

CSCE 771: Computer Processing of Natural Language

Lecture 16: Invited Talk - Transformer LLM, Using LLM in NLP Tasks

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

11TH OCTOBER, 2022

***Carolinian Creed: “I will practice personal and
academic integrity.”***

Organization of Lecture 16

- Opening Segment
 - Recap of recent classes
- Main Lecture
- Concluding Segment
 - About Next Lecture – Lecture 17



Main Section

- Invited Talk: Amitava Das
 - Basics of transformers
 - Using Transformers
- Paper Reading

Recent Classes

Sep 27 (Tu)	Guest Lecture – Dr. Amitava Das: Glove, Word2Vec
Sep 29 (Th)	Representation: Embeddings, Language Models, Reasoning, QUIZ,
Oct 4 (Tu)	Review: Reasoning and Representation for NLP: Ontology, Knowledge Graph, PROJ REVIEW
Oct 6 (Th)	PROJ REVIEW
Oct 11 (Tu)	Guest Lecture – Dr. Amitava Das: Using lang models to solve NLP tasks
Oct 13 (Th)	
Oct 18 (Tu)	Entity extraction, Entity linking, Events extraction, spatio-temporal analysis
Oct 20 (Th)	Topic Analysis, QUIZ

Review of Lecture 15

- Project Overviews
- Ontologies – please go through material

Main Lecture

Invited Talk

- Dr. Amitava Das, Assoc. Research Prof., AI Institute, UoSC
- Material: <https://prezi.com/view/amx5hBo8UhMOn1rPyJ02/>

Review: Transformer

- RNN/ LSTM with
 - Attention
 - attention layer can access all previous states and weighs them according to some learned measure of relevancy to the current token, providing sharper information about far-away relevant tokens
 - **Query** vector, **Key** vector, and **Value** vectors introduced during encoding and decoding phase
 - Parallelization of learning
 - See Dr. Amitava Das's slide for Attention/ BERT video

Source and details: [https://en.wikipedia.org/wiki/Transformer_\(machine_learning_model\)](https://en.wikipedia.org/wiki/Transformer_(machine_learning_model)),
<http://jalammargithub.io/illustrated-transformer/>

BERT - Bidirectional Encoder Representations from Transformers

Learns with two tasks

- Predicting missing words in sentences
 - mask out 15% of the words in the input, predict the masked words.
- Given two sentences A and B, is B the actual next sentence that comes after A, or just a random sentence from the corpus?

(12-layer to 24-layer Transformer)
on (Wikipedia + [BookCorpus](#))

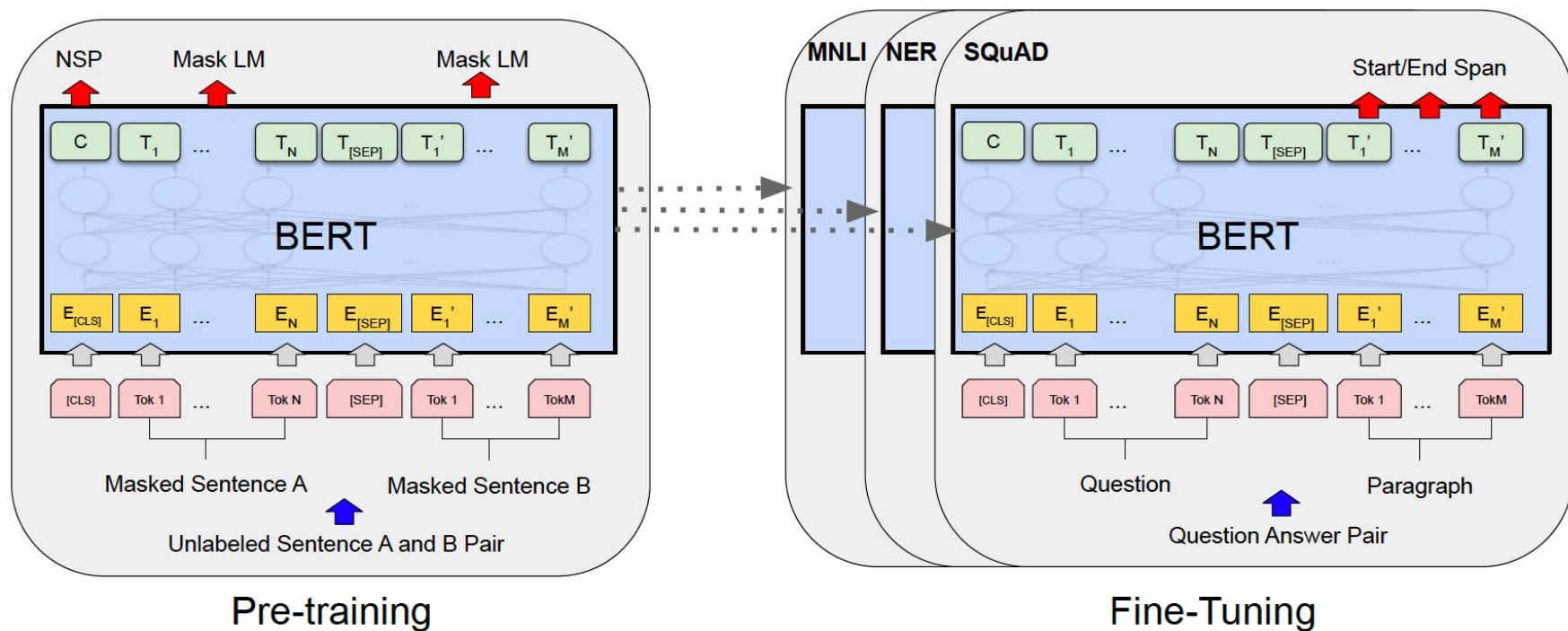
Input: the man went to the [MASK1] . he bought a [MASK2] of milk.
Labels: [MASK1] = store; [MASK2] = gallon

Sentence A: the man went to the store .
Sentence B: he bought a gallon of milk .
Label: IsNextSentence

Sentence A: the man went to the store .
Sentence B: penguins are flightless .
Label: NotNextSentence

Credit and details: <https://github.com/google-research/bert>

BERT: Before and During Usage



Credit and details: **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**
[Jacob Devlin](#), [Ming-Wei Chang](#), [Kenton Lee](#), [Kristina Toutanova](#), 2018

Using BERT in Practice – Huggingface Libraries

- Transformers – <https://github.com/huggingface/transformers>
- APIs to download and use pre-trained models, fine-tune them on own datasets and tasks
 - Code Sample

```
# Loading BERT
model_class, tokenizer_class, pretrained_weights = (ppb.DistilBertModel, ppb.DistilBertTokenizer, 'distilbert-base-uncased')

# Load pretrained model/tokenizer
tokenizer = tokenizer_class.from_pretrained(pretrained_weights)
model = model_class.from_pretrained(pretrained_weights)
```
- Provides pretrained models in 100+ languages.
- Use with popular deep learning libraries, [PyTorch](#) and [TensorFlow](#),
 - Possible to train / fine-tune models with one, and load it for inference with another

Using BERT in Practice – Huggingface Libraries

- DistilBERT
 - Details: <https://medium.com/huggingface/distilbert-8cf3380435b5>
 - Teacher-student learning, also called model distillation
 - Teacher: bert-base-uncased
 - Student: distilBERT - BERT without *the token-type embeddings and the pooler* , and half the layers
 - “**DistilBERT**, has **about half** the total number of parameters of BERT base and retains 95% of BERT’s performances on the language understanding benchmark GLUE”
- Sample code of usage for sentiment classification:
<https://github.com/biplav-s/course-nl/blob/master/l12-langmodel/UsingLanguageModel.ipynb>

Example Pre-Trained Models

1. ALBERT (from Google Research and the Toyota Technological Institute at Chicago) released with the paper ALBERT: A Lite BERT for Self-supervised Learning of Language Representations, by Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, Radu Soricut.
2. BART (from Facebook) released with the paper BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension by Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov and Luke Zettlemoyer.
3. BERT (from Google) released with the paper BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.
4. BERT For Sequence Generation (from Google) released with the paper Leveraging Pre-trained Checkpoints for Sequence Generation Tasks by Sascha Rothe, Shashi Narayan, Aliaksei Severyn.
5. CamemBERT (from Inria/Facebook/Sorbonne) released with the paper CamemBERT: a Tasty French Language Model by Louis Martin*, Benjamin Muller*, Pedro Javier Ortiz Suárez*, Yoann Dupont, Laurent Romary, Éric Villemonte de la Clergerie, Djamé Seddah and Benoît Sagot.
6. CTRL (from Salesforce) released with the paper CTRL: A Conditional Transformer Language Model for Controllable Generation by Nitish Shirish Keskar*, Bryan McCann*, Lav R. Varshney, Caiming Xiong and Richard Socher.
7. DeBERTa (from Microsoft Research) released with the paper DeBERTa: Decoding-enhanced BERT with Disentangled Attention by Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen.
8. DialoGPT (from Microsoft Research) released with the paper DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation by Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, Bill Dolan.
9. DistilBERT (from HuggingFace), released together with the paper DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter by Victor Sanh, Lysandre Debut and Thomas Wolf. The same method has been applied to compress GPT2 into DistilGPT2, RoBERTa into DistilRoBERTa, Multilingual BERT into DistilMBERT and a German version of DistilBERT.
10. DPR (from Facebook) released with the paper Dense Passage Retrieval for Open-Domain Question Answering by Vladimir Karpukhin, Barlas Öğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih.
11. ELECTRA (from Google Research/Stanford University) released with the paper ELECTRA: Pre-training text encoders as discriminators rather than generators by Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning.
12. FlauBERT (from CNRS) released with the paper FlauBERT: Unsupervised Language Model Pre-training for French by Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, Didier Schwab.
13. Funnel Transformer (from CMU/Google Brain) released with the paper Funnel-Transformer: Filtering out Sequential Redundancy for Efficient Language Processing by Zihang Dai, Guokun Lai, Yiming Yang, Quoc V. Le.
14. GPT (from OpenAI) released with the paper Improving Language Understanding by Generative Pre-Training by Alec Radford, Karthik Narasimhan, Tim Salimans and Ilya Sutskever.
15. GPT-2 (from OpenAI) released with the paper Language Models are Unsupervised Multitask Learners by Alec Radford*, Jeffrey Wu*, Rewon Child, David Luan, Dario Amodei** and Ilya Sutskever**.
16. LayoutLM (from Microsoft Research Asia) released with the paper LayoutLM: Pre-training of Text and Layout for Document Image Understanding by Yiheng Xu, Minghao Li, Lei Cui, Shaohan Huang, Furu Wei, Ming Zhou.
17. Longformer (from AllenAI) released with the paper Longformer: The Long-Document Transformer by Iz Beltagy, Matthew E. Peters, Arman Cohan.
18. LXMERT (from UNC Chapel Hill) released with the paper LXMERT: Learning Cross-Modality Encoder Representations from Transformers for Open-Domain Question Answering by Hao Tan and Mohit Bansal.
19. MarianMT Machine translation models trained using OPUS data by Jörg Tiedemann. The Marian Framework is being developed by the Microsoft Translator Team.
20. MBart (from Facebook) released with the paper Multilingual Denoising Pre-training for Neural Machine Translation by Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, Luke Zettlemoyer.
21. MMBT (from Facebook), released together with the paper a Supervised Multimodal Bitransformers for Classifying Images and Text by Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Davide Testuggine.
22. Pegasus (from Google) released with the paper PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization> by Jingqing Zhang, Yao Zhao, Mohammad Saleh and Peter J. Liu.
23. Reformer (from Google Research) released with the paper Reformer: The Efficient Transformer by Nikita Kitaev, Łukasz Kaiser, Anselm Levskaya.
24. RoBERTa (from Facebook), released together with the paper a Robustly Optimized BERT Pretraining Approach by Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. ultilingual BERT into DistilMBERT and a German version of DistilBERT.
25. SqueezeBert released with the paper SqueezeBERT: What can computer vision teach NLP about efficient neural networks? by Forrest N. Iandola, Albert E. Shaw, Ravi Krishna, and Kurt W. Keutzer.
26. T5 (from Google AI) released with the paper Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer by Colin Raffel and Noam Shazeer and Adam Roberts and Katherine Lee and Sharan Narang and Michael Matena and Yanqi Zhou and Wei Li and Peter J. Liu.
27. Transformer-XL (from Google/CMU) released with the paper Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context by Zihang Dai*, Zhilin Yang*, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov.
28. XLM (from Facebook) released together with the paper Cross-lingual Language Model Pretraining by Guillaume Lample and Alexis Conneau.
29. XLM-RoBERTa (from Facebook AI), released together with the paper Unsupervised Cross-lingual Representation Learning at Scale by Alexis Conneau*, Kartikay Khandelwal*, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov.
30. XLNet (from Google/CMU) released with the paper XLNet: Generalized Autoregressive Pretraining for Language Understanding by Zhilin Yang*, Zihang Dai*, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, Quoc V. Le.

Contextual Representations

	Name	Description	URL, References
1.	TF-IDF		Ramos, J. (1999). Using TF-IDF to Determine Word Relevance in Document Queries
2	Word2Vec	Word-based, prediction focus	Mikolov, Tomas; et al. (2013). "Efficient Estimation of Word Representations in Vector Space". arXiv:1301.3781 [cs.CL]. Mikolov, Tomas (2013). "Distributed representations of words and phrases and their compositionality". <i>Advances in Neural Information Processing Systems</i> . arXiv:1310.4546 .
3	Glove	Word-based, count	https://nlp.stanford.edu/projects/glove/ , Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation . [pdf] [bib]
4	BERT / DistilBERT		BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding J. Devlin , M. Chang , K. Lee , and K. Toutanova . (2018)cite arxiv:1810.04805
5	Elmo (embeddings from language models)	Contextual, deep, character-based	https://allennlp.org/elmo ; Deep contextualized word representations, Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, Luke Zettlemoyer. NAACL 2018.
6	Fasttext	Variation of word2vec, works with N-gram, words not in vocabulary	https://fasttext.cc/ , P. Bojanowski*, E. Grave*, A. Joulin, T. Mikolov, Enriching Word Vectors with Subword Information

Commentaries:

<https://jalammar.github.io/illustrated-bert/> , <https://cai.tools.sap/blog/glove-and-fasttext-two-popular-word-vector-models-in-nlp/>

To Do:

- Read papers on language models. (Full pdf available from reading list page on GitHub: <https://github.com/biplav-s/course-nl-f22/tree/main/reading-list>)
 - Hang Li, [Language Models: Past, Present, and Future](#), Communications of the ACM, July 2022, Vol. 65 No. 7, Pages 56-63 10.1145/3490443
 - [A Primer in BERTology: What We Know About How BERT Works](#) (Rogers et al., TACL 2020)
- Will be asked in Quiz3 (Oct 20)

Lecture 16: Concluding Comments

- Introduced transformers and their motivation
- Compliment with reading

Concluding Segment

About Next Lecture – Lecture 17

Lecture 17 Outline

- Entity extraction
- Entity linking
- Representation of extracted entities and links