From back to the roots into the gated woods: Deep learning for NLP

Barbara Plank
IT University of Copenhagen
`bplank@gmail.com`

Abstract

Deep neural networks have revolutionized many fields, including Natural Language Processing. This paper outlines teaching materials for an introductory lecture on deep learning in Natural Language Processing (NLP). The main submitted material covers a summer school lecture on *encoder-decoder models*. Complementary to this is a set of jupyter notebook slides from earlier teaching, on which parts of the lecture were based on. The main goal of this teaching material is to provide an overview of neural network approaches to natural language processing, while linking modern concepts back to the roots showing traditional essential counterparts. The lecture departs from count-based statistical methods and spans up to gated recurrent networks and attention, which is ubiquitous in today's NLP.

1 Introduction

In 2015, the "deep learning tsunami" hit our field (Manning, 2015). In 2011, neural networks were not really "a thing" yet in NLP; they were even hardly taught. I remember when in 2011 Andrew Ng introduced the free online course on machine learning, which soon after led Daphne Koller and him to start massive online education: "Many scientists think the best way to make progress [..] is through learning algorithms called *neural networks*".[1] Ten years later, it is hard to imagine any NLP education not touching upon neural networks and in particular, representation learning. While neural networks have undoubtedly pushed the field, it has led to a more homogenized field; some lecturers even question whether to include pre-neural methods. I believe it is essential to teach statistical foundations besides modern DL, ie., to go back to the roots, to better be equipped for the future.[2]

2 Structure of the summer school lecture

This paper outlines teaching material (Keynote slides, Jupyter notebooks), which I would like to provide to the community for reuse when teaching an introductory course on deep learning for NLP.[3]

The main material outlined in this paper is a Keynote presentation slide deck for a 3-h lecture on encoder-decoder models. An overview of the lecture (in non-temporal form), from foundations, to representations to neural networks (NNs) beyond vanilla NNs is given in Figure 1. Moreover, I provide complementary teaching material in form of Jupyter notebook (convertable to slides). I've used these notebooks (slides and exercises) during earlier teaching and in parts formed the basis of the summer school lecture (see Section 3).
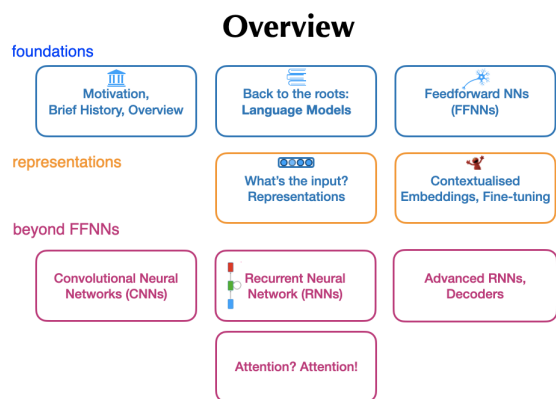

Figure 1: Overview of the core concepts covered.

The lecture covers NLP methods broadly from introduction concepts of more traditional NLP methods to recent deep learning methods. A particular focus is to contrast traditional approaches from the statistical NLP literature (sparse repre-

---

[1] Accessed March 15, 2021: `https://bit.ly/2OZH2MZ`

[2] Disclaimer: This submitted teaching material is limited as it spans a single lecture. However, I believe it to be essential to

teach traditional methods besides modern neural approaches (e.g. count-based vs prediction-based word representations; statistical n-grams vs neural LMs; naive Bayes and logistic regression vs neural classifiers, to name a few examples).

[3] Available at: `https://github.com/bplank/teaching-dl4nlp`

sentations, n-grams) to their deep learning-based counterparts (dense representations, 'soft' ngrams in CNNS and RNN-based encoders). Consequently, the following topics are covered in the lecture:

- Introduction to NLP and DL (deep learning), what makes language so difficult; traditional versus neural approaches

- N-gram Language Models,

- Feedforward Neural Networks (FFNNs)

- What's the input? Sparse traditional vs dense representations; Bag of words (BOW) vs continuous BOW (CBOW)

- Neural Language Models (LMs)

- Convolutional Neural Networks (CNNs) for Text Classification ('soft n-grams')

- Recurrent Neural Networks (RNNs) for Text Processing, RNNs as LMs

- Bidirectional RNNs, stacking, character representations

- Gated RNNs (GRU, LSTMs)

- Deep contextualized embeddings (Embeddings as LMs, Elmo)

- Attention

The lecture was held as 3-h lecture at the first Athens in NLP (AthNLP) summer school in 2019. In the overall schedule of the summer school, this lecture was the 3rd in a row of six. It was scheduled after an introduction to classification (by Ryan McDonald), and a lecture on structured prediction (by Xavier Carreras). The outlined encoder-decoder lecture was then followed by a lecture on machine translation (by Arianna Bisazza; the lecture build upon this lecture here and included the transformer), machine reading (by Sebastian Riedel) and dialogue (by Vivian Chen).[4] Each lecture was enhanced by unified lab exercises.[5]

---

[4]Videos of all lectures are available at: http://athnlp2019.iit.demokritos.gr/schedule/index.html

[5]Exercises were kindly provided as unified framework by the AthNLP team, and hence are not provided here. They included, lab 1: Part-of-Speech tagging with the perceptron; lab 2: POS tagging with the structured perceptron; lab 3: neural encoding for text classification; lab 4: neural language modeling; lab 5: machine translation; lab 6: question answering

As key textbook references, I would like to refer the reader to chapters 3-9 of Jurafsky and Martin's 3rd edition (under development) (Jurafsky and Martin, 2020),[6] and Yoav Goldberg's NLP primer (Goldberg, 2015). Besides these textbooks, key papers include (Kim, 2014) for CNNs on texts, attention (Luong et al., 2015) and Elmo (Peters et al., 2018).

3 Complementary notebooks of earlier material

This lecture evolved from a series of lectures given earlier, amongst which a short course given in Malta in 2019, and a MSc-level course I taught at the University of Groningen (Language Technology Project). To complement the Keynote slides of the summer school lecture provided here, earlier Jupyter notebooks can be found at the website. These cover a subset of the material above.

4 Conclusions

This short paper outlines teaching material for an introductory lecture on deep learning for NLP. By releasing this teaching material, I hope to contribute material that fellow researchers find useful when teaching introductory courses on DL for NLP. For comments and suggestions to improve upon this material, please reach out to me.

---

[6]https://web.stanford.edu/~jurafsky/slp3/

References

Yoav Goldberg. 2015. A primer on neural network models for natural language processing.

Dan Jurafsky and James H. Martin. 2020. *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Christopher D Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.