Modeling Hierarchical OPS Labels In Multilabel Recurrent Neural Network Based Document Classification

Lennart Grosser

Exposé

July 2019

Contents

1 Introduction

A common task in machine learning research is binary classification where an object described by a set of features is associated with one out of two classes, e.g. classifying a text to be either positvely or negatively phrased [7, 15]. Multi-class classification problems describe scenarios where the object is associated with one out of three or more classes. If an object can be part of multiple classes, one usually refers to multi-label problems. A more challenging scenario is when the labels are organized in a hierarchical structure where objects then refer to one or more given superclasses and their corresponding subclasses. If an object can be tagged with multiple labels which occur in a hierarchical structure, one refers to hierarchical multi-label classification (HMC) [18]. Neural network architectures have been shown to be an effective way for HMC [4, 16, 18, 19]. An algorithm that performs HMC can tackle the problem in two steps: first, it must understand and learn an accurate representation of the text it's trying to classify. Second, it needs to use an appropriate classification approach that considers the hierarchical label structure and avoids unnecessary computation. A typical real-life scenario where HMC is performed is assigning a surgical report to OPS codes. OPS codes describe surgical procedures, naturally organized in a hierarchical structure [1, 8].

2 Thesis Objective

The goal of the master thesis is to explore the nature of surgical reports and their relation to corresponding OPS codes. An existing hierarchical neural network architecture is supposed to be improved in terms of mapping surgical reports to the correct OPS codes. The architecture is based on the Hierarchical Attention Network proposed in 2016 by Yang et al. [17]. The implementation supposed to be used as a starting point deviates from the original HAN [17] mainly in two aspects: the Word2Vec algorithm which is used for embedding the words into a vector space was replaced by the FastText algorithm presented in [11] and the Gated Recurrent Units [5] were replaced by LSTM cells [10]. The replacements led to improved accuracy and precision. Three possible improvement dimensions are supposed to be considered regarding better hierarchical multi-label performance: First, the replacement of the hierarchical attention network architecture by a flat recurrent neural network architecture with attention mechanism shall be considered, similar to the approach proposed in [18]. Second, the hierarchical attention network architecture shall be extended by a loss function which considers the hierarchical structure when being minimized during training. Third, a non-neural network architecture, e.g. a Support Vector Machine with Bag of Words is supposed to be implemented solving the same task. The performance of all three approaches will be tested and evaluated against each other as well as against the previous implementation. The models will be trained on a data set of german surgical reports and tested for their classification accuracy and training / prediction time. Furthermore, the precision and ROC curve will be considered. The goal is to increase the amount of correctly classified OPS codes compared to the baseline model. Such an evaluation would be fairly useful in the field of clinical text analysis since a surgical report could be associated to the correct set of OPS codes with a higher probability.

3 Related Work

3.1 Text Embedding

Machine Learning models commonly expect the input data to be of a numerical type which directly yields the challenge of finding an accurate vectoral numerical representation of documents. In 2013 Mikolov et al. proposed two model architectures capable of learning vectoral representations of words while keeping word similarity by training a feed forward neural network on the relation of words and their context [13]. After training, the one-hot-encoding of a word may then be used as an index to extract its vectoral representation from the hidden dimension of the network. By considering the cosine distance one is able to calculate the similarity of two vectors respectively words. Word embeddings using this approach outperform previous approaches like N-grams and Latent Semantic Analysis Similarity [13, 20]. A rather broad overview of vectoral representations of text in deep learning is presented in [9]. An extension of the approach above was presented in 2016 by Joulin et al. in [11] where the proposed algorithm breaks words into their n-grams and uses the n-grams as inputs. In 2017 Bojanowski et al. present an approach to learn word embeddings through an extended skip-gram model [3] where a word is broken into a set of n-grams and the word representation is learned as the sum of its n-grams. This enables consideration of character-level similarities of words.

3.2 Text Representation

Once an similarity-preserving embedding of a document is found, a machine learning model still needs to learn its meaning. A common challenge in natural language models is to capture the relation of words which collectively form a meaning but don't necessarily appear next to each other. Furthermore, the last sentence of a document may refer to the first sentence in the same document. These long-term relationships between words or sentences have be captured by the model. A popular neural network architecture tackling this problem is the Hierarchical Attention Network (HAN) [17] proposed in 2016 by Yang et al.. The HAN uses bidirectional Gated Recurrent Units [5] in a two-level architecture. The first level processes the words of a sentence and encodes their relationship. The second level processes the resulting encodings for all sentences in a document to learn a document vector which may serve as the input for a further classification task. An alternative approach to the Gated Recurring Unit for capturing long-term relationships in sequences is presented in [10]. In 2017 the HAN has been extended to multilingual use by sharing several parameters across different training processes (one for each language) in a single network and constructing a joint multilingual objective [14]. A different version of the HAN using convolutional layers was presented in 2018 by Gao et al. [6]. A combination of convolutional layers and Gated Recurrent Units is presented in [2] by Abreu et al..

3.3 Hierarchical Multi-Label Classification

Many approaches to multi-label-classification exist, some of them referrring to extreme multi-label text classification (XMTC) where the label hierarchy is flattened and treated as a larger set of labels. In 2018 You et al. proposed a neural network architecture for multi-label text classification which uses word embeddings as input for a self-attention mechanism inspired by Lin et al. [12] to capture differently important parts of a text. For each label there is a separate attention layer followed by a fully connected layer which produces the output, that is the probability for the label. In [16] a neural network architecture of hierarchical multi-label classification is proposed by Wehrmann et al. which performs optimization on both local and global level. Local approaches train a classifier for every split in the label hierarchy and are much more computationally expensive. Global approaches train a single classifier which associates an object with its corresponding classes, these approaches tend to miss local information.

4 Methodological Foundation

4.1 Methodology

The thesis means to give an overview of current approaches related to natural language processing, especially text representation learning and hierarchical multi-label classification. An existing network architecture is meant to be modified in three different ways. Two neural network architectures and a non-neural network approach will be implemented and explained. The networks will be implemented using Python and Keras (Tensorflow Backend) as they are common tools for such tasks. Once implemented, the models will be tested, evaluated and compared.

4.2 Data

The data that will be used is a subset of 430.000 surgical reports from german hospitals. Each surgical report is a german text describing the disease of the patient as well as the corresponding surgical procedure applied. The description covers the used surgical techniques and information about the condition of the patient's body (part) before and during the surgery. The reports usually contain latin medical terms. For all surgical reports the correct OPS codes assigned by an accredited expert are known.

1. Amount of unique OPS codes: 134165
2. Minimum number of OPS codes per document: 1
3. Maximum number of OPS codes per document: 33
4. Average amount of OPS codes per document: 2.2

References

[1] Icd-10-gm version 2019, systematisches verzeichnis, internationale statistische klassifikation der krankheiten und verwandter gesundheitsprobleme, 10. revision, stand: 21.september 2018. *Deutsches Institut für Medizinische Dokumentation und Information (DIMDI) im Auftrag des Bundesministeriums für Gesundheit (BMG) unter Beteiligung der Arbeitsgruppe ICD des Kuratoriums für Fragen der Klassifikation im Gesundheitswesen (KKG)*, 2019. URL `https://www.dimdi.de/dynamic/de/startseite/`.

[2] Jader Abreu, Luis Fred, David Macêdo, and Cleber Zanchettin. Hierarchical attentional hybrid neural networks for document classification. *CoRR*, abs/1901.06610, 2019. URL `http://arxiv.org/abs/1901.06610`.

[3] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017. doi: 10.1162/tacl_a_00051. URL `https://doi.org/10.1162/tacl_a_00051`.

[4] Ricardo Cerri, Rodrigo C. Barros, and André C.P.L.F. de Carvalho. Hierarchical multi-label classification using local neural networks. *Journal of Computer and System Sciences*, 80(1):39 – 56, 2014. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2013.03.007. URL `http://www.sciencedirect.com/science/article/pii/S0022000013000718`.

[5] KyungHyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. *CoRR*, abs/1409.1259, 2014. URL `http://arxiv.org/abs/1409.1259`.

[6] Shang Gao, Arvind Ramanathan, and Georgia Tourassi. Hierarchical convolutional attention networks for text classification. In *Proceedings of The Third Workshop on Representation Learning for NLP*, pages 11–23, Melbourne, Australia, July 2018. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/W18-3002`.

[7] L.A. Gottschalk and G.C. Gleser. *The measurement of psychological states through the content analysis of verbal behavior*. University of California Press, 1969. URL `https://books.google.de/books?id=5JWBAAAAIAAJ`.

[8] Bernd Graubner. Icd und ops. *Bundesgesundheitsblatt - Gesundheitsforschung - Gesundheitsschutz*, 50(7):932–943, Jul 2007. ISSN 1437-1588. doi: 10.1007/s00103-007-0283-x. URL `https://doi.org/10.1007/s00103-007-0283-x`.

[9] Karol Grzegorczyk. Vector representations of text data in deep learning. *CoRR*, abs/1901.01695, 2019. URL `http://arxiv.org/abs/1901.01695`.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735. URL `https://doi.org/10.1162/neco.1997.9.8.1735`.

[11] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *CoRR*, abs/1607.01759, 2016. URL `http://arxiv.org/abs/1607.01759`.

[12] Zhouhan Lin, Minwei Feng, Cícero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. A structured self-attentive sentence embedding. *CoRR*, abs/1703.03130, 2017. URL `http://arxiv.org/abs/1703.03130`.

[13] Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. Efficient estimation of word representations in vector space. pages 1–12, 01 2013.

[14] Nikolaos Pappas and Andrei Popescu-Belis. Multilingual hierarchical attention networks for document classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1015–1025, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing. URL `https://www.aclweb.org/anthology/I17-1102`.

[15] Philip J. Stone and Earl B. Hunt. A computer approach to content analysis: Studies using the general inquirer system. In *Proceedings of the May 21-23, 1963, Spring Joint Computer Conference*, AFIPS '63 (Spring), pages 241–256, New York, NY, USA, 1963. ACM. doi: 10.1145/1461551.1461583. URL `http://doi.acm.org/10.1145/1461551.1461583`.

[16] Jonatas Wehrmann, Ricardo Cerri, and Rodrigo Barros. Hierarchical multi-label classification networks. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5075–5084, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/wehrmann18a.html`.

[17] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. pages 1480–1489, 01 2016. doi: 10.18653/v1/N16-1174.

[18] Ronghui You, Suyang Dai, Zihan Zhang, Hiroshi Mamitsuka, and Shanfeng Zhu. Attentionxml: Extreme multi-label text classification with multi-label attention based recurrent neural networks. 11 2018.

[19] L. Zhang, S.K. Shah, and I.A. Kakadiaris. Hierarchical multi-label classification using fully associative ensemble learning. *Pattern Recognition*, 70: 89 – 103, 2017. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2017.05.007. URL `http://www.sciencedirect.com/science/article/pii/S0031320317301899`.

[20] Geoffrey Zweig and Chris J.C. Burges. The microsoft research sentence completion challenge. Technical Report MSR-TR-2011-129, December 2011. URL `https://www.microsoft.com/en-us/research/publication/the-microsoft-research-sentence-completion-challenge/`.