admission Note => Drug Name)
- Article Classifier (Admission Note => Linked Wikipedia Article)
- Training mit und ohne Long-Tail

Bi-Encoder

- BERT-Encoder (BioBERT)
- Dot-Product Similarity
- Single Label Loss (crossentropy)
- Multi Label Loss (binary crossentropy)
- Mit und ohne Long-Tail

Single Label Loss		
Aspirin		
Paracetamol		
Multi Label Loss		
Aspirin		
Paracetamol		

Auswertung

- Hyphothesen
- HPO
- Experimente
- Vergleich Classifier und Bi-Encoder

Hyphothesen

- Bi-Encoder besser als Classifier
- Training mit Long-Tail schlechter
- Multi-Label Bi-Encoder besser als Single-Label
- Recall besser als Precision
- Größere Batch-Sizes Besser

HPO Classifier

- Batch-Sizes kein signifikanter Unterschied

Model	Batch Size	Learning Rate
Drug Classifier	16	$4 \cdot 10^{-5}$
Drug Classifier All	16	$4 \cdot 10^{-5}$
Article Classifier	32	$4 \cdot 10^{-5}$
Article Classifier All	32	$4 \cdot 10^{-5}$

HPO Bi-Encoder

- Größere Batch-Sizes funktionieren besser
- Sigmoid funktioniert deutlich schlechter

Model	Batch Size	Learning Rate	Activation Function
Bi-Encoder	16	$4 \cdot 10^{-5}$	mish
Bi-Encoder All	16	$4 \cdot 10^{-5}$	mish
Bi-Encoder MC	32	$4 \cdot 10^{-5}$	mish
Bi-Encoder MC All	32	$4 \cdot 10^{-5}$	mish

Experimente Classifier

- Beste Resultate um Bester AUROC Wert
- Varianten mit Long-Tail nicht signifikant schlechter
- Precision bessert als Recall

Model	Accuracy	Precission	Recall	F1	AUROC
Drug Classifier	0.9735	0.7550	0.4165	0.5123	**0.8392**
Drug Classifier All	**0.9940**	0.7000	0.4294	0.5080	0.7799
Article Classifier	0.9224	**0.7604**	**0.5371**	**0.6135**	0.8112
Article Classifier All	0.9799	0.7590	0.5208	0.5972	0.8083

Experimente Bi-Encoder

- Beste Resultate um besten NDCG Wert

Model	AUROC	Recall@1	Recall@5	Recall@10	Recall@32	NDCG
Bi-Encoder	0.6675	0.3888	0.3394	0.3234	0.4219	0.6614
Bi-Encoder All	0.6853	0.3294	0.2669	0.2278	0.1932	0.5270
Bi-Encoder MC	**0.7666**	**0.9550**	**0.8599**	**0.7780**	**0.7282**	**0.8870**
Bi-Encoder MC All	0.7474	0.9514	0.8584	0.7766	0.6454	0.8523

Vergleich Classifier und Bi-Encoder

- Drug Classifier nicht vergleichbar
- Clasifier Besser als Bi-Encoder

Model	AUROC
Article Classifier	**0.8112**
Bi-Encoder	0.6675
Bi-Encoder MC	0.7666
Article Classifier All	**0.8083**
Bi-Encoder All	0.6853
Bi-Encoder MC All	0.7474

Zusammenfassung

- Bi-Encoder schlechter als Classifier
- Long-Tail nur unsignifikant schlecher
- Multi-Label Loss Bi-Encoder besser als Single Label
- Precision besser als Recall
- Größere Bi-Encoder Batch-Sizes besser
- Bi-Encoder deutlich schlechter mit sigmoid

Ausblick

- Andere Architekturen (Poly-Encoder, Cross-Encoder) (Humeau et al. (2020))
- Bessere Quelle für Medikament Beschreibungen (DailyMed(FDA))
- Mehr Aktivirungsfunktionen für Bi-Encoder
- Dosierung
- Längere Texte (Beltagy et al. (2020), Kitaev et al. (2020))
- Bessere Verknüpfung von Medikamenten

Quellen

- D. Baptista, P. G. Ferreira, and M. Rocha. Deep learning for drug response prediction in cancer. Briefings in Bioinformatics, 22(1):360–379, Jan. 2020. doi: 10.1093/bib/bbz171. URL https://doi.org/10.1093/bib/bbz171.
- I. Beltagy, M. E. Peters, and A. Cohan. Longformer: The long-document transformer. ArXiv:2004.05150, 2020.
- E. Choi, M. T. Bahadori, A. Schuetz, W. F. Stewart, and J. Sun. Doctor ai: Predicting clinical events via recurrent neural networks. In F. Doshi-Velez, J. Fackler, D. Kale, B. Wallace, and J. Wiens, editors, Proceedings of the 1st Machine Learning for Healthcare Conference, volume 56 of Proceedings of Machine Learning Research, pages 301–318, Northeastern University, Boston, MA, USA, 18–19 Aug 2016. PMLR. URL https://proceedings.mlr.press/v56/Choi16.html.

Quellen

- J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 27, 2019, Volume 1 (Long and Short Papers), pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-1423. URL https://aclanthology.org/N19-1423.
- I. Deznabi, M. Iyyer, and M. Fiterau. Predicting in-hospital mortality by combining clinical notes with time-series data. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4026–4031, Online, Aug. 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.352. URL https://aclanthology.org/2021.findings-acl.352.

Quellen

- B. van Aken, J.-M. Papaioannou, M. Mayrdorfer, K. Budde, F. A. Gers, and A. Löser. Clinical outcome prediction from admission notes using self-supervised knowledge integration. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021, pages 881–893. Association for Computational Linguistics, 2021. URL https://www.aclweb.org/anthology/2021.eacl-main.75/.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, pages 5998–6008, 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.

Quellen

- D. Zhang, C. Yin, J. Zeng, X. Yuan, and P. Zhang. Combining structured and unstructured data for predictive models: a deep learning approach. BMC Medical Informatics and Decision Making, 20(1), Oct. 2020. doi: 10.1186/s12911-020-01297-6. URL https://doi.org/10.1186/s12911-020-01297-6.
- G. Wu, J. Liu, and X. Yue. Prediction of drug-disease associations based on ensemble meta paths and singular value decomposition. BMC Bioinformatics, 20(S3), Mar. 2019. doi: 10.1186/s12859-019-2644-5. URL https://doi.org/10.1186/s12859-019-2644-5.
- S. Humeau, K. Shuster, M. Lachaux, and J. Weston. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, 2020. URL https://openreview.net/forum?id=SkxgnnNFvH.

Quellen

- G. Jang, T. Lee, S. Hwang, C. Park, J. Ahn, S. Seo, Y. Hwang, and Y. Yoon. PISTON: Predicting drug indications and side effects using topic modeling and natural language processing. Journal of Biomedical Informatics, 87:96–107, Nov. 2018. doi: 10.1016/j.jbi.2018.09.015. URL https://www.sciencedirect.com/science/article/pii/S1532046418301898.
- Y. Kim, S. Zheng, J. Tang, W. J. Zheng, Z. Li, and X. Jiang. Anticancer drug synergy prediction in understudied tissues using transfer learning. Journal of the American Medical Informatics Association, 28(1):42–51, Oct. 2020. doi: 10.1093/jamia/ocaa212. URL https://doi.org/10.1093/jamia/ocaa212.
- S. Liu, T. Li, H. Ding, B. Tang, X. Wang, Q. Chen, J. Yan, and Y. Zhou. A hybrid method of recurrent neural network and graph neural network for next-period prescription prediction. International Journal of Machine Learning and Cybernetics, 11(12):2849–2856, June 2020. doi: 10.1007/s13042-020-01155-x. URL https://doi.org/10.1007/s13042-020-01155-x.

Quellen

- J. Song, Y. Wang, S. Tang, Y. Zhang, Z. Chen, Z. Zhang, T. Zhang, and F. Wu. Local–global memory neural network for medication prediction. IEEE Transactions on Neural Networks and Learning Systems, 32(4):1723–1736, Apr. 2021. doi: 10.1109/tnnls.2020.2989364. URL https://doi.org/10.1109/tnnls.2020.2989364.
- A. E. Johnson, T. J. Pollard, L. Shen, L. wei H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark. MIMIC-III, a freely accessible critical care database. Scientific Data, 3(1), May 2016. doi: 10.1038/sdata.2016.35. URL https://www.nature.com/articles/sdata201635.
- J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, Sep 2019. ISSN 1460-2059. doi: 10.1093/bioinformatics/btz682. URL https://academic.oup.com/bioinformatics/article/36/4/1234/5566506.

Fragen?