Layer-wise Guided Training for BERT: Learning Incrementally Refined Document Representations Supplementary Material

Nikolaos Manginas † **Ilias Chalkidis** †‡ **Prodromos Malakasiotis** †‡
† Institute of Informatics & Telecommunications, NCSR "Demokritos"
‡ Department of Informatics, Athens University of Economics and Business
[nmanginas,ichalkidis,pmalakasiotis]@iit.demokritos.gr

A Data manipulation

Hierarchy truncation: In order to directly apply all our methods, we truncate both hierarchies and reduce their depth to six. We believe this truncation is justified since in EUROVOC the last two layers contain a very small number of labels, which are rarely, if at all, assigned and in ICD-9 the first layer also contains a very small number of labels which are very general and can be trivially classified (Table 1). In both cases it seems that only minimal information is lost which would have small practical use in the classification tasks.

Depth	EUROVOC	ICD-9
1	21	4*
2	127	79
3	568	589
4	4,545	3,982
5	2,335	9,640
6	497	7,234
7	79*	867
8	6*	-
Overall	8,178 / 8,093	22,395 / 22,391

Table 1: Label distribution across EUROVOC and ICD-9 hierarchy levels. Concepts (labels) are arranged from more abstract (level 1-2) to more specialized ones (levels 6-8). Labels with an asterisk are truncated in our experiments.

Document Truncation: Documents in both datasets are often above the 512 token limit of BERT. To reduce document size, we perform a number of pre-processing normalizations, including removal of numeric tokens, punctuation and stop-words.[1] In EURLEX documents have been tokenized using SpaCy's default tokenizer,[2] while in MIMIC-III, we use regular expressions tailored for the biomedical domain. While document length is severely reduced post normalization, if a document still has a larger number of tokens, i.e. more than 512, we use the first 512 tokens and ignore the rest.

[1] Similar procedures are very common in classification, thus we believe they do not harm text semantics.

[2] https://spacy.io

B Experimental Setup

All our methods build on BERT-BASE and are implemented in Tensorflow 2. For EURLEX we use the original BERT-BASE (Devlin et al., 2019), while for MIMIC-III we use SCIBERT (Beltagy et al., 2019), which has the same architecture (12 layers, 768 hidden units, 12 attention heads), and better suits biomedical documents.[3] Our models are tuned by grid searching three learning rates (2e-5, 3e-5, 5e-5) and two drop-out rates $(0, 0.1)$. We use the Adam optimizer (Kingma and Ba, 2015) with early stopping on validation loss. In preliminary experiments, we found that weighting individual losses with respect to the number of labels in each level is crucial. We therefore weigh each loss by the percentage of labels at the corresponding level, i.e., $w_n = \frac{|L_n|}{|L|}$, where $|L_n|$ is the number of labels in the n^{th} level of the hierarchy and $|L|$ is the total number of labels across all levels, e.g., in EURLEX57K, $w_1 = \frac{21}{8093} \approx 0.0026$.

C Evaluation in LMTC

The literature of LMTC (Rios and Kavuluru, 2018; Chalkidis et al., 2019) mostly uses information retrieval evaluation measures. We support the premise that when the number of labels is that large the problem mimics retrieval with each document acting as a query and the model having to score relevant labels higher than the rest. However in our study, it would be really confusing to report the standard retrieval metrics $\text{Recall@}R$, $\text{Precision@}K$, $\text{nDCG@}K$ since we evaluate our classifiers at each hierarchy depth and reasonable values for K have large fluctuations between levels, as the number of labels per level vastly varies (see Table 1). Instead, we prefer R-Precision (Manning et al., 2009), which is the $\text{Precision@}R$ where R

[3] We use the Transformers library of Huggingface (https://github.com/huggingface/transformers).

is the number of gold labels associated with each document. It follows that R-Precision can neither under-estimate (penalize) nor over-estimate the performance of the models (Chalkidis et al., 2019).

D Peculiarities of MIMIC-III dataset

In our experiments we observe a hindered performance in MIMIC-III, which can be attributed to a number of characteristics of the dataset. Firstly, documents contain a lot of non-trivial biomedical terminology which naturally makes the classification task more difficult. Further, discharge summaries describe a patient's condition during their hospitalization and therefore proper label annotations change throughout the document as the patient's diagnosis changes or as they exhibit new symptoms, e.g., "*the patient was admitted to the hospital with no heart issues, [. . .] the patient had a heart failure and died.*". Both the in-domain language and the constant change of events make the dataset more challenging than EURLEX57K, where documents are more organized and well-written also with simpler language.

It therefore seems reasonable that in MIMIC-III allowing lower BERT layers to retain and enhance the preliminary functionality, without explicitly guiding them, is of utmost importance. We would like to highlight that even though we use SCIBERT (Beltagy et al., 2019), which is based on a new scientific vocabulary, we observe that specialized biomedical terms are often over-fragmented in multiple sub-word units, e.g. 'atelectasis' splits into ['ate', '##lect', '##asis']. Thus, the initial layers need to decipher these over-fragmented sub-word units and reconstruct the original word semantics. On the contrary, in EURLEX57K, classifying general concepts in the initial layers, even considering only the sub-word unit embeddings is plausible.

E Discussion on model utilization

We present additional results for the rest of the methods (IN-PAIRS, HYBRID). Figure 1 shows the average angular distances between the `[cls]` representations of each layer (Figure 1) for all considered methods. We observe that the distances of IN-PAIRS between consecutive `[cls]` representations follow a similar pattern with those of ONE-BY-ONE, with the exception of 0.25+ distances which are more dense in the upper layers for IN-PAIRS. This is reasonable, since in IN-PAIRS all layers directly contribute to the classification tasks. The pattern of HYBRID is very similar to ONE-BY-ONE and IN-PAIRS, except for the first three non-guided layers in which distances bear close resemblance to those of the corresponding layers in LAST-SIX. Similar observations hold for MIMIC-III (Figure 3). Finally, Figure 2 shows the KL-Divergence of the average (across heads) attention for all layers on the development data. All structured methods show better utilization of the attention mechanism than FLAT, having higher KL-Divergence across layers. Contrary, in MIMIC-III, all structured methods follow a similar pattern of low KL-Divergence across layers (Figure 4), even lower than the upper layers of FLAT, i.e., the models attend to similar sub-word positions across layers. We aim to further study and explain this behaviour in future work.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3606–3611, Hong Kong, China.

Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-Scale Multi-Label Text Classification on EU Legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6314–6322, Florence, Italy.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, abs/1810.04805.

Diederik P. Kingma and Jim Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 5th International Conference on Learning Representations*.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2009. *Introduction to Information Retrieval*. Cambridge University Press.

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. Association for Computational Linguistics.

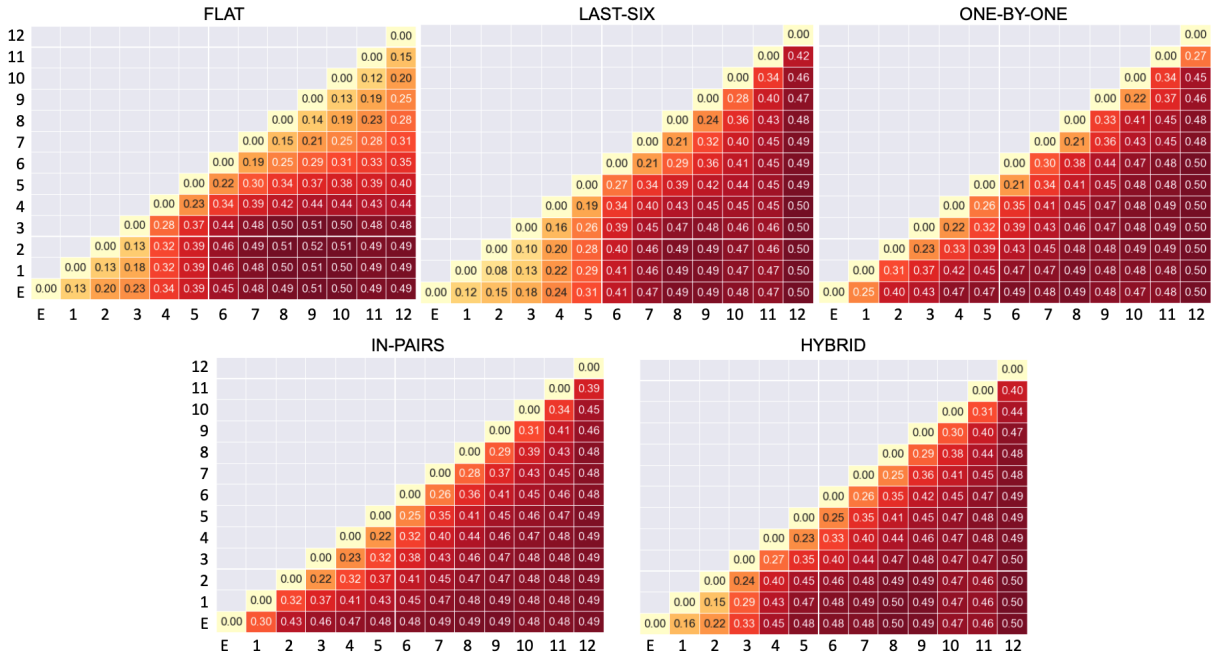

Figure 1: Angular distance between `[cls]` representations across layers in the development dataset of EURLEX57K.

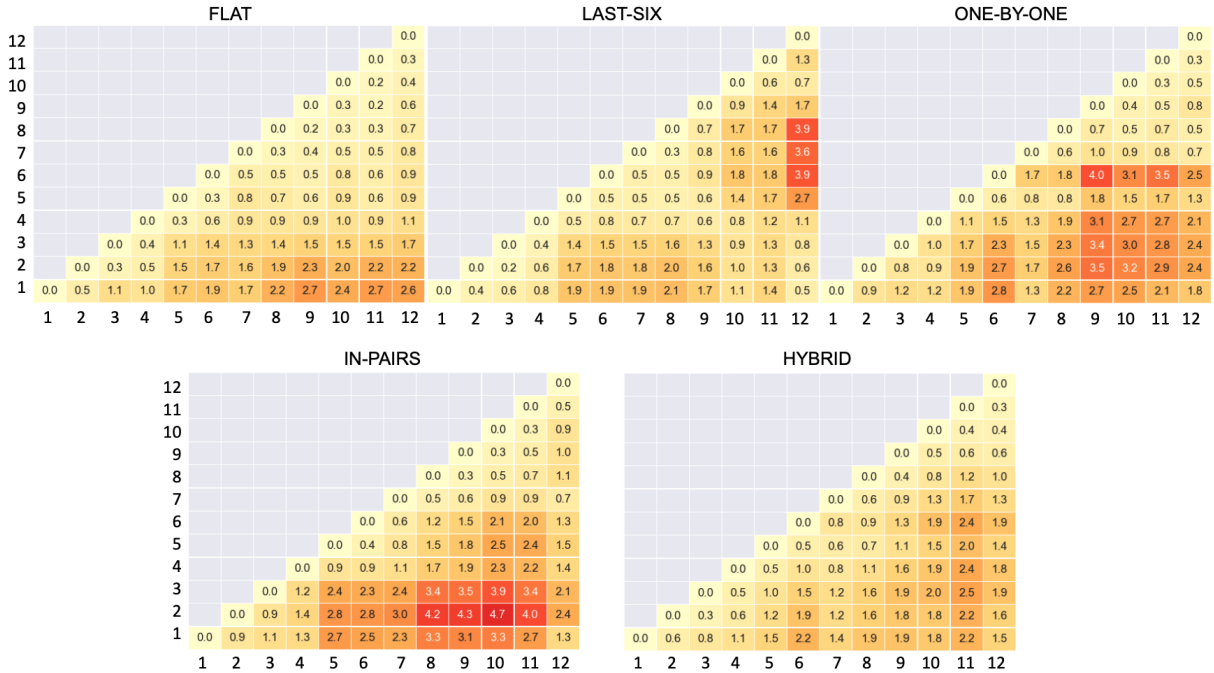

Figure 2: KL-Divergence between attention distributions across layers in the development dataset of EURLEX57K.

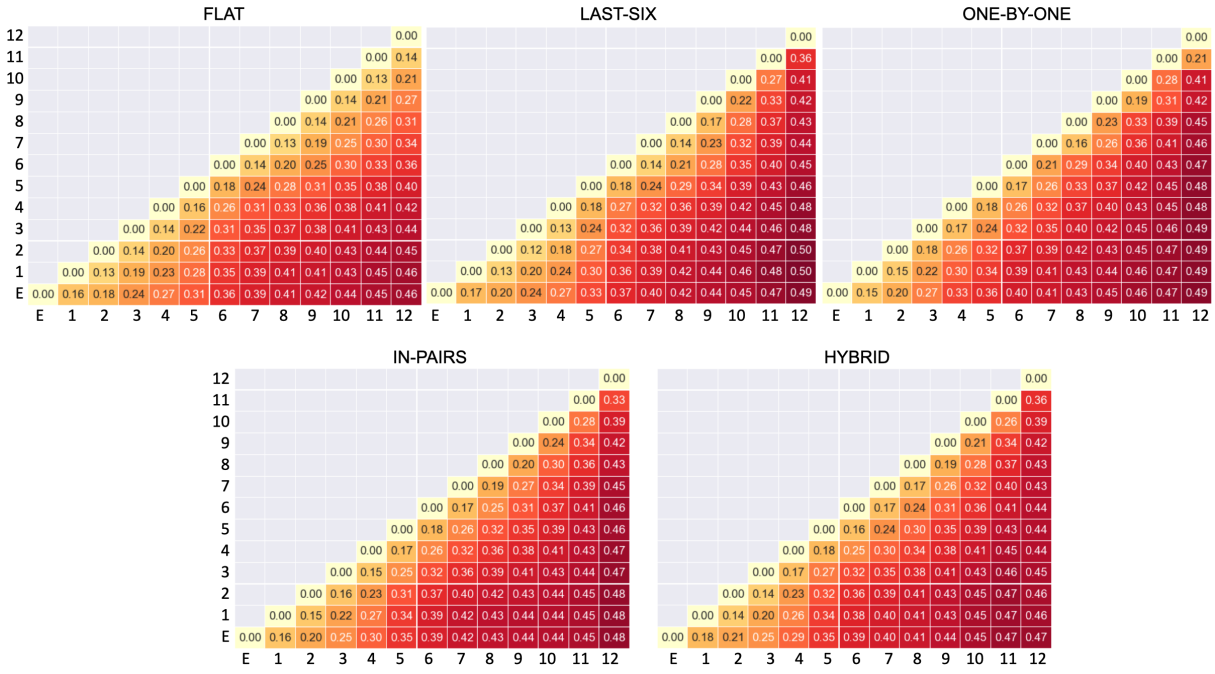

Figure 3: Angular distance between `[cls]` representations across layers in the development dataset of MIMIC-III.

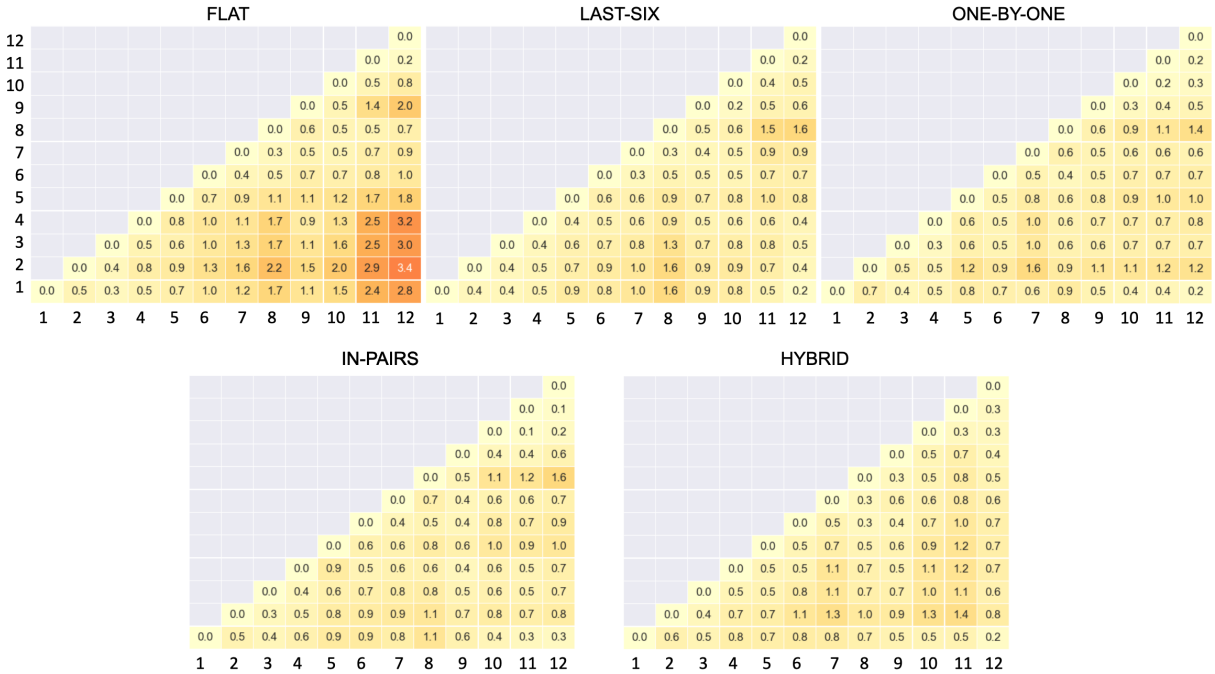

Figure 4: KL-Divergence between attention distributions across layers in the development dataset of MIMIC-III.