Description-based Label Attention Classifier for Explainable ICD-9 Classification

Malte Feucht[1], **Zhiliang Wu**[2,3], **Sophia Althammer**[4], **Volker Tresp**[2,3]
[1]Department of Informatics, TU Munich [2]Institute of Informatics, LMU Munich
[3]Technology, Siemens AG [4]Faculty of Computer Science, TU Wien
malte.feucht@tum.de, {zhiliang.wu,volker.tresp}@siemens.com,
sophia.althammer@tuwien.ac.at

Abstract

ICD-9 coding is a relevant clinical billing task, where unstructured texts with information about a patient's diagnosis and treatments are annotated with multiple ICD-9 codes. Automated ICD-9 coding is an active research field, where CNN- and RNN-based model architectures represent the state-of-the-art approaches. In this work, we propose a description-based label attention classifier to improve the model explainability when dealing with noisy texts like clinical notes. We evaluate our proposed method with different transformer-based encoders on the MIMIC-III-50 dataset. Our method achieves strong results together with augmented explainablilty.

1 Introduction

Physicians are obliged to thoroughly document every patient encounter. Structured and semi-structured reports become more common, which contain comprehensive information about performed treatments, procedures and diagnoses. They are typically annotated with multiple billing codes, the international classification of diseases codes (ICD-9 in the US, ICD-10 in Europe). Annotating the noisy discharge summaries with ICD-9 codes is not only manual and labor-intensive, but also error-prone, which has attracted much attention both from clinical and technical perspectives. To facilitate the clinical workflow, we propose a new approach with state-of-the-art annotation performance while providing an explanation for the proposed annotation.

de Lima et al. (1998) introduced automated ICD-9 coding as a text-based multi-label classification problem. Deep learning-based approaches, which exploit convolutional (CNNs) and recurrent neural networks (RNNs) with attention mechanisms (Shi et al., 2017; Mullenbach et al., 2018; Vu et al., 2020) define the current state-of-the-art. Meanwhile, large-scale pre-trained language models based on transformer (Vaswani et al., 2017) architectures have demonstrated considerable performance improvements for text-based tasks, e.g., BERT (Devlin et al., 2019). Especially their ability to model long-range dependencies within an input sequence would potentially benefit the task of ICD-9 coding since the information for a certain label prediction can be distributed across the whole text. Unlike other areas of natural language processing (NLP), little research on applying transformer-based architectures on the task of ICD-9 coding has been explored (Pascual et al., 2021; Biswas et al., 2021; Ji et al., 2021). Sun and Lu (2020) argue that attention scores are able to capture global, absolute importance of work tokens and can thus provide some degree of explainability for text classification.

In this work, we propose a **description-based label attention classifier (DLAC)**. We show that it can be applied to different transformer-based encoders and provides explainable predictions on noisy texts. DLAC learns ICD-9 code embeddings by integrating the descriptions of the ICD-9 codes and applies the embeddings to the respective text representations to obtain label-specific representations for each code classification. We evaluate different model architectures on the MIMIC-III-50 dataset, a benchmark dataset for the task of ICD-9 coding, in order to answer the following research questions (RQs).

RQ1: *Which transformer-based encoder is best suited for ICD-9 coding?*

We evaluate and compare BERT (Devlin et al., 2019), hierarchical BERT (Pappagari et al., 2019) and Longformer (Beltagy et al., 2020) as transformer-based encoders and find that the Longformer (Beltagy et al., 2020) yields the best results for ICD-9 coding. Furthermore, we investigate:

RQ2: *How does our proposed description-based label attention classifier perform on ICD-9 coding?*

We compare DLAC with a common logistic regression classifier (LRC) on top of the transformer-based encoder for ICD-9 coding. While adding ex-

Proceedings of the 2021 EMNLP Workshop W-NUT: The Seventh Workshop on Noisy User-generated Text, pages 62–66

plainability to the predictions, DLAC outperforms the corresponding LRC by 1-4 %. Since explainable predictions are crucial for noisy texts, like discharge summaries, we investigate:

RQ3: *To which extent can the DLAC provide explainable predictions for ICD-9 codes?*

Since the attention scores in DLAC offer a way to explain the predictions with respect to different text segments, we analyze the top attention scores of DLAC and demonstrate the utility as part of a graphical interface.

2 Methods

As a multi-label text classification problem, each discharge summary is represented by a tokenized input sequence of words $\boldsymbol{X}_i := [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t_i}] \in \mathbb{R}^{l \times t_i}$, where l denotes the vector dimension and t_i is the length of the i-th input sequence. The goal is to predict a binary vector $\boldsymbol{y}_i \in \mathbb{R}^m$, where m represents the set size of ICD-9 codes. Each element in the predicted vector $\boldsymbol{y}$ is of value 0 or 1. In the following, we denote scalars with lowercase letters like x, vectors with bold lowercase letters like $\boldsymbol{x}$, and matrices with bold uppercase letters like $\boldsymbol{X}$.

2.1 Description-based Label Attention

An overview of the proposed model architecture with DLAC is illustrated in Figure 1. It includes a transformer-based encoder to represent the discharge summaries into a word embedding matrix $\boldsymbol{E}$. Meanwhile, the descriptions of the ICD-9 codes are represented with a description embedding matrix $\boldsymbol{D}$, which is initialized by ICD-9 code descriptions with Word2vec embeddings (Mikolov et al., 2013).

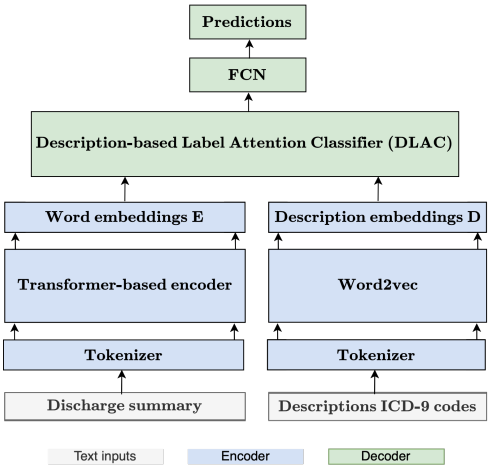

Figure 1: Overall model architecture

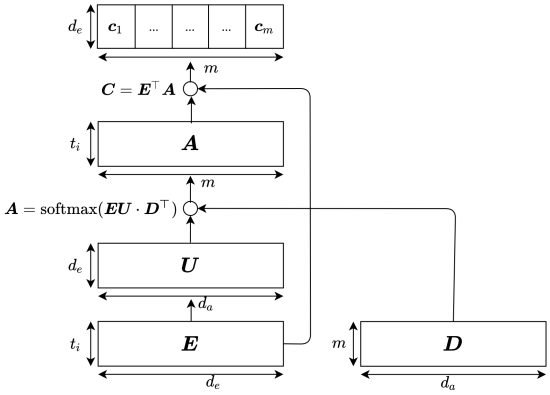

Figure 2: Description-based Label Attention Classifier

More specifically, as shown in Figure 2, DLAC is fed with two matrices: The word embedding matrix $\boldsymbol{E} \in \mathbb{R}^{t_i \times d_e}$ is computed from a transformer-based encoder and the ICD-9 code description embedding matrix $\boldsymbol{D} \in \mathbb{R}^{m \times d_a}$. Here, d_e, d_a denote the dimension of the word embeddings and the description embeddings, respectively. The description embedding matrix $\boldsymbol{D}$ is set to be trainable. To compute the attention score a_{jk} for the i-th word on the j-th label, we first apply a dimension transformation on the word embedding matrix with $\boldsymbol{U} \in \mathbb{R}^{d_e \times d_a}$ to match the shape of description matrix $\boldsymbol{D}$. The ICD-9 code description vectors stored in the description matrix $\boldsymbol{D}$ can be seen as queries that include the essential information from the label description of respective ICD-9 codes. Formally, we compute the label attention matrix $\boldsymbol{A} \in \mathbb{R}^{t_i \times m}$ as

$$\boldsymbol{A} = \text{softmax}(\boldsymbol{EU} \cdot \boldsymbol{D}^\top).$$

After that, we compute contextual embeddings for each label by aggregating information from the word embedding matrix $\boldsymbol{E}$ with attention scores in $\boldsymbol{A}$. More concretely, the contextual embedding matrix $\boldsymbol{C} \in \mathbb{R}^{d_e \times m}$ is computed as

$$\boldsymbol{C} = \boldsymbol{E}^\top \boldsymbol{A}.$$

2.2 Classification

Each label specific contextual embedding is fed into a single layer fully-connected network (FCN) for the prediction of the respective ICD-9 code label. A sigmoid activation function is applied to have a probabilistic prediction. The training objective is the binary cross-entropy loss computed from predictions $\hat{\boldsymbol{y}}_i \in \mathbb{R}^m$ and ground-truth labels $\boldsymbol{y}_i \in \mathbb{R}^m$.

2.3 Transformer-based Encoders

Since DLAC is agnostic to encoders, we conduct experiments with different transformer-based encoders by integrating them into the proposed model architecture. The first one is a pre-trained **BERT$_{\text{BASE}}$** model (Devlin et al., 2019), which can only consume input sequences with a length of up to 512 tokens. Longer discharge summaries are simply truncated. The second architecture is a **hierarchical BERT$_{\text{BASE}}$** model (Pappagari et al., 2019), which aims to overcome the input sequence length limitation. The discharge summary of length t_i is split into k overlapping chunks, where $k = \lceil \frac{t_i}{512} \rceil$, where $\lceil \cdot \rceil$ is a ceiling operation. The chunks are fed sequentially into the **BERT$_{\text{BASE}}$** model to obtain the word embedding matrices for every chunk, which are then averaged using mean-pooling across all chunks. The third architecture is a pre-trained **Longformer$_{\text{BASE}}$** model (Beltagy et al., 2020). A limitation of transformer-based language models such as BERT is their inability to process long input sequences due to the computational cost of the self-attention mechanism, which scales quadratically with the input sequence length. The **Longformer$_{\text{BASE}}$** model overcomes this limitation by offering a "sparsified" self-attention mechanism, making it more suitable to process longer input sequences. As a result, the **Longformer$_{\text{BASE}}$** model can process input sequences of lengths of up to 4096 tokens.

3 Data

MIMIC-III (Johnson et al., 2016) is a large, freely available clinical database. Similar to Mullenbach et al. (2018), we create the subset MIMIC-III-50 from the full dataset. It includes discharge summaries containing the 50 most frequent ICD-9 codes, as otherwise, the label distribution is extremely imbalanced. After pre-processing, MIMIC-III-50 contains 11.368 samples and 50 ICD-9 codes, where a summary of different statistics can be found in Table 1. Following the pre-processing in Mullenbach et al. (2018), we first lowercase all tokens, remove punctuations and remove numerical tokens-only. The MIMIC-III-50 dataset is split into a training, validation and test set with 8.066, 1.573 and 1.729 samples, respectively.

4 Experiments

We train the proposed description-based label attention classifier (DLAC) using BERT, hierarchical BERT, and Longformer as encoders. As a baseline classifier, we choose a simple logistic regression classifier (LRC) on top of the different encoders. In contrast to DLAC, LRC does not take the description embeddings $\boldsymbol{D}$, into account. We set $d_a = 600$ and $d_e = 768$. We train all architectures using Adam optimizer with a learning rate of $\alpha = 1.41 \times 10^{-5}$ and a global batch size of 64. Furthermore, we use k-fold cross-validation with $k = 5$ folds and train every fold for 25 epochs. For regularization, we use dropout layers with a probability set to $p = 0,1$ and early stopping. We use the widely adopted micro-and macro averaged area under the ROC curve (AUC), F1 and Precision@n as evaluation metrics to ensure comparability with other works. For P@n we choose $n = 5$ because this roughly equals the average number of ICD-9 codes one discharge summary is annotated with, which is $5,77$ for the MIMIC-III-50 dataset. The implementation is made available to ensure the reproducibility of the work[1].

MIMIC-III-50	# Words	# ICD-9 Codes
mean	1.612	5,77
std	788	3,37
min	105	1
max	7.567	24
25%	1.065	3
50%	1.478	5
75%	1.992	5

Table 1: Descriptive statistics of MIMIC-III-50 dataset

5 Results and Discussions

Table 2 presents the results for all proposed model architectures. The Longformer+DLAC yields the best results across all metrics for the architectures in this work.

5.1 RQ1: Transformer-based Encoders

Among the transformer-based encoders that are combined with the simple LRC classifier (BERT+LRC, H-BERT+LRC, Longformer+LRC), the Longformer encoder yields the best results across all metrics. This can be attributed to the inability of BERT to process sequences longer than 512, where over 75% of the discharge summaries are truncated and potentially important information is disregarded. For the micro- and macro F1 scores, H-BERT+LRC model yields even poorer results

[1] https://git.io/JzOyk

Model	AUC		F1		P@5
	Macro	Micro	Macro	Micro	
BERT+LRC	0.80 ± 0.007	0.84 ± 0.006	0.33 ± 0.033	0.45 ± 0.026	0.51 ± 0.011
H-BERT+LRC	0.82 ± 0.006	0.86 ± 0.006	0.29 ± 0.030	0.41 ± 0.032	0.51 ± 0.012
Longformer+LRC	0.85 ± 0.003	0.89 ± 0.003	0.48 ± 0.005	0.58 ± 0.003	0.59 ± 0.012
BERT+DLAC	0.80 ± 0.006	0.84 ± 0.004	0.35 ± 0.032	0.46 ± 0.026	0.51 ± 0.013
H-BERT+DLAC	0.83 ± 0.035	0.87 ± 0.004	0.32 ± 0.020	0.43 ± 0.013	0.52 ± 0.010
Longformer+DLAC*	0.87 ± 0.008	0.91 ± 0.006	0.52 ± 0.020	0.62 ± 0.024	0.61 ± 0.013
JointLAAT**	**0.93**	**0.95**	**0.67**	**0.72**	**0.68**
TransICD	0.89	0.92	0.56	0.64	0.62

Table 2: Test results on the MIMIC-III-50 dataset for all proposed model architectures compared to the state-of-the-art architectures. ** marks the best overall model architecture. * marks the best model architecture of this work.

than the BERT+LRC model. This indicates that the way of aggregating the chunks k using mean-pooling is suboptimal, and thus the model fails to create rich input feature representations.

5.2 RQ2: Longformer+DLAC

The results of the Longformer+DLAC model show that it performs well on the task of ICD-9 coding. In addition, DLAC outperforms the LRC for all encoder architectures on the task of ICD-9 coding across all metrics by $1-4\%$. Meanwhile, JointLAAT (Vu et al., 2020) is one state-of-the-art CNN-based model architecture. Our Longformer+DLAC underperforms it by $\Delta - 0.06, \Delta - 0.04, \Delta - 0.15, \Delta - 0.10, \Delta - 0.07$ for Macro AUC, Micro AUC, Macro F1, Micro F1 and P@5 respectively. In comparison to state-of-the-art transformer-based architectures, like TransICD (Biswas et al., 2021), our model shows a comparable performance with the difference being $\Delta - 0.02, \Delta - 0.01, \Delta - 0.04, \Delta - 0.02, \Delta - 0.01$ for Macro AUC, Micro AUC, Macro F1, Micro F1 and P@5, respectively. Transformer-based models haven't reached state-of-the-art dominated by more lightweight CNN-based architectures. This can be partially attributed to the fact that the MIMIC-III-50 dataset does not hold enough training samples for training such a large architecture, e.g., the Longformer+DLAC has up to 152 million trainable parameters. Furthermore, the Longformer+DLAC model could potentially be improved by using a domain-specific, pre-trained Longformer architecture (Gu et al., 2020) and by developing a regularization mechanism (Cao et al., 2019) that helps to classify rare labels more accurately. On the other hand, as an encoder agnostic classifier, DLAC can be combined with other models to improve the performance while keeping explainable predictions.

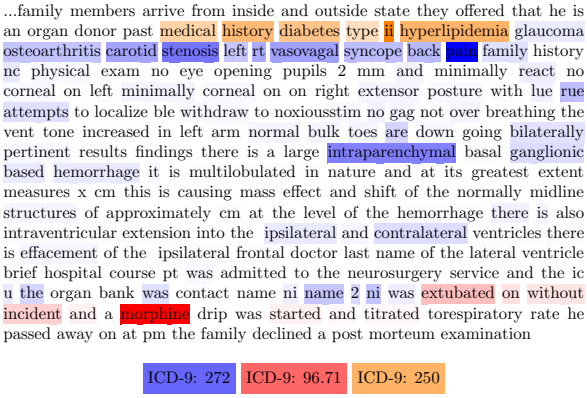

Figure 3: The top attention scores for the predicted ICD-9 codes **272, 96.71, 250** are highlighted with color intensity. Higher color intensity represents larger attention scores and vice versa.

5.3 RQ3: Explainability

In a setting where a machine learning model would serve as a decision support tool for medical workers, explainability of the obtained model predictions are of utmost importance. In contrast to LRC, DLAC provides explainable predictions using the attention scores. DLAC can retrieve the top attention scores for each ICD-9 code prediction. As an example, the text segments that lead to a certain ICD-9 code prediction are highlighted with respective color intensity in Figure 3. This can be useful for working with noisy texts in general because it provides some extent of explainability.

6 Conclusion

Transformer-based architectures show promising performance on the task of ICD-9 coding. We find

that the Longformer encoder is the best suitable encoder architecture for processing long, noisy input sequences such as discharge summaries. We show that our proposed description-based label attention classifier (DLAC) can be applied to various transformer-based encoders and the resulting model outperforms a common decoder architecture like logistic regressions by 1-4%. In addition, our proposed DLAC is especially suitable for a practical use case when working with fuzzy, long texts such as the discharge summaries, where explainability for the predicted ICD-9 codes is necessary.

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Biplob Biswas, Thai-Hoang Pham, and Ping Zhang. 2021. Transicd: Transformer based code-wise attention model for explainable icd coding. In *International Conference on Artificial Intelligence in Medicine*, pages 469–478. Springer.

Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Aréchiga, and Tengyu Ma. 2019. Learning imbalanced datasets with label-distribution-aware margin loss. In *NeurIPS*.

Luciano RS de Lima, Alberto HF Laender, and Berthier A Ribeiro-Neto. 1998. A hierarchical approach to the automatic categorization of medical documents. In *Proceedings of the seventh international conference on Information and knowledge management*, pages 132–139.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *CoRR*, abs/2007.15779.

Shaoxiong Ji, Matti Hölttä, and Pekka Marttinen. 2021. Does the magic of bert apply to medical code assignment? a quantitative study. *arXiv preprint arXiv:2103.06511*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana. Association for Computational Linguistics.

Raghavendra Pappagari, Piotr Zelasko, Jesús Villalba, Yishay Carmiel, and Najim Dehak. 2019. Hierarchical transformers for long document classification. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 838–844. IEEE.

Damian Pascual, Sandro Luck, and Roger Wattenhofer. 2021. Towards BERT-based automatic ICD coding: Limitations and opportunities. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 54–63, Online. Association for Computational Linguistics.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2017. Towards automated ICD coding using deep learning. *CoRR*, abs/1711.04075.

Xiaobing Sun and Wei Lu. 2020. Understanding attention for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3418–3428.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thanh Vu, Dat Quoc Nguyen, and Anthony Nguyen. 2020. A label attention model for icd coding from clinical text. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3335–3341. International Joint Conferences on Artificial Intelligence Organization. Main track.