Processamento de Linguagem Natural por meio de redes neurais profundas: Teoria e Aplicações

Fabiano Luz Felipe Salvatore Marcelo Finger

April 15, 2018

1 Resumo

O tutorial vai ser dividido em duas partes. Vamos começar revisando a base teórica sobre redes neurais profundas com o foco em processamento de linguagem natural. Na segunda parte, vamos desenvolver o arcabouço teórico com novas arquiteturas e falar sobre suas aplicações.

1.1 Processamento de linguagem com redes profundas (1h30)

Diferentes arquiteturas de aprendizado profundo (*deep learning*) são usadas em processamento de linguagem natural: redes convolucionais possuem uma boa performance em tarefas onde queremos encontrar certos indicadores linguísticos independente da sua posição (e.g. classificação de documentos, análise de sentimento, etc.); vetores de palavras (*word embeddings*) são apreendidos por uma arquitetura que se assemelha a uma rede neural (*feedforward neural network*). Mas para uma variedade de atividades em processamento de linguagem queremos capturar regularidades e similaridades na estrutura do texto. Dessa forma os modelos recorrentes e recursivos tem sido usados amplamente no campo. Nessa primeira parte do tutorial vamos nos focar em modelos generativos baseados em redes recorrentes (*recurrent neural network - RNNs*), dado a eficacia de tais modelos.

Vamos começar apresentando a arquitetura básica de uma RNN juntamente com sua aplicação em linguística computacional: a construção de modelos de linguagem. Também vamos explicar o processo de retropropagação para essa rede (*backpropagation through time*) e mostrar o problema da explosão e desaparecimento do gradiente. Por fim vamos explicar as arquiteturas que solucionam esse problema GRU (*gated recurrent unit*) e LSTM (*Long short-term memory*).

Também vamos mencionar como estas arquiteturas podem ser implementadas em plataformas como Pytorch.

1.2 Arquitetura de codificação-decodificação com mecânismo de Atenção Neural (1h30)

Nesta parte do tutorial vamos abordar arquiteturas neurais de transformação de texto, com foco nas arquiteturas neurais de codificação-decodificação. Este tipo de arquitetura de codificação-decodificação foi proposto para a tarefa de tradução estatística, e atualmente pode ser aplicada a diversas tarefas de processamento de linguagem natural como por exemplo: sumarização, extração de entidades nomeadas, análise semântica, etc.

Este modelo trata a transformação de uma sequência de entrada em uma sequência de saída como um processo que codifica uma seqüência de comprimento variável em uma representação vetorial de comprimento fixo, e outro processo que decodifica uma dada representação vetorial de comprimento fixo em uma outra seqüência de comprimento variável. Cada parte deste processo utiliza uma arquitetura neural recorrente, vista na parte anterior deste tutorial. Este modelo não depende de regras nem de modelos construídos manualmente, nem de léxicos de alta qualidade ou outras complexas estruturas artesanais. Para aplicá-lo, só precisamos de dados rotulados, ou córpus paralelos, elementos de processo simples em um aprendizado supervisionado. Nesta parte do tutorial, vamos mostrar e explicar um modelo capaz de fazer tradução do português para o inglês e também tradução do português para a linguagem de consulta a ontologias, o SPARQL, vamos falar sobre as diferenças entre traduções de linguagens naturais e artificiais.

Por fim, falaremos de modelos de atenção neural, um recurso aplicado para refinar o modelo de codificação-decodificação, por exemplo, para resolver problemas de alinhamento na tradução entre linguagens. Vamos mostrar como performa esta arquitetura aplicada nas tarefas de tradução de língua portuguesa para inglesa. Abordaremos então alguns avanços que conseguimos em nossas pesquisas, como a melhora da representação do léxico de linguagens artificiais e de nossa arquitetura especial que codifica linguagem natural e decodifica linguagens livre de contexto. Também vamos apresentar exemplos de aplicação desse modelo para a tarefa de pergunta e resposta (*Question-answering*) e memória de diálogo.

2 Estrutura

- Introdução a *Deep Learning* ([5, Capítulo 6])
- Representação vetorial de palavras ([8, 4])
- Redes convolucionais (CNN) para análise de sentimento ([6, 4])
- Redes recorrentes (RNN) e suas extensões: GRU e LSTM ([5, Capítulo 10])
- Modelos de linguagem ([4])
- *Coffe break*

- Mapeando sequências em sequências: os modelos Seq2seq ([9])
- Mecanismos de atenção ([3, 7])
- Exemplo prático i): tradução automática ([1])
- Exemplo prático ii): geração de diálogo ([2])

A bibliografia indicada é apenas para quem quiser se preparar antes do tutorial, mas não é necessária a leitura para entender o conteudo que será exposto.

3 Sobre os autores

Fabiano Luz é aluno de doutorado em Ciências da Computação na Universidade de São Paulo (USP) onde trabalha com aprendizagem computacional aplicado a processamento de linguagem humana.

Felipe Salvatore é aluno de doutorado em Ciências da Computação. Depois de ter trabalhado com lógica matemática no mestrado ele se voltou para Processamento de linguagem natural.

Marcelo Finger é professor titular de ciência da Computação no IME-USP, com experiência em diversas áreas de Inteligência Artificial e Processamento de Linguagem Natural.

References

[1] N3LP. `https://github.com/hassyGo/N3LP`.

[2] Parlai. `https://github.com/facebookresearch/ParlAI`.

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *ICLR*, 2015.

[4] Y. Goldberg. A primer on neural network models for natural language processing. *CoRR*, abs/1510.00726, 2015.

[5] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2017.

[6] Y. Kim. Convolutional neural networks for sentence classification. *CoRR*, abs/1408.5882, 2014.

[7] T. Luong, H. Pham, and C. D. Manning. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, 2015.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[9] I. Sustskever, O. Vinyals, and Q. V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 3104–3112, 2014.