Dilated Convolutional Attention Network for Medical Code Assignment from Clinical Text

Shaoxiong Ji[†], Erik Cambria[‡] and Pekka Marttinen[†]
[†] Department of Computer Science, Aalto University, Finland
{shaoxiong.ji; pekka.marttinen}@aalto.fi
[‡] School of Computer Science and Engineering,
Nanyang Technological University, Singapore
cambria@ntu.edu.sg

Abstract

Medical code assignment, which predicts medical codes from clinical texts, is a fundamental task of intelligent medical information systems. The emergence of deep models in natural language processing has boosted the development of automatic assignment methods. However, recent advanced neural architectures with flat convolutions or multi-channel feature concatenation ignore the sequential causal constraint within a text sequence and may not learn meaningful clinical text representations, especially for lengthy clinical notes with long-term sequential dependency. This paper proposes a Dilated Convolutional Attention Network (DCAN), integrating dilated convolutions, residual connections, and label attention, for medical code assignment. It adopts dilated convolutions to capture complex medical patterns with a receptive field which increases exponentially with dilation size. Experiments on a real-world clinical dataset empirically show that our model improves the state of the art.

1 Introduction

Medical code assignment categorizes clinical documents with sets of codes to facilitate hospital management and improve health record searching (Hsia et al., 1988; Farkas and Szarvas, 2008). These clinical texts comprise physiological signals, laboratory tests, and physician notes, where the International Classification of Diseases (ICD) coding system is widely used for annotation. Most hospitals rely on manual coding by human coders to assign standard diagnosis codes to the discharge summaries for billing purposes. However, this work is and error-prone (Hsia et al., 1988; Farzandipour et al., 2010). Incorrect coding can cause billing mistakes and mislead other general practitioners when patients are readmitted. Intelligent automated coding systems could act as a recommendation system to help coders to allocate correct medical codes to clinical notes.

Automatic medical code assignment has been intensively researched during the past decades (Crammer et al., 2007; Stanfill et al., 2010). Recent advances in natural language processing (NLP) with deep learning techniques have inspired many methods for automatic medical code assignment (Shi et al., 2017; Mullenbach et al., 2018; Li and Yu, 2020). Zhang et al. (2020) incorporated structured knowledge into medical text representations by preserving translational property of concept embeddings. However, several challenges remain in medical text understanding. Diagnosis notes contain complex diagnosis information, which includes a large number of professional medical vocabulary and noisy information such as non-standard synonyms and misspellings. Free text clinical notes are lengthy documents, usually from hundreds to thousands of tokens. Thus, medical text understanding requires effective feature representation learning and complex cognitive process to enable multiple diagnosis code assignment.

Previous neural methods for medical text encoding generally fall into two categories. Medical text modeling is commonly regarded as a synonym of recurrent neural networks (RNNs) that capture the sequential dependency. Such works include AttentiveLSTM (Shi et al., 2017), Bi-GRU (Mullenbach et al., 2018) and HA-GRU (Baumel et al., 2018). The other category uses convolutional neural networks (CNNs) such as CAML (Mullenbach et al., 2018) and MultiResCNN (Li and Yu, 2020). These methods only capture locality but have achieved the optimal predictive performance on medical code assignment.

Inspired by the generic temporal convolutional network (TCN) architecture (Bai et al., 2018), we consider medical text modeling with causal con-

straints, where the encoding of the current token only depends on previous tokens, using the dilated convolutional network. We combine it with the label attention network for fine-grained information aggregation.

Distinction of Our Model   The MultiResNet is currently the state-of-the-art model. It applies multi-channel CNN with different filters to learn features and further concatenates these features to produce a final prediction. In contrast, our model extends the TCN to sequence modeling that uses a single filter and the dilation operation to control the receptive field. In addition, instead of weight tying used in the TCN, we customize it with label attention pooling to extract relevant rich features.

Our Contributions   We contribute to the literature in three ways. (1) We consider medical text modeling from the perspective of imposing the sequential causal constraint in medical code assignment using dilated convolutions, which effectively captures long sequential dependencies and learns contextual representations in the long clinical notes. (2) We propose a dilated convolutional attention network (DCAN), coupling residual dilated convolution, and label attention network for more effective and efficient medical text modeling. (3) Experiments in real-world medical data show improvement over the state of the art. Compared with multi-channel CNN and RNN models, our model also offers a smaller computational cost.

2   Proposed Model

This section describes the proposed model - Dilated Convolutional Attention Network (DCAN). It includes three main components, i.e., dilated convolution for learning features from word embeddings of clinical notes, residual connection for stacking a deep neural architecture, and label attention module for prioritizing relevant representation for different labels. The architecture of our proposed model is illustrated in Fig. 1.

Our model benefits from the effective integration of these three neural modules. Dilated convolutions are widely used in audio signal modeling (Oord et al., 2016) and semantic segmentation (Yu and Koltun, 2015). Yu and Koltun (2015) proposed dilated convolutions with an exponentially large receptive filed. Bai et al. (2018) utilized causal convolutions with dilation and tied weighting for sequence modeling. Following their works, we

integrated dilated convolution with label attention network for better medical text encoding to predict diagnosis codes. The dilated convolution follows the causal constraint of sequence modeling. By stacking dilated convolutions with residual connection (He et al., 2016), our DCAN model can be built as a very deep neural network to learn different levels of features. And the final label attention module further extracts the most relevant information to the label space.
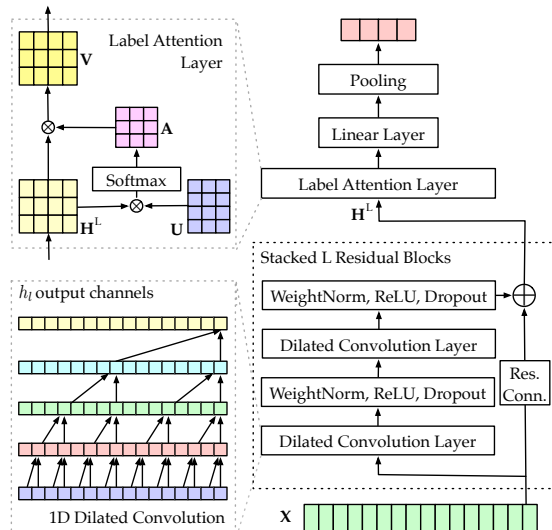

Figure 1: Model architecture of dilated convolutional attention network

2.1   Dilated Convolution Layers

A clinical note with n words is denoted as $\{x_0, \ldots, x_n\}$. We use word2vec (Mikolov et al., 2013) to train word embeddings from raw tokens. Word embedding matrix of a clinical note is denoted as $[\mathbf{w}_1, \ldots, \mathbf{w}_n]^T \in \mathbb{R}^{n \times d_e}$, where d_e is the dimension of word vectors. The word embeddings are then inputted into the dilated convolution layers, which are also called convolutions with dilated filters. Specifically, we use a 1D convolution operator to each dimension (i.e. channel) of the word vectors. Given a sequence of one-dimensional elements $\mathbf{x} \in \mathbb{R}^n$ and a convolutional filter $f : \{0, \ldots, k-1\} \to \mathbb{R}$, the one-dimensional dilated convolution $\mathcal{F}_d$ is denoted as

$$\mathcal{F}_d(s) = (\mathbf{x} *_d f)(s) = \sum_{i=0}^{k-1} f(i) \cdot \mathbf{x}_{s-d\cdot i}, \quad (1)$$

where d is the dilation size of the spacing between kernel elements, s is the element of input sequence,

k is the convolving kernel (aka, the filter) size, and $s - d \cdot i$ refers to past time steps. The 1D dilated convolution has h_l output channels, i.e., for each of the d_e input channels h_l features are learned and summed over the input channels. The dilated convolution is followed by a weight normalization, an activation function, and a dropout operation. Two dilated convolution layers ared stacked to a dilated convolution block. It outputs a hidden representation $\mathbf{H}^l \in \mathbb{R}^{n \times h_l}$ of the l-th layer, where the dimension d_h of the hidden representation is the number of output channels in the last dilated convolution layer. To expand the receptive field, the dilation size is exponentially increased, i.e., $d_i \in \{2^i\}$ for $i = 0, 1, \ldots, l - 2$.

2.2 Residual Connections

Residual connections (He et al., 2016) of l residual blocks are built upon the dilated convolution layers to create deep neural networks. Given the input encoding vector $\mathbf{x}$, the output of residual connection is denoted as $o = \sigma(\mathbf{x} + \mathcal{G}(\mathbf{x}))$, where $\mathcal{G}$ represents neural layers and σ is a non-linear activation function. We use residual mechanism between two stacked dilated convolution layers, which is formalized as:

$$\mathbf{H}^{l+1} = \sigma(\mathbf{H}^l + \mathcal{G}(\mathbf{H}^l)). \qquad (2)$$

2.3 Label Attention Layer

We apply the label attention layer to prioritize important information in the hidden representation relevant to ICD codes. Specifically, the dot product attention is used to calculate the attention score $\mathbf{A} \in \mathbb{R}^{n \times m}$ as:

$$\mathbf{A} = \mathrm{Softmax}(\mathbf{H}^L \mathbf{U}), \qquad (3)$$

where $\mathbf{H}^L$ (the superscript represents the ordinal of the layer $\mathbf{H}$ and not the power) is the hidden encoding of the L-th layer, $\mathbf{U} \in \mathbb{R}^{h_L \times m}$ is the parameter matrix of the label attention layer (also known as the query), and m is the number of ICD codes. The attention matrix $\mathbf{A}$ captures the importance of ICD code and hidden word representation pair. The output of the attention layer is then calculated by multiplying attention $\mathbf{A}$ with the hidden representation from residual dilated convolution layers. The attentive representation $\mathbf{V} \in \mathbb{R}^{m \times h_L}$ is formalized as

$$\mathbf{V} = \mathbf{A}^\mathrm{T} \mathbf{H}^L. \qquad (4)$$

With features representing sequential dependency and label awareness, the final representation is further used for medical code classification.

2.4 Classification Layer

The classification layer is a linear fully-connected layer. The i-th projected representation $\mathbf{Y}_i \in \mathbb{R}^k$ is calculated as:

$$\mathbf{Y}_i = \mathbf{V}_i \mathbf{W}^\mathrm{T} + \mathbf{b}, \qquad (5)$$

where $\mathbf{W} \in \mathbb{R}^{k \times h_L}$ is the linear weight, $\mathbf{b} \in \mathbb{R}^{1 \times k}$ is the bias, and $\mathbf{Y}_i$ and $\mathbf{V}_i$ are the i-th row of $\mathbf{Y}$ and $\mathbf{V}$ for $i \in \{1, \ldots, m\}$. The predicted logits $\hat{\mathbf{y}} \in \mathbb{R}^m$ between 0 and 1 are produced by a pooling operation over the linearly projected matrix $\mathbf{Y}$ and passed into the Sigmoid activation function, denoted as:

$$\hat{\mathbf{y}} = \mathrm{Sigmoid}(\mathrm{Pooling}(\mathbf{Y})). \qquad (6)$$

2.5 Training

ICD code assignment is a typical multi-label multi-class classification problem. We adopt the binary cross entropy loss denoted as:

$$\mathcal{L} = \sum_{i=1}^{m} \left[-y_i \log(\hat{y}_i) - (1 - y_i) \log(1 - \hat{y}_i) \right], \qquad (7)$$

where $y_i \in \{0, 1\}$ is the ground-truth label, $\hat{y}_i$ is the sigmoid score for prediction, and m is the number of ICD codes. To mitigate the effect of noisy labels, we apply label smoothing over ground-truth labels y penalizing model from over-confident predictions. The modified targets $\tilde{y}$ are calculated as:

$$\tilde{y}_i = y_i(1 - \alpha) + \alpha/m, \qquad (8)$$

where α is the smoothing coefficient. We use Adam optimizer (Kingma and Ba, 2014) to train the model with backpropagation.

3 Experiments

This section introduces the experimental analysis of real-world clinical datasets. Our proposed models are compared with several recent strong baselines. The code is publicly available at `https://agit.ai/jsx/DCAN`.

3.1 Dataset and Settings

This paper focuses on textual discharge summaries from a hospital stay. Following Shi et al. (2017)

Table 1: Results on MIMIC-III dataset with top-50 ICD codes. "-" indicates no results reported in the original paper.

Model	AUC-ROC		F1		
	Macro	Micro	Macro	Micro	P@5
CNN (Kim, 2014)	87.6	90.7	57.6	62.5	62.0
C-MemNN (Prakash et al., 2017)	83.3	-	-	-	42.0
Attentive LSTM (Shi et al., 2017)	-	90.0	-	53.2	-
Bi-GRU (Mullenbach et al., 2018)	82.8	86.8	48.4	54.9	59.1
CAML (Mullenbach et al., 2018)	87.5	90.9	53.2	61.4	60.9
DR-CAML (Mullenbach et al., 2018)	88.4	91.6	57.6	63.3	61.8
LEAM (Wang et al., 2018)	88.1	91.2	54.0	61.9	61.2
MultiResCNN (Li and Yu, 2020)	89.9±0.4	92.8±0.2	60.6±1.1	67.0±0.3	64.1±0.1
DCAN (Ours)	**90.2**±0.6	**93.1**±0.1	**61.5**±0.7	**67.1**±0.1	**64.2**±0.2

and Mullenbach et al. (2018), additional experiment on the subset of MIMIC-III (Johnson et al., 2016) with the top 50 frequent labels is conducted. Free-text discharge summaries are extracted, including raw notes, ICD diagnoses, and procedures for patients. Textual notes related to the same admission are concatenated to a single document to be used as input to our model. Each document is labeled with a set of ICD-9 diagnosis and procedure codes, which are the prediction targets. We use the standard train-test partition. The MIMIC-III dataset with top-50 codes contains 8,066 training, 1,573 development, and 1,729 test instances.

Settings We preprocess the textual documents following the preprocessing procedures developed by Mullenbach et al. (2018) and Li and Yu (2020). The NLTK package is utilized for tokenization and all tokens are converted into lowercase. Alphabetic characters such as numbers and punctuations are removed. All documents are truncated at the length of 2500 tokens. We choose some common settings from prior publications. For example, the word embedding dimension is 100, the dropout rate is 0.2. The Adam optimizer Kingma and Ba (2014) is used to optimize our model parameters. The rest choices of hyper-parameters are configured via random search.

3.2 Baselines

Baselines models include memory network based C-MemNN (Prakash et al., 2017), the joint embedding model (LEAM) (Wang et al., 2018), RNN-based models like Attentive LSTM (Shi et al., 2017) and Bi-GRU (Mullenbach et al., 2018), and CNN-based models such as vanilla CNN (Kim, 2014), CAML (Mullenbach et al., 2018) and MultiResCNN (Li and Yu, 2020).

3.3 Results

We evaluate the F1-score and area under the receiver operating characteristic curve (AUC-ROC) with both micro and macro averaging, and the precision at k codes with $k = 5$ (P@5). The results are shown in Table 1. Our model outperforms the state-of-the-art in all the metrics. To compare with the MultiResCNN model, we follow its setting and run our model for three times. We average the predictive scores and calculate their standard deviation. Our model has a clear improvement in the macro F1-score when the macro score calculates the label-wise average by treating all codes equally. For the other metrics, our model still has a marginal improvement with a lower or comparable standard deviation. We also try the pre-trained Bidirectional Encoder Representations from Transformers(BERT) model (Devlin et al., 2019) for sequence classification. However, the BERT model does not work well in this task. This conclusion is also reported by Li and Yu (2020).

Table 2: Computational cost comparison

Model	# params.	training time	training ep.
CAML	6.2M	673 s/ep	85 epochs
MultiResCNN	11.9M	1161 s/ep	26 epochs
DCAN (Ours)	8.7M	951 s/ep	23 epochs

Computational efficiency We compared the computational efficiency from two perspectives, i.e., number of parameters and convergence epochs, results are shown in Table 2. Not relying on concatenated multi-channel features, our model has fewer trainable parameters and takes less training time than the state-of-the-art MultiResCNN. Moreover, our model converges faster.

4   Conclusion

Recent years extensively studies the automatic medical code assignment. Neural clinical text encoding models use CNNs to extract local features and RNNs to preserve sequential dependency. This paper combines both by using dilated convolution. The dilated convolutional attention network (DCAN) consists of dilated convolution layers, residual connections, and the label attention layer. The DCAN model obeys the causal constraint of sequence encoding and learns rich representations to capture label-aware importance. Through experiments on the MIMIC-III dataset, our model shows better predictive performance than the state-of-the-art methods.

Acknowledgments

References

Shaojie Bai, J Zico Kolter, and Vladlen Koltun. 2018. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv preprint arXiv:1803.01271*.

Tal Baumel, Jumana Nassour-Kassis, Raphael Cohen, Michael Elhadad, and Noemie Elhadad. 2018. Multi-label Classification of Patient Notes: Case Study on ICD Code Assignment. In *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.

Koby Crammer, Mark Dredze, Kuzman Ganchev, Partha Talukdar, and Steven Carroll. 2007. Automatic Code Assignment to Medical Text. In *Biological, Translational, and Clinical language processing*, pages 129–136.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT*.

Richárd Farkas and György Szarvas. 2008. Automatic Construction of Rule-based ICD-9-CM Coding Systems. In *BMC Bioinformatics*, volume 9(Suppl 3), pages 1–9. Springer.

Mehrdad Farzandipour, Abbas Sheikhtaheri, and Farahnaz Sadoughi. 2010. Effective Factors on Accuracy of Principal Diagnosis Coding based on International Classification of Diseases, the 10th Revision (ICD-10). *International Journal of Information Management*, 30(1):78–84.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

David C Hsia, W Mark Krushat, Ann B Fagan, Jane A Tebbutt, and Richard P Kusserow. 1988. Accuracy of Diagnostic Coding for Medicare Patients under the Prospective-payment System. *New England Journal of Medicine*, 318(6):352–355.

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*, 3:160035.

Yoon Kim. 2014. Convolutional Ceural Networks for Sentence Classification. *arXiv preprint arXiv:1408.5882*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Fei Li and Hong Yu. 2020. ICD Coding from Clinical Text Using Multi-Filter Residual Convolutional Neural Network. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable Prediction of Medical Codes from Clinical Text. In *Proceedings of NAACL-HLT*, pages 1101–1111.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. *arXiv preprint arXiv:1609.03499*.

Aaditya Prakash, Siyuan Zhao, Sadid A Hasan, Vivek Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2017. Condensed Memory Networks for Clinical Diagnostic Inferencing. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P Xing. 2017. Towards Automated ICD Coding Using Deep Learning. *arXiv preprint arXiv:1711.04075*.

Mary H Stanfill, Margaret Williams, Susan H Fenton, Robert A Jenders, and William R Hersh. 2010. A Systematic Literature Review of Automated Clinical Coding and Classification systems. *Journal of the American Medical Informatics Association*, 17(6):646–651.

Guoyin Wang, Chunyuan Li, Wenlin Wang, Yizhe Zhang, Dinghan Shen, Xinyuan Zhang, Ricardo Henao, and Lawrence Carin. 2018. Joint Embedding of Words and Labels for Text Classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2321–2331.

Fisher Yu and Vladlen Koltun. 2015. Multi-scale Context Aggregation by Dilated Convolutions. *arXiv preprint arXiv:1511.07122*.

Xiao Zhang, Dejing Dou, and Ji Wu. 2020. Learning Conceptual-Contexual Embeddings for Medical Text. *AAAI Conference on Artificial Intelligence*.