This Patient Looks Like That Patient: Prototypical Networks for Interpretable Diagnosis Prediction from Clinical Text

Betty van Aken[1], Jens-Michalis Papaioannou[1], Marcel G. Naik[2],
Georgios Eleftheriadis[2], Wolfgang Nejdl[3], Felix A. Gers[1], Alexander Löser[1]

[1] Berliner Hochschule für Technik (BHT),

[2] Charité Berlin,

[3] Leibniz University Hannover

{bvanaken,michalis.papaioannou,gers,aloeser}@bht-berlin.de,
{marcel.naik,georgios.eleftheriadis}@charite.de, nejdl@L3S.de

摘要

使用深度神经模型从临床文本进行诊断预测已显示出可喜的结果。然而，在临床实践中，此类模型不仅要准确，还要为医生提供可解释和有用的结果。我们介绍了 ProtoPatient，这是一种基于原型网络和具有这两种能力的标签注意的新方法。ProtoPatient 根据与原型患者相似的文本部分进行预测——提供医生理解的理由。我们在两个公开可用的临床数据集上评估该模型，并表明它优于现有基线。与医生的定量和定性评估进一步证明该模型为临床决策支持提供了有价值的解释。

关键词：辅助诊断预测；可解释性

1 引言

医疗专业人员每天都要面对大量的文本患者信息。临床决策支持系统 (CDSS) 旨在帮助临床医生根据此类数据进行决策。我们专门研究 CDSS 的一个子任务，即根据患者入院记录预测临床诊断。当临床医生处理诊断预测任务时，他们通常会考虑（根据他们自己的经验、临床数据库或通过与同事交谈）表现出典型或非典型疾病体征的相似患者。然后，他们将手头的病人与这些以前的遭遇进行比较，并确定病人患上相同病症的风险。

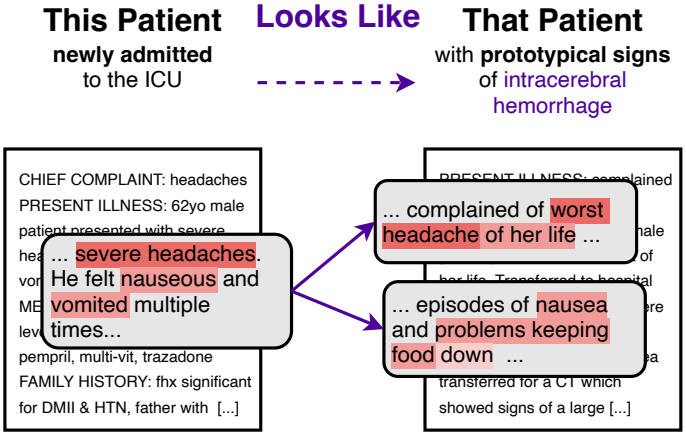


图 1: ProtoPatient 方法的基本概念。该模型根据与早期患者原型部分（右侧）的比较对患者（左侧）进行预测.

在这项工作中，我们提出了 ProtoPatient，这是一种模仿临床医生推理过程的深度神经方法：我们的模型从以前的患者那里学习诊断的原型特征，并根据与这些原型的相似性对当前患者进行预测。这引出了一个模型，该模型本身是可解释的，并为临床医生提供了指向以前原型患者的指针。我们的方法受到 Chen et al.[1] 的启发，他引入了用于图像分类的原型部分网络 (PPN)。PPN 学习图像类的原型部分，并根据与这些原型部分的相似性进行分类。我们将这项工作转移到文本域，并将其应用于诊断预测的极端多标签分类任务。对于这种转移，我们应用了一种额外的标签注意机制，通过突出显示与诊断相关的临床记录中最相关的部分，进一步提高了我们方法的可解释性。

虽然深度神经模型在过去[2] 中已广泛应用于结果预测任务，但它们的黑盒性质仍然是临床应用[3]的一大障碍。我们认为，只有当模型预测伴随着使临床医生能够遵循线索或可能放弃预测的理由时，决策支持才有可能。通过 ProtoPatient，我们引入了一种允许此类决策支持的架构。我们对公开数据的评估表明，该模型可以进一步提高预测临床结果的最新性能。

贡献　我们将这项工作的贡献总结如下：

1。我们引入了一种基于原型网络和标签注意的新型模型架构，可实现可解释的诊断预测。系统学习文本中的相关部分，并指向导致做出特定决定的原型患者。

2. 我们将我们的模型与几个最先进的基线进行比较，表明它优于早期的方法。性能提升在罕见诊断中尤为明显。

3. 我们进一步评估我们的模型提供的解释。定量结果表明，我们的模型产生的解释比事后解释更忠实于其内部工作。医生进行的人工分析进一步显示了原型患者在临床决策过程中的帮助。

2　相关工作

临床笔记的诊断预测　已经使用不同的方法研究了从临床文本预测诊断风险。Fakhraie[4] 使用词袋和词嵌入分析了临床笔记的预测价值。Jain et al.[5] 尝试将注意力模块添加到循环神经模型中。最近，使用 Transformer 模型进行诊断预测的效果优于早期方法。van Aken et al.[6] 应用基于 BERT 的模型进一步对临床病例进行预训练以预测患者结果。然而，这些模型的黑盒性质阻碍了它们在临床实践中的应用。因此，我们引入了 ProtoPatient，它使用 Transformer 表示，但提供可解释的预测。

小样本学习的原型网络　原型网络首先由 Snell et al.[7] 引入，用于小样本学习任务。他们将原型初始化为每集支持样本的质心，并将该方法应用于图像分类任务。Sun et al.[8] 将该方法应用于具有分层注意力层的文本文档。最近，基于原型网络的相关方法已被用于多个小样本文本分类任务[9-13]。与这项工作相比，我们不使用情景学习在少数情况下训练我们的模型。但是，我们的模型通过改进可用样本很少的诊断结果来显示相关功能。。

原型网络的可解释模型　Chen et al.[1]在一个不同的设置中使用了原型网络，以建立一个可解释的图像分类模型。为此，他们学习图像的原型部分来模仿人类的推理。我们调整了他们的想法，并展示了如何将其应用于临床自然语言。最近，Ming et al.[14]和 Das et al.[15]将原型网络的概念应用于文本分类，

并展示了原型文本如何帮助解释预测。与他们的工作相反，继[1]之后，我们通过使用标签关注来识别原型 textitparts 而不是整个文档。这使得解释结果更容易，并能用一千多个标签进行多标签分类。

Label-wise attention Mullenbach et al.[16]用 CAML 模型为临床文本引入了标签明智的关注。此后，该方法通过分层注意的方法得到了进一步的改进[17-19]。标签式注意力主要用于 ICD 编码，这是一项与诊断预测有关的任务，在输入数据方面有所不同。ICD 编码是在描述整个住院期间的笔记上完成的。相比之下，结果诊断预测使用入院笔记作为输入，并确定诊断 *risks* 而不是文本中已经提到的诊断。我们的方法–结合原型网络和标签明智的关注–特别关注检测和突出这些风险，以实现临床决策支持。

3 本文方法

3.1 本文方法概述

我们提出了一个新的模型架构，叫做 ProtoPatient，它通过使用标签关注和降维，将原型网络的概念[1]调整到极端的多标签场景中。图 2展示了一个示意性的概述。我们进一步展示了我们的模型如何有效地初始化以提高速度和性能。此部分对本文将要复现的工作进行概述，如图 2所示：

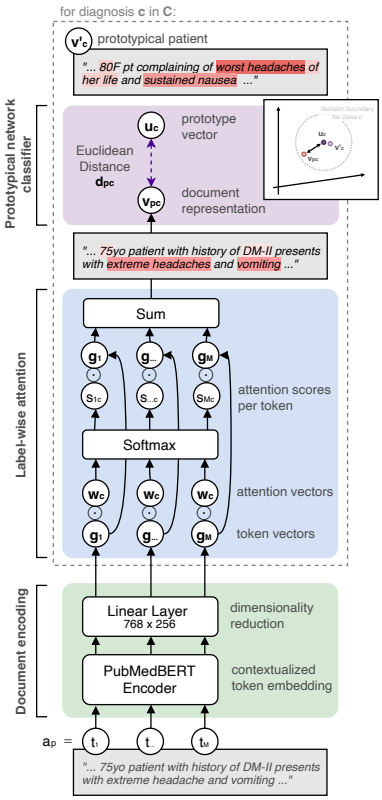


图 2: ProtoPatient 方法的示意图。从底部开始，文档标记得到了一个上下文的编码，然后被转换为一个标签式的文档表示，$\mathbf{v_{pc}}$。分类器只需考虑该表示与所学的原型向量 $\mathbf{u_c}$ 的距离。原型病人 $\mathbf{v}'_\mathbf{c}$ 是最接近原型向量的训练实例。

3.2 特征提取模块

3.3 Learning Prototypical Representations

我们将输入的文档 a_p（p 索引病人）编码为维度为 D 的向量 $\mathbf{v_p}$，并测量它们与一组学习的原型向量的距离。每个原型向量 $\mathbf{u_c}$ 代表数据集中的一个诊断 $c \in C$。原型向量是与文档编码器共同学习的，以便有诊断的病人与没有诊断的病人能够最好地区分开来。作为距离度量，我们使用欧氏距离 $d_{pc} = ||\mathbf{v_p} - \mathbf{u_c}||_2$。，该指标被认为最适合于原型网络。然后我们计算负距离的 sigmoid σ，得到预测值

$\hat{y}_{pc} = (-d_{pc})$，这样离原型向量更近的文档得到的预测分数更高。我们将损失 L 定义为 □$haty_{pc}$ 和真实标签 $y_{pc} \in \{0, 1\}$ 之间的二元交叉熵（BCE）。。

$$L = \sum_{p} \sum_{c} BCE(\hat{y}_{pc}, y_{pc}) \tag{1}$$

Prototype initialization Snell et al.[7] 将每个原型定义为嵌入支持集文档的平均值。相比之下，我们在优化多标签分类的同时，从头到尾地学习标签的原型向量。这导致了更好的原型表征，因为并不是所有的文档都能平等地代表一个类别，就像取平均值那样。然而，使用所有 support document 的平均值是一个合理的起点。我们将一个类别的初始原型向量设定为 $\mathbf{u}_{\mathbf{c}_{\mathbf{init}}} = \langle \mathbf{v_c} \rangle$, i.e. 即训练集中具有类标签 c 的所有文件向量 $\mathbf{v_c}$ 的平均值。然后我们在训练过程中对其表示进行微调。最初的实验表明，与随机初始化相比，这种初始化导致模型收敛的步骤只有一半。

Contextualized document encoder 对于文档的编码，我们选择了一个基于 Transformer 的模型，因为 Transformer 能够对上下文的标记表示进行建模。为了初始化文档编码器，我们使用预先训练过的语言模型的权重。在我们实验的时候，PubMedBERT[20]模型在一系列生物医学 NLP 任务上达到了最好的结果。因此，我们用 PubMedBERT 的权重初始化我们的文档编码器[1] ，并在训练期间用小的学习率进一步优化它。

3.4 Encoding Relevant Document Parts with Label-wise Attention

由于我们面临的是一个多标签问题，每个文件只有一个联合表示，往往会产生位于向量空间中多个原型中心的文件向量。这样一来，单一诊断的重要特征就会变得模糊，特别是当这些诊断很罕见的时候。为了防止这种情况，我们遵循原型部分网络的想法，选择笔记中对某一诊断有意义的部分。与 Chen et al.[1]相比，我们使用基于注意力的方法，而不是卷积过滤器，因为注意力是选择文本相关部分的有效方法。对于每个诊断 c，我们学习一个注意力向量 $\mathbf{w_c}$。为了编码与 c 有关的病人笔记，我们在 $\mathbf{w_c}$ 和每个嵌入的标记 $\mathbf{g_{pj}}$ 之间进行点乘，其中 j 是标记索引。然后，我们应用一个 softmax。

$$s_{pcj} = softmax(\mathbf{g_{pj}^T}\, \mathbf{w_c}) \tag{2}$$

我们使用所得到的分数 s_{pcj} 来创建一个文档表征 $\mathbf{v_{pc}}$ 作为标记向量的加权和。

$$\mathbf{v_{pc}} = \sum_{j} s_{pcj}\, \mathbf{g_{pj}} \tag{3}$$

这样，某项诊断的文件表述是基于与该诊断最相关的部分。然后，我们测量 $d_{pc} = ||\mathbf{v_{pc}} - \mathbf{u_c}||_2$ 与原型向量 $\mathbf{u_c}$ 的距离, 基于特定于诊断的文档表述

3.5 损失函数定义

$$L = \sum_{p} \sum_{c} BCE(\hat{y}_{pc}, y_{pc}) \tag{4}$$

[1]Model weights from: https://huggingface.co/microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext

4 复现细节

4.1 与已有开源代码对比

参考了原作者的代码以 pytorch_lighting 模块的框架重新实现，并进行了改进, 加入了对比学习表征模块。

4.2 实验环境搭建

Python 版本为 3.7-3.9

其他所需要的 Package:

scikit-learn==0.23.2

pandas==1.1.4

tensorboard==2.7.0

fire==0.4.0

matplotlib==3.5.1

pytorch-lightning==1.4.9

transformers==4.11.2

torchmetrics==0.6.2

4.3 界面分析与使用说明

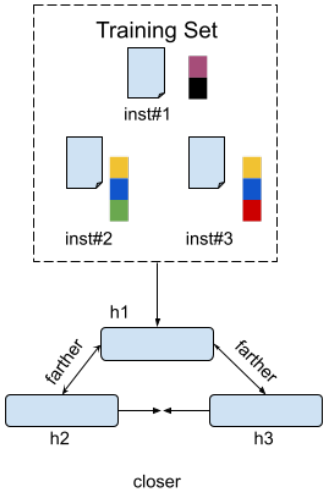
采用命令行进行训练。python training.py 通过更改命令行的指令如:–gpus 和–train 等更改所需要 gpu 参数和训练文件夹的位置。Pytorclighting 的框架使我们能够看到训练时候的状况以及测试的时候详细指标。

```
python training.py
            --model_type PROTO
            --train_file {TRAIN.csv}
            --val_file {VAL.csv}
            --test_file {TEST.csv}
            --num_warmup_steps 5000
            --num_training_steps 5000
            --lr_features 0.000005
            --lr_prototypes 0.001
            --lr_others 0.001
            --use_attention True
            --reduce_hidden_size 256
            --all_labels_path {ALL_LABELS.txt}
```

4.4 创新点

基于 Prototype 的分类学习相当于一个度量学习如 KNN, 最近邻算法等，加入对比学习让实例的特征在特征空间中得到更好的聚类特征，当实例的特征获得更好的聚类特征时，Prototype 会在特征空间中获得更好的表示即形成更好的聚类中心。在训练过程中 attention 向量会占用大量内存，为了检测我们的对比学习算法的可行性，我们取消了 attention 模块并将其训练的 batch size 增大。现有的监督对比学习方法试图缩小同一类别的实例之间的距离，并推开不同类别的实例。然而，在 MLTC 中，两个实例可能共享一些共同的标签，同时也可能有一些标签是每个实例所独有的。如何处理这些情况是在 MLTC 中利用对比性学习的关键。因此，为了模拟多标签实例之间的复杂关联，我们设计了一个基

于标签相似性的动态系数基于标签的相似性。我们的做法直观如图所示:



考虑到一个大小为 b 的数据 minibatch，我们定义一个函数来输出特定实例的所有其他实例，对于一个特例来说:

$$i : g(i) = \{k | k \in \{1, 2, \cdots, b\}, k \neq i\}$$

每个实例对 (i, j) 的对比损失可计算为:

$$C_{con}^{ij} = -\beta_{ij} \log \frac{e^{-d(z_i;_i z_j)/\tau'}}{\sum_{k \in g(i)} e^{-d(z_i, z_k)}/\tau'} \tag{5}$$

$$C_{ij} = y_i^T \cdot y_j, \beta_{ij} = \frac{C_{ij}}{\sum_{k \in g(i)} C_{ik}} \tag{6}$$

5 实验结果分析

本部分对实验所得结果进行分析，详细对实验内容进行说明，实验结果进行描述并分析。可以看到 ProtoPatient 比纯 PubMebBERT 和纯 Prototype(不加 Label-attention) 的分类方法在三个指标上都达到超越的效果。这充分说明了 Label-Attention 的重要性，但是实验过程中也发现了该框架的巨大问题就是内存占用过大的问题，这使得在实际场景的应用有着巨大的挑战。那么我们的目标也变得更为明确就是做出一个可解释性，性能尚可，能够符合应用场景的深度学习模型。消融研究表明，所有组件都在改善结果方面发挥作用。没有标签关注的原型网络无法捕获极端的多标签数据。使用标准分类头的 PubMedBERT 也受益于标签关注，但程度不同。因此，将原型网络和标签式关注相结合会带来额外的好处。尺寸大小的选择是另一个重要因素。使用 768 维（标准 BERT 基本尺寸）似乎会导致注意力和原型向量的过度参数化 Label-wise attention

	ROC AUC macro	ROC AUC micro	PR AUC macro
PubMedBERT	83.48 ±0.21	95.47 ±0.22	13.42 ±0.57
Prototypical Network	81.89 ±0.22	95.23 ±0.01	9.94 ±0.36
ProtoPatient	86.10 ±0.24	**97.23** ±0.00	**22.72** ±0.21

表 1: 基于 MIMIC-III 数据，诊断预测任务（1266 个标签）的 AUC 为%。ProtoPatient 模型在微 ROC AUC 和 PR AUC 方面优于基线.

	ROC AUC macro
Dimensionality reduction	
ProtoPatient 768	83.56 ±0.17
ProtoPatient (our proposed model with D=256)	**86.93** ±0.24
Transformer vs. Prototypical	
PubMedBERT 768	83.48 ±0.21
PubMedBERT 768 + Label Attention	**84.10** ±0.25
ProtoPatient 768	83.56 ±0.17
Label-wise attention	
PubMedBERT 256	83.61 ±0.04
PubMedBERT 256 + Label Attention	**84.68** ±0.52

表 2: **Ablation studies** 比较不同尺寸的变压器以及标准变压器（PubMedBERT）的性能，并验证了 Label attention 的有效性.

6 总结与展望

复现了 ProtoPatient 这一论文的结果并且对下一步改进有了比较明确的方向，接下来就是根据自己的 idea 进行模型的改进。

参考文献

[1] CHEN C, LI O, TAO D, et al. This Looks Like That: Deep Learning for Interpretable Image Recognition [C/OL]//WALLACH H M, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. 2019: 8928-8939. https://proceedings.neurips.cc/paper/2019/hash/adf7ee2dcf142b0e11888e72b43fcb75-Abstract.html.

[2] SHAMOUT F, ZHU T, CLIFTON D A. Machine learning for clinical outcome prediction[J]. IEEE reviews in Biomedical Engineering, 2020, 14: 116-126.

[3] Van AKEN B, HERRMANN S, LÖSER A. What Do You See in this Patient? Behavioral Testing of Clinical NLP Models[C/OL]//Proceedings of the 4th Clinical Natural Language Processing Workshop. Seattle, WA: Association for Computational Linguistics, 2022: 63-73. https://aclanthology.org/2022.clinicalnlp-1.7. DOI: 10.18653/v1/2022.clinicalnlp-1.7.

[4] FAKHRAIE N. What's in a Note? Sentiment Analysis in Online Educational Forums[M]. University of Toronto (Canada), 2011.

[5] JAIN S, MOHAMMADI R, WALLACE B C. An Analysis of Attention over Clinical Notes for Predictive Tasks[C/OL]//Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics, 2019: 15-21. https://aclanthology.org/W19-1902. DOI: 10.18653/v1/W19-1902.

[6] Van AKEN B, PAPAIOANNOU J M, MAYRDORFER M, et al. Clinical Outcome Prediction from Admission Notes using Self-Supervised Knowledge Integration[C/OL]//Proceedings of the 16th Confer-

ence of the European Chapter of the Association for Computational Linguistics: Main Volume. Online: Association for Computational Linguistics, 2021: 881-893. https://aclanthology.org/2021.eacl-main.75.

[7] SNELL J, SWERSKY K, ZEMEL R S. Prototypical Networks for Few-shot Learning[C/OL]//GUYON I, von LUXBURG U, BENGIO S, et al. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. 2017: 4077-4087. https://proceedings.neurips.cc/paper/2017/hash/cb8da6767461f2812ae4290eac7cbc42-Abstract.html.

[8] SUN S, SUN Q, ZHOU K, et al. Hierarchical Attention Prototypical Networks for Few-Shot Text Classification[C/OL]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China: Association for Computational Linguistics, 2019: 476-485. https://aclanthology.org/D19-1045. DOI: 10.18653/v1/D19-1045.

[9] WEN W, LIU Y, OUYANG C, et al. Enhanced prototypical network for few-shot relation extraction [J/OL]. Inf. Process. Manag., 2021, 58(4): 102596. https://doi.org/10.1016/j.ipm.2021.102596. DOI: 10.1016/j.ipm.2021.102596.

[10] ZHANG J, ZHU J, YANG Y, et al. Knowledge-Enhanced Domain Adaptation in Few-Shot Relation Classification[C/OL]//ZHU F, OOI B C, MIAO C. KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021. ACM, 2021: 2183-2191. https://doi.org/10.1145/3447548.3467438. DOI: 10.1145/3447548.3467438.

[11] REN H, CAI Y, CHEN X, et al. A Two-phase Prototypical Network Model for Incremental Few-shot Relation Classification[C/OL]//SCOTT D, BEL N, ZONG C. Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020. International Committee on Computational Linguistics, 2020: 1618-1629. https://doi.org/10.18653/v1/2020.coling-main.142. DOI: 10.18653/v1/2020.coling-main.142.

[12] DENG S, ZHANG N, KANG J, et al. Meta-Learning with Dynamic-Memory-Based Prototypical Network for Few-Shot Event Detection[C/OL]//CAVERLEE J, HU X (, LALMAS M, et al. WSDM '20: The Thirteenth ACM International Conference on Web Search and Data Mining, Houston, TX, USA, February 3-7, 2020. ACM, 2020: 151-159. https://doi.org/10.1145/3336191.3371796. DOI: 10.1145/3336191.3371796.

[13] FENG J, WEI Q, CUI J. Prototypical networks relation classification model based on entity convolution [J/OL]. Comput. Speech Lang., 2023. https://doi.org/10.1016/j.csl.2022.101432. DOI: 10.1016/j.csl.2022.101432.

[14] MING Y, XU P, QU H, et al. Interpretable and Steerable Sequence Learning via Prototypes[C/OL]//TEREDESAI A, KUMAR V, LI Y, et al. Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019.

ACM, 2019: 903-913. https://doi.org/10.1145/3292500.3330908. DOI: 10.1145/3292500.3330908.

[15] DAS A, GUPTA C, KOVATCHEV V, et al. ProtoTEx: Explaining Model Decisions with Prototype Tensors[C/OL]//MURESAN S, NAKOV P, VILLAVICENCIO A. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022. Association for Computational Linguistics, 2022: 2986-2997. https://doi.org/10.18653/v1/2022.acl-long.213. DOI: 10.18653/v1/2022.acl-long.213.

[16] MULLENBACH J, WIEGREFFE S, DUKE J, et al. Explainable Prediction of Medical Codes from Clinical Text[C/OL]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, Louisiana: Association for Computational Linguistics, 2018: 1101-1111. https://aclanthology.org/N18-1100. DOI: 10.18653/v1/N18-1100.

[17] BAUMEL T, NASSOUR-KASSIS J, COHEN R, et al. Multi-label classification of patient notes: case study on ICD code assignment[C]//Workshops at the thirty-second AAAI conference on artificial intelligence. 2018.

[18] YANG Z, YANG D, DYER C, et al. Hierarchical Attention Networks for Document Classification [C/OL]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, California: Association for Computational Linguistics, 2016: 1480-1489. https://aclanthology.org/N16-1174. DOI: 10.18653/v1/N16-1174.

[19] DONG H, SUÁREZ-PANIAGUA V, WHITELEY W, et al. Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation[J]. Journal of biomedical informatics, 2021.

[20] TINN R, CHENG H, GU Y, et al. Fine-Tuning Large Neural Language Models for Biomedical Natural Language Processing[J/OL]. CoRR, 2021, abs/2112.07869. arXiv: 2112.07869. https://arxiv.org/abs/2112.07869.