Ontological attention ensembles for capturing semantic concepts in ICD code prediction from clinical text

Matúš Falis[1]*, Maciej Pajak[1]*, Aneta Lisowska[1]*, Patrick Schrempf[1,3], Lucas Deckers[1], Shadia Mikhael[1], Sotirios A. Tsaftaris[1,2] and Alison Q. O'Neil[1,2]
[1] Canon Medical Research Europe, Edinburgh, UK
[2] University of Edinburgh, Edinburgh, UK
[3] University of St Andrews, St Andrews, UK
{matus.falis, maciej.pajak, aneta.lisowska}@eu.medical.canon,
{patrick.schrempf, lucas.deckers, shadia.mikhael}@eu.medical.canon,
s.tsaftaris@ed.ac.uk, alison.oneil@eu.medical.canon

Abstract

We present a semantically interpretable system for automated ICD coding of clinical text documents. Our contribution is an ontological attention mechanism which matches the structure of the ICD ontology, in which shared attention vectors are learned at each level of the hierarchy, and combined into label-dependent ensembles. Analysis of the attention heads shows that shared concepts are learned by the lowest common denominator node. This allows child nodes to focus on the differentiating concepts, leading to efficient learning and memory usage. Visualisation of the multi-level attention on the original text allows explanation of the code predictions according to the semantics of the ICD ontology. On the MIMIC-III dataset we achieve a 2.7% absolute (11% relative) improvement from 0.218 to 0.245 macro-F1 score compared to the previous state of the art across 3,912 codes. Finally, we analyse the labelling inconsistencies arising from different coding practices which limit performance on this task.

1 Introduction

Classification of clinical free-text documents poses some difficult technical challenges. One task of active research is the assignment of diagnostic and procedural International Classification of Diseases (ICD) codes. These codes are assigned retrospectively to hospital admissions based on the medical record, for population disease statistics and for reimbursements for hospitals in countries such as the United States. As manual coding is both time-consuming and error-prone, automation of the coding process is desirable. Coding errors may result in unpaid claims and loss of revenue (Adams et al., 2002).

Automated matching of unstructured text to medical codes is difficult because of the large number of possible codes, the high class imbalance in the data, and the ambiguous language and frequent lack of exposition in clinical text. However, the release of large datasets such as MIMIC-III (Johnson et al., 2016) has paved the way for progress, enabling rule-based systems (Farkas and Szarvas, 2008) and classical machine learning methods such as support vector machines (Suominen et al., 2008), to be superseded by neural network-based approaches (Baumel et al., 2017; Karimi et al., 2017; Shi et al., 2018; Duarte et al., 2018; Rios and Kavuluru, 2018). The most successful reported model on the ICD coding task is a shallow convolutional neural network (CNN) model with label-dependent attention introduced by Mullenbach et al. (2018) and extended by Sadoughi et al. (2018) with multi-view convolution and a modified label regularisation module.

One of the common features of the aforementioned neural network models is the use of attention mechanisms (Vaswani et al., 2017). This mirrors advances in general representation learning. In the text domain, use of multi-headed attention has been core to the development of *Transformer*-based language models (Devlin et al., 2018; Radford et al., 2019). In the imaging domain, authors have had success with combining attention vectors learned at the global and local levels with *Double Attention* networks (Chen et al., 2018). In the domain of structured (coded) medical data, Choi et al. (2017) leveraged the ontological structure of the ICD and SNOMED CT coding systems in their *GRAM* model, to combine the attention vectors of a code and its ancestors in order to predict the codes for the next patient visit based on the codes assigned in the previous visit.

Our contributions are:

1. A structured ontological attention ensemble mechanism which provides improved accuracy, efficiency, and interpretability.

*equal contribution

Proceedings of the 10th International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pages 168–177

https://doi.org/10.18653/v1/D19-62

Dataset	# Documents	# Unique patients	# ICD-9 Codes	# Unique ICD-9 codes
Training	47,719	36,997	758,212	8,692
Development	1,631	1,374	28,896	3,012
Test	3,372	2,755	61,578	4,085
Total	**52,722**	**41,126**	**848,686**	**8,929**

Table 1: Distribution of documents and codes in the MIMIC-III dataset.

2. An analysis of the multi-level attention weights with respect to the text input, which allows us to interpret the code predictions according to the semantics of the ICD ontology.
3. An analysis of the limitations of the MIMIC-III dataset, in particular the labelling inconsistencies arising from variable coding practices between coders and between timepoints.

2 Dataset

We used the MIMIC-III dataset (Johnson et al., 2016) ("Medical Information Mart for Intensive Care") which comes from the intensive care unit of the Beth Israel Deaconess Medical Center in Boston. We concatenated the hospital discharge summaries associated with each admission to form a single document and combined the corresponding ICD-9 codes. The data was split into training, development, and test patient sets according to the split of Mullenbach et al. (2018) (see Table 1).

3 Methods

We formulate the problem as a multi-label binary classification task, for which each hospital discharge summary is labelled with the presence or absence of the complete set of ICD-9 codes for the associated admission. Our model is a CNN similar to those of (Mullenbach et al., 2018; Sadoughi et al., 2018). Inspired by the graph-based attention model of (Choi et al., 2017), we propose a hierarchical attention mechanism (mirroring the ICD ontology) which yields a multi-level, label-dependent ensemble of attention vectors for predicting each code. Our architecture is shown in Figure 1 and described below.

3.1 Embedding

Documents were pre-processed by lower-casing the text and removing punctuation, followed by tokenisation during which purely numeric tokens were discarded. We used a maximum input length of 4500 tokens and truncated any documents longer than this (260 training, 16 development, and 22 test). Tokens were then embedded with a 100-dimensional word2vec model. For each document, token embeddings were concatenated to give a $100 \times N$ document embedding matrix D, where N is the document length.

We pre-trained the word2vec model on the training set using continuous bag-of-words (CBOW) (Mikolov et al., 2013). The vocabulary comprises tokens which occur in at least 3 documents (51,847 tokens). The embedding model was fine-tuned (not frozen) during subsequent supervised training of the complete model.

3.2 Convolutional module

The first part of the network proper consists of a multi-view convolutional module, as introduced by Sadoughi et al. (2018). Multiple one-dimensional convolutional kernels of varying size with stride = 1 and weights W are applied in parallel to the document embedding matrix D along the N dimension. The outputs of these kernels are padded at each end to match the input length N. This yields outputs of size $C \times M \times N$ where C is the number of kernel sizes ("views"), M is the number of filter maps per view, and N is the length of the document. The outputs are max-pooled in the C dimension i.e., across each set of views, to yield a matrix E of dimensions $M \times N$:

$$E = \tanh(\max_{C=[0,3]} W_C * D) \quad (1)$$

Optimal values were $C = 4$ filters of lengths $\{6, 8, 10, 12\}$ with $M = 256$ filter maps each.

3.3 Prediction via label-dependent attention

Label-specific attention vectors are employed to collapse the variable-length E document representations down to fixed-length representations. For each label l, given the matrix E as input, a token-wise linear layer u_l is trained to generate a vector of length N. This is normalised with a softmax operation, resulting in an attention vector a_l:

$$a_l = \text{softmax}(E^T u_l) \quad (2)$$

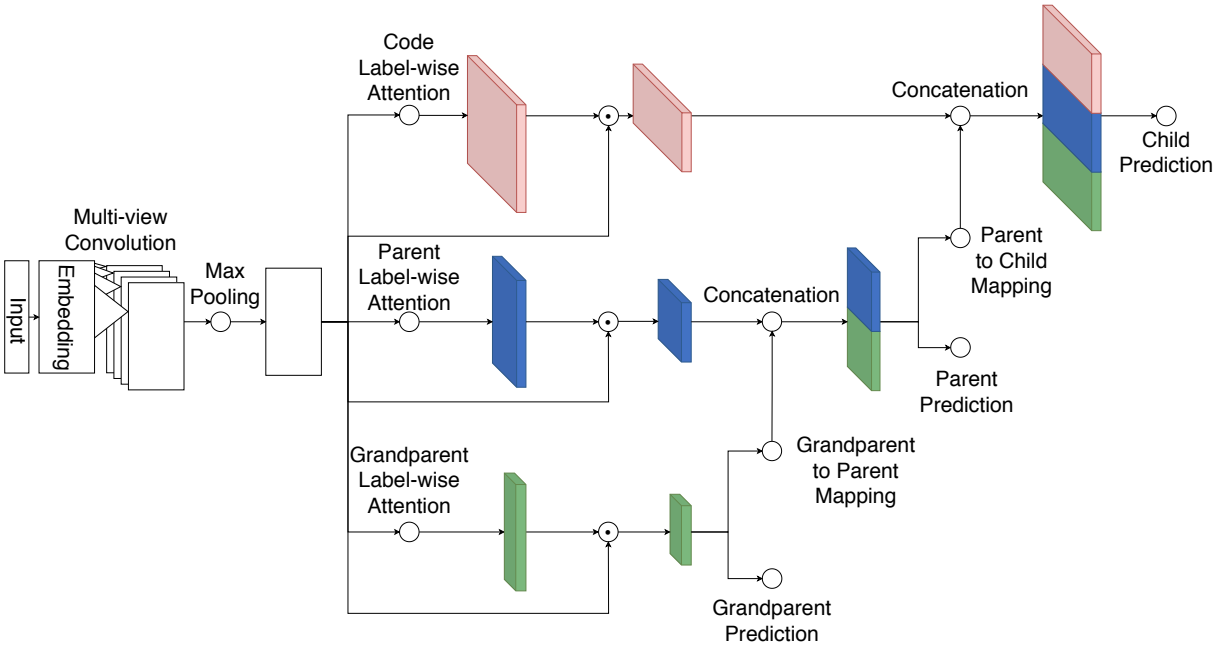

Figure 1: Network architecture. The output of the convolutional module is fed into the ensemble of ancestral attention heads for multi-task learning. Circles with dots represent matrix product operations. Ancestors are mapped to descendants by multiplication with a mapping connectivity matrix based on the ontology structure.

The attention vector is then multiplied with the matrix E which yields a vector v_l of length M, a document representation specific to a label:

$$v_l = a_l E \tag{3}$$

If multiple linear layers $u_{l,0}, u_{l,1}, \ldots$ are trained for each label at this stage, multiple attention vectors (or "heads") will be generated. Thus, multiple document representations v_l could be made available, each of length M, and concatenated together to form a longer label-specific representation for the document. We experimented with multiple attention vectors and found two vectors per label to be optimal. To make a prediction of the probability of each label, $P(l)$, there is a final dense binary classification layer with sigmoid activation. This is shown for two attention vectors:

$$P(l) = \sigma(W_l[v_{l,0}; v_{l,1}] + \beta_l) \tag{4}$$

3.4 Prediction via label-dependent ontological attention ensembles

The ICD-9 codes are defined as an ontology, from more general categories down to more specific descriptions of diagnosis and procedure. Rather than simply training two attention heads per code as shown in Section 3.3, we propose to exploit the ontological structure to train shared attention heads between codes on the same branch of the tree, thus pooling information across labels which share ancestry. In this work, we use two levels of ancestry, where the first level corresponds to the pre-floating-point portion of the code. For instance, for the code *425.11 Hypertrophic obstructive cardiomyopathy*, the first-degree ancestor is *425 Cardiomyopathy* and the second-degree ancestor is *420-429 Other forms of heart disease* (the chapter in which the parent occurs). This is illustrated in Figure 2.

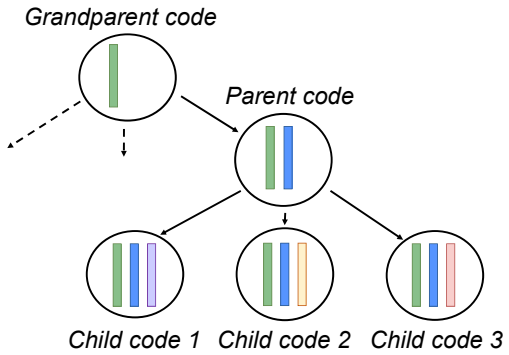

Figure 2: Illustration of inheritance of the linear layers u_l. This yields label-specific ontological attention ensembles of the attention heads a_l and subsequently the document representations v_l.

For the entire set of 8929 labels, we identi-

fied 1167 first-degree ancestors and 179 second-degree ancestors. Compared to two attention vectors per code, this reduces the parameter space and memory requirements from 17,858 attention heads (8929 x 2) to 10,275 attention heads (8929 + 1167 + 179) as well as increasing the number of training samples for each attention head.

The label prediction for each code is now derived from the concatenated child (c), parent (p) and grandparent (gp) document representations:

$$P(l_{\text{child}}) = \sigma(W_{l_c}[v_{l,c}; v_{l,p}; v_{l,gp}] + \beta_{l_c}) \quad (5)$$

In order to facilitate learning of multiple attention heads, we employ deep supervision using the ancestral labels, adding auxiliary outputs for predicting the parent and grandparent nodes:

$$P(l_{\text{parent}}) = \sigma(W_{l_p}[v_{l,p}; v_{l,gp}] + \beta_{l_p}) \quad (6)$$

$$P(l_{\text{grandparent}}) = \sigma(W_{l_{gp}}[v_{l,gp}] + \beta_{l_{gp}}) \quad (7)$$

3.5 Training process

We trained our model with weighted binary cross entropy loss using the Adam optimiser (Kingma and Ba, 2014) with learning rate 0.0005.

Stratified shuffling: The network accepts input of any length but all instances within a single batch need to be padded to the same length. To minimise the amount of padding, we used length-stratified shuffling between epochs. For this, documents were grouped by length and shuffled only within these groups; groups were themselves then shuffled before batch selection started.

Dampened class weighting: We employed the standard practice of loss weighting to prevent the imbalanced dataset from affecting performance on rare classes. We used a softer alternative to empirical class re-weighting, by taking the inverse frequencies of positive (label= 1) and negative (label= 0) examples for each code c, and adding a damping factor α. In the equations below, $n_{label_c=1}$ stands for the number of positive examples for the ICD code c, and n stands for the total number of documents in the dataset.

$$\begin{aligned} \omega_{(c,1)} &= \left(\frac{n}{n_{\text{label}_c=1}}\right)^{\alpha} \\ \omega_{(c,0)} &= \left(\frac{n}{n_{\text{label}_c=0}}\right)^{\alpha} \end{aligned} \quad (8)$$

Upweighting for codes with 5 examples or fewer, where we do not expect to perform well in any case, was removed altogether as follows:

$$\omega_{(c,1)} = \begin{cases} \left(\frac{n}{n_{\text{label}_c=1}}\right)^{\alpha} & , n_{\text{label}_c=1} > 5 \\ 1 & , \text{otherwise} \end{cases} \quad (9)$$

Deep supervision: The loss function was weighted in favour of child codes, with progressively less weight given to the codes at higher levels in the ICD ontology. A weighting of 1 was used for the child code loss, a weighting w_h for the parent code auxiliary loss, and w_h^2 for the grandparent code auxiliary loss, i.e.,

$$\text{Loss} = L_{\text{c}} + w_h L_{\text{p}} + w_h^2 L_{\text{gp}} \quad (10)$$

Optimal values were $\alpha = 0.25$ and $w_h = 0.1$.

3.6 Implementation and hyperparameters

The word2vec embedding was implemented with Gensim (Řehůřek and Sojka, 2010) and the ICD coding model was implemented with PyTorch (Paszke et al., 2017). Experiments were run on Nvidia V100 16GB GPUs. Hyperparameter values were selected by maximising the development set macro-F1 score for codes with more than 5 training examples.

4 Experiments

4.1 Results

In our evaluation, we focus on performance across all codes and hence we prioritise macro-averaged metrics, in particular macro-averaged precision, recall, and F1 score. Micro-averaged F1 score and Precision at k ($P@K$) are also reported in order to directly benchmark performance against previously reported metrics. All reported numbers are the average of 5 runs, starting from different random network initialisations.

We compare our model to two previous state-of-the-art models: Mullenbach et al. (2018), and Sadoughi et al. (2018) (published only on arXiv). We trained these models with the hyperparameter values quoted in the respective publications, and used the same early stopping criteria as for our model. Both Mullenbach *et al.* and Sadoughi *et al.* use label regularisation modules, at the output and at the attention layer respectively. In line with their published results, we found that only the method of Sadoughi *et al.* gave an improvement and thus it

Method	R_{macro}	P_{macro}	$F1_{macro}$	$F1_{micro}$	$P@8$
Mullenbach et al. (2018)	0.218	0.195	0.206	0.499	0.651
Sadoughi et al. (2018)	0.261	0.186	0.218	0.498	0.662
Ontological Attention	0.341	0.192	**0.245**	0.497	0.681

Table 2: Benchmark results for the models trained with $F1_{macro}$ stopping criterion.

Method	R_{micro}	P_{micro}	$F1_{macro}$	$F1_{micro}$	$P@8$
Mullenbach et al. (2018)	0.469	0.593	0.172	0.523	0.685
Sadoughi et al. (2018)	0.516	0.560	0.173	0.537	0.695
Ontological Attention	0.514	0.617	0.206	**0.560**	0.727

Table 3: Benchmark results for the models trained with $F1_{micro}$ stopping criterion.

Method	$F1_{macro}$	Relative $F1_{macro}$ change (%)
Ontological Attention	**0.245**	0
Efficacy of ontological attention ensemble		
1. No deep supervision	0.243	-0.82
2. No ontology: One attention head for each label	0.234	-4.5
3. No ontology: Two attention heads for each label	0.242	-1.2
4. Partial ontology: Randomised ontological connections	0.231	-5.7
Efficacy of additional modifications		
5. No class weighting	0.232	-5.3
6. Reduced convolutional filters (70, as in Sadoughi et al. (2018))	0.236	-3.7

Table 4: Ablation study of individual components of the final method. All models are trained with the $F1_{macro}$ stopping criterion. Experiments 2 and 3 do not use the ontological attention mechanism, and instead have one or two attention heads respectively per code-level label. For experiment 4, child-parent and parent-grandparent connections were randomised, removing shared semantics between codes across the full 3 levels.

is included in the model reported here. However, this regularisation is not used in our own model where we observed no benefit.

Overall results are shown in Table 2. Our method significantly outperforms the benchmarks on macro-F1 and $P@8$.

Previous models have optimised for F1 micro-average. Different target metrics require different design choices: after removal of the class weighting in the loss function and when using $F1_{micro}$ as our stopping criterion, we are also able to surpass previous state-of-the-art results on micro-F1. The results are presented in Table 3; our method achieves the highest $F1_{micro}$ score, as well as the highest $P@8$ score. We note that $P@8$ score is consistently higher for models stopped using the $F1_{micro}$ criterion.

In Table 4 we present an ablation study. It can be seen that the improvement in performance of the ontological attention model is not simply due to increased capacity of the network, since even with 73% greater capacity (17,858 compared to 10,275 attention vectors), the two-vector multi-headed model has a 1.2% drop in performance. Experiments with deep supervision and randomisation of the ontology graph connections show the benefit of each component of the ontological architecture. We also measure the effect of additional changes made during optimisation of the architecture and training.

Levels of the ontology: Three levels of the ontology (including the code itself) were found to be optimal for the Ontological Attention model (see Figure 3). Adding parent and grandparent levels provide incremental gains in accuracy. Adding a level beyond the grandparent node (i.e., the great-grandparent level) does not provide further improvement. Since we identified only 22 ancestral nodes at the level directly above the grandparent, we hypothesise that the grouping becomes too coarse to be beneficial. In fact, all procedure codes share the same ancestor at this level; the remaining

21 nodes are split between diagnostic codes.

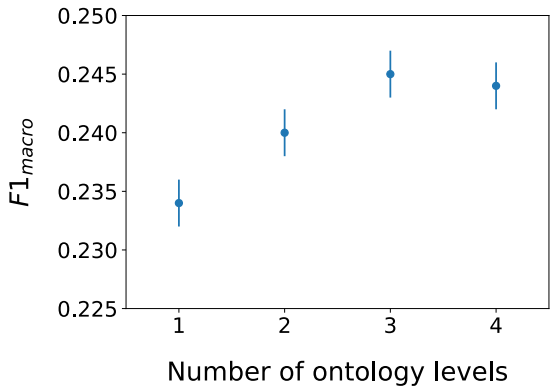

Figure 3: $F1_{macro}$ for models using attention ensembles across different levels of the ontological tree. Error bars represent the standard deviation across 5 different random weight initialisations. The model with 1 level has only the code-level attention head, the model with 2 levels also includes the shared parent attention heads; the model with 3 levels adds the shared grandparent attention heads (this is our reported Ontological Attention model), and finally, the model with 4 levels adds shared great-grandparent attention heads.

4.2 Analysis of the attention weights

In Figure 4 we show how the weights of code-level u_l vectors (which give rise to the attention heads) change when the ontological attention ensemble mechanism is introduced. As expected, we observe that in the case of a single attention head, the weights for different codes largely cluster together based on their position in the ontology graph. Once the parent and grandparent attention heads are trained, the ontological similarity structure on the code level mostly disappears. This suggests that the common features of all codes within a parent group are already extracted by the parent attention. thus, the capacity of the code-level attention is spent on the representation of the differences between the descendants of a single parent.

4.3 Interpretability of the attention heads

In Section 4.2, we showed the links between the ontology and the attention heads within the space of the u_l vector weights. We can widen this analysis to links between the predictions and the input, by examining which words in the input documents are attended by the three levels of attention heads for a given label. A qualitative visual example is shown in Figure 5. We performed quantitative frequency analysis of high-attention terms (keywords) in the training set. A term was considered a keyword if its attention weight in a document surpassed the threshold t_{kw}:

$$t_{kw}(N, \gamma_{kw}) = \gamma_{kw} \frac{1}{N}, \tag{11}$$

where N is the length of a document and γ_{kw} is a scalar parameter controlling the strictness of the threshold. With $\gamma_{kw} = 1$, a term is considered a keyword if its attention weight surpasses the uniformly distributed attention. In our analysis we chose $\gamma_{kw} = 17$ for all documents.

We aggregated these keywords across all predicted labels in the training set, counting how many times a term is considered a keyword for a label. The results of this analysis are in line with our qualitative analysis of attention maps. The most frequent keywords for the labels presented in the example in Figure 5 include "cancer", "ca", "tumor", at the grandparent level (focusing on the concept of cancer); "metastatic", "metastases" and "metastasis" at the parent level (focusing on the concept of metastasis); and "brain", "craniotomy", "frontal" at the code-level (focusing on terms relating to specific anatomy). A sibling code (*198.5 Secondary malignant neoplasm of bone and bone marrow*) displays similar behaviour in focusing on anatomy, with "bone", "spine", and "back" being among the most frequent keywords.

Not all codes display such structured behaviour. For instance, the grandparent *401-405 Hypertensive disease* attended to the term "hypertension" most frequently. The parent code *401 Essential hypertension*, does not attend to "hypertension", but neither does it attend to any useful keywords — this may be due to the code being simple compared to its sibling codes, which are more specific (e.g., *402 Hypertensive heart disease*). Interestingly, the children of *401 Essential hypertension* attend to the word "hypertension" again, while also focusing on terms that set them apart from each other — e.g., *401.0 Malignant essential hypertension* focuses on terms implying malignancy, such as "urgency", "emergency", and "hemorrhage".

5 Limitations due to labelling variability

Since performance on this task appears to be much lower than might be acceptable for real-world use, we investigated further. Figure 6 shows the per-label F1 scores; it can be seen that there is high

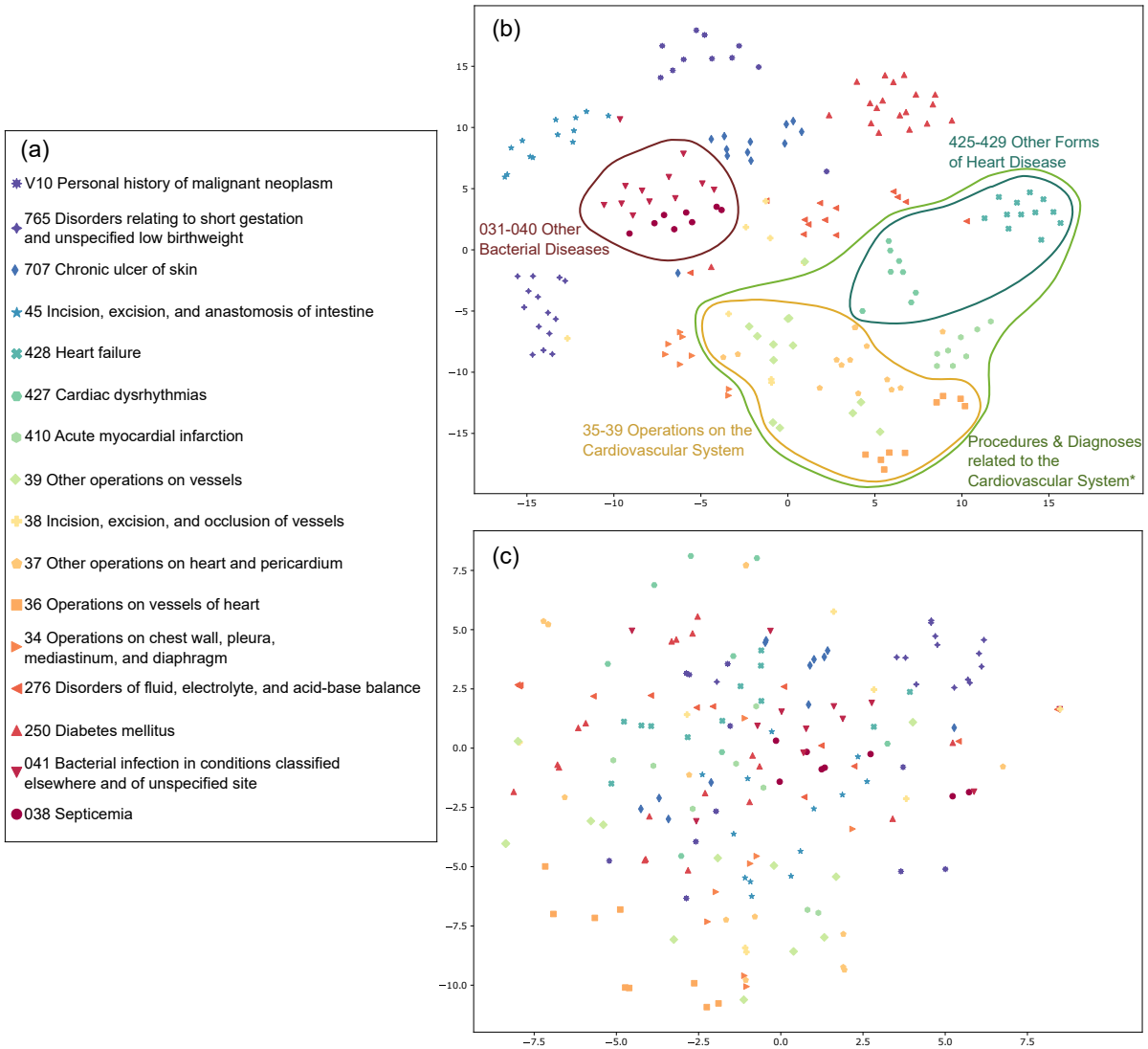

Figure 4: Two-dimensional t-SNE (Maaten and Hinton, 2008) representation of the u_l vectors (which give rise to the attention heads) for a subset of 182 codes with at least 100 occurrences each, in the data belonging to 16 different parent nodes. (a) Legend for the annotation of data points according to their parent node. (b) u_l vectors from the model with only a single attention head for each code (i.e., no ontology). It can be seen that codes naturally cluster by their parent node. Selected higher-level alignments are indicated by additional contours — for grandparent nodes (3 nodes) and for diagnoses/procedure alignment (in the case of cardiovascular disease). (c) u_l vectors in the ontological attention ensemble model for the same set of codes (and the same t-SNE hyperparameters). In most cases the clustering disappears, indicating that the attention weights for the ancestral codes have extracted the similarities from descendants' clusters.

Method	R_{macro}	P_{macro}	$F1_{macro}$	$F1_{micro}$	$P@8$
Mullenbach et al. (2018)	0.226	0.200	0.212	0.500	0.651
Sadoughi et al. (2018)	0.272	0.187	0.222	0.497	0.662
Ontological Attention	0.347	0.199	**0.252**	0.507	0.686

Table 5: Benchmark results for the models trained with $F1_{macro}$ stopping criterion.

variability in accuracy, that is only partially correlated with the number of training examples.

Inspection of examples for some of the poorly performing codes revealed some variability in coding policy, described further below.

5.1 Misreporting of codes

The phenomenon of human coding errors is reported in the literature; for instance, Kokotailo

(a) Brief Hospital Course:
Mr. John Doe is a 68-year-old male with metastatic NSCLC and brain metastases. He presented with 2/7 of palpitations and feeling generally unwell. Diagnosed with a saddle PE.

(b) Brief Hospital Course:
Mr. John Doe is a 68-year-old male with metastatic NSCLC and brain metastases. He presented with 2/7 of palpitations and feeling generally unwell. Diagnosed with a saddle PE.

(c) Brief Hospital Course:
Mr. John Doe is a 68-year-old male with metastatic NSCLC and brain metastases. He presented with 2/7 of palpitations and feeling generally unwell. Diagnosed with a saddle PE.

Figure 5: Discharge summary snippet with highlights generated from attention heads for (a) the grandparent code (*190-199 Malignant neoplasm of other and unspecified sites*), (b) the parent code (*198 Secondary malignant neoplasm of other specified sites*), and (c) the specific code (*198.3 Secondary malignant neoplasm of brain and spinal cord*). Different words and phrases are attended at each level.

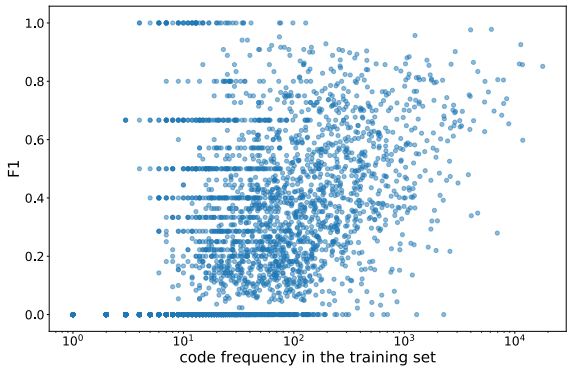

Figure 6: Per-code frequency of training examples v.s. $F1_{macro}$ score from the ontological attention model

and Hill estimated sensitivity and specificity to be 80% and 100% respectively for ICD codes relating to stroke and its risk factors (Kokotailo and Hill, 2005). In the MIMIC-III dataset, we inspected the assignment of smoking codes (current smoker *305.1*, past smoker *V15.82*, or never smoked i.e., no code at all), using regular expression matching to identify examples of possible miscoding, followed by manual inspection of 60 examples (10 relating to each possible miscoding category) to verify our estimates. We estimated that 10% of patients had been wrongly assigned codes, and 30% of patients who had a mention of smoking in their record had not been coded at all. We also observed that often the "correct" code is not clear-cut. For instance, many patients had smoked in the distant past or only smoke occasionally, or had only recently quit; in these cases, where the narrator reliability may be questionable, the decision of how to code is a matter of subjective clinical judgement.

5.2 Revisions to the coding standards

Another limitation of working with the MIMIC-III dataset is that during the deidentification process, information about absolute dates was discarded. This is problematic when we consider that the MIMIC-III dataset contains data that was collected between 2001 and 2012, and the ICD-9 coding standard was reviewed and updated annually between 2006 and 2013 (Centers for Medicare & Medicaid Services) i.e., each year some codes were added, removed or updated in their meaning.

To investigate this issue, we took the 2008 standard and mapped codes created post-2008 back to this year. In total, we identified 380 codes that are present in the dataset but were not defined in the 2008 standard. An example can be seen in Figure 7. We report our results on the 2008 codeset in Table 5. It can be seen that there is an improvement to the metrics on this dataset, which we expect would increase further if all codes were mapped back to the earliest date of 2001. Without time data, it is an unfair task to predict codes which are fundamentally time-dependent. This is an interesting example of conflicting interests between (de)identifiability and task authenticity.

During real-world deployment, codes should be assigned according to current standards. In order to use older data, codes should be mapped forwards rather than backwards. The backwards operation was possible by automated re-mapping of the codes, however the forwards operation is more arduous. Newly introduced codes may require annotation of fresh labels or one-to-many conversion — both operations requiring manual inspection of the original text. A pragmatic approach would be to mask out codes for older documents where they cannot be automatically assigned.

6 Conclusions

We have presented a neural architecture for automated clinical coding which is driven by the ontological graph of relationships between codes. This model establishes a new state-of-the-art result for the task of automated clinical coding with MIMIC-III dataset. Compared to simply doubling the number of attention heads, our ontological attention ensemble mechanism provides improve-

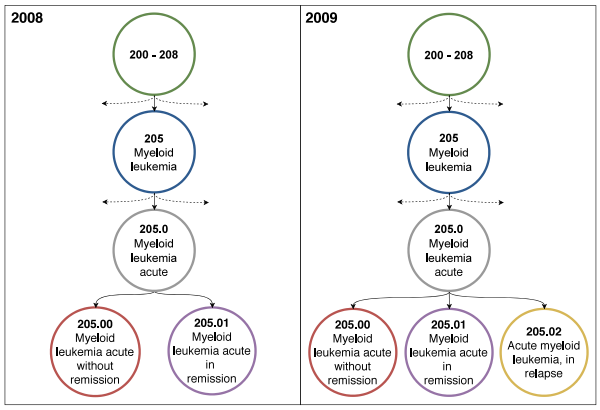

Figure 7: Example code added to the ICD-9 standard.

ments in *accuracy*, in *memory efficiency*, and in *interpretability*. Our method is not specific to an ontology, and in fact could be used for a graph of any formation. If we were to exploit further connections within the ICD ontology e.g., between related diagnoses and procedures, and between child codes which share modifier digits, we would expect to obtain a further performance boost.

We have illustrated that labels may not be reliably present or correct. Thus, even where plenty of training examples are available, the performance may (appear to) be low. In practice, the most successful approach may be to leverage a combination of automated techniques and manual input. An active learning setup would facilitate adoption of new codes by the model as well as allowing endorsement of suggested codes which might otherwise have been missed by manual assignment, and we propose this route for future research.

References

Diane L Adams, Helen Norman, and Valentine J Burroughs. 2002. Addressing medical coding and billing part ii: a strategy for achieving compliance. a risk management approach for reducing coding and billing errors. *Journal of the National Medical Association*, 94(6):430.

Tal Baumel, Jumana Nassour-Kassis, Michael Elhadad, and Noemie Elhadad. 2017. Multi-label classification of patient notes: Case study on ICD code assignment. *ArXiv*.

Centers for Medicare & Medicaid Services. New, deleted, and revised codes - summary tables. https://www.cms.gov/Medicare/Coding/ICD9ProviderDiagnosticCodes/summarytables.html.

Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. 2018. Aˆ 2-nets: Double attention networks. In *Advances in Neural Information Processing Systems*, pages 352–361.

Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F Stewart, and Jimeng Sun. 2017. Gram: graph-based attention model for healthcare representation learning. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 787–795. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Francisco Duarte, Bruno Martins, Càtia Sousa Pinto, and Màrio J. Silva. 2018. Deep neural models for icd-10 coding of death certificates and autopsy reports in free-text. *Journal of Biomedical Informatics*, 80:64–77.

Richárd Farkas and György Szarvas. 2008. Automatic construction of rule-based icd-9-cm coding systems. In *BMC bioinformatics*, volume 9, page S10. BioMed Central.

AEW Johnson, TJ Pollard, L Shen, L Lehman, M Feng, M Ghassemi, B Moody, P Szolovits, LA Celi, and RG Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3.

Sarvnaz Karimi, Xiang Dai, Hamed Hassanzadeh, and Anthony Nguyen. 2017. Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. *In Proc. of the BioNLP 2017 Workshop*, pages 328–332.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Rae A Kokotailo and Michael D Hill. 2005. Coding of stroke and stroke risk factors using international classification of diseases, revisions 9 and 10. *Stroke*, 36(8):1776–1781.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam

Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8).

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3132–3142.

Najmeh Sadoughi, Greg P. Finley, James Fone, Vignesh Murali, Maxim Korenevski, Slava Baryshnikov, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. Medical code prediction with multi-view convolution and description-regularized label-dependent attention. *arXiv*.

Haoran Shi, Pengtao Xie, Zhiting Hu, Ming Zhang, and Eric P. Xing. 2018. Towards automated icd coding using deep learning. *ArXiv*.

Hanna Suominen, Filip Ginter, Sampo Pyysalo, Antti Airola, Tapio Pahikkala, Sanna Salanterä, and Tapio Salakoski. 2008. Machine learning to automate the assignment of diagnosis codes to free-text radiology reports: a method description. In *Proceedings of the ICML/UAI/COLT Workshop on Machine Learning for Health-Care Applications*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.