ML Theory Journal

Aritz Bercher

November 6, 2019

Abstract

In this journal, I will try to gather good references (mainly webpages) about some Machine Learning (and in particular Deep Learning and Natural Language Processing) algorithms as well as some personal thoughts on the subject. The aim of this journal is to keep gather in a same place the *theory* of Machine Learning

Contents

1 Useful Resources

- I found this **blog** by Tobias Sterbak, which seem to cover a lot of **advanced topic in machine learning/deep learning, natural language processing, and computer vision** with both theory and implementation aspects treated:
  `https://www.depends-on-the-definition.com/classify-toxic-comments-on-wikipedia/`
- There is this **coursera mook about NLP**:
  `https://www.coursera.org/learn/language-processing`
- There is this **repo** from Sebastian Ruder which keeps track of the **best models for many NLP tasks**:
  `https://github.com/sebastianruder/NLP-progress`

- There is this **review** by two guys from Microsoft and one from Google about the **state of the art in Conversational AI** (as of 13.12.18):
  `https://arxiv.org/pdf/1809.08267.pdf`
- (17.09.19) I found this blog which provides both **theoretical explanations and implementation tutorials for NLP models**:
  `https://mlexplained.com/2019/`
- (17.09.19) Here is a useful website presenting **papers with their implementation**:
  `https://paperswithcode.com/`
- (09.10.19) **Latest arxiv publications** in specific domains, life for instance:
  `https://arxiv.org/list/cs.CL/recent`

2 Specific technical questions which should be answered

1. (17.09.19) I should understand exactly how BERT can generate an answer. This paper touches the topic (but not in detail):
   `https://arxiv.org/pdf/1906.05416.pdf`

3 Journal

04.11.17 ~ Principal Component Analysis (PCA) and Singular Value Decomposition
PCA comes again and again, so I think I should try to learn it properly. I quickly reviewed what I had already seen in the first lecture of CIL. This being said I'm not use that it is exactly the PCA, since it's called SVD (Singular Value Decomposition). I read this very interesting page which gives a good explanation of the two different quantities and their relations:
Stack Exchange: PCA and SVD

21.02.18 ~ Skip-gram model for words embedding
I read these two pages of a same tutorial on the **Skip-gram** model:
`http://mccormickml.com/2016/04/19/word2vec-tutorial-the-skip-gram-model/`
`http://mccormickml.com/2017/01/11/word2vec-tutorial-part-2-negative-sampling/`
Edit (27.07.18): The skip-gram model is also presented in the coursera mook on NLP, in a quite different way.
Edit (21.10.18): Basically, the skip-gram model does the following: it builds a neural network to predict for a given word, its context. The input is a one-hot encoded vector representing a word and the output is a vector of probability of size equal to the size of the vocabulary (let's call it n) indicating the probability of having any word as neighbor. The embeddings are obtained by taking the intermediary layer (which has no activation

function and can be seen as matrix of size $n \times 300$ (if the embedding dimension is 300)).

14.03.18 $\sim$ Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM)
I read this introduction to **Recurrent Neural Networks in NLP**, and their applications to **Language Modeling and Generating Test**, **Machine Translation**, **Speech Recognition**, **Generating Image Descriptions** (when coupled with some CNN):
`http://www.wildml.com/2015/09/recurrent-neural-networks-tutorial-part-1-introduction-to-`
I also read this page on **Long Short Term Memory (LSTM)** networks:
`https://colah.github.io/posts/2015-08-Understanding-LSTMs/`
which is a very successful special case of RNN.

17.03.18 $\sim$ Bag of words model
I read an introduction to the **Bag of word model (BOW)** here:
`https://machinelearningmastery.com/gentle-introduction-bag-words-model/`
I also read the beginning of the "Example: Logistic Regression Bag-of-Words classifier" here:
`http://pytorch.org/tutorials/beginner/nlp/deep_learning_tutorial.html`

21.03.18 $\sim$ Character embedding vs Word embedding
I read this post on Quora which compares word embedding and **character embedding**:
Quora: How does character embedding work in comparison to word embedding?

22.03.18 $\sim$ More on character embedding, GloVe
I read the beginning of this **article on character embedding**:
`http://minimaxir.com/2017/04/char-embeddings/`
but I didn't find it so good when it comes to explaining what character embedding is. But the little implementation with Keras seems like a good exercise. Maybe I could reproduce something like this with the dataset of fast ai.
I read this page about *GloVe*:
`https://nlp.stanford.edu/projects/glove/`
I think it is just a library doing this word embedding. It looks like the fundamental ideas are the same as for Word2Vec, i.e. train a model to predict co-occurrence of words (and then use the weights obtained as your embeddings).
The I read this page:
analyticsvidhya.com: Word representations-text classification using Fasttext (an NLP library from facebook)
The introduction explains quite well what **character embedding** is and what are its advantages compared to classical (i.e. Word2Vec already mentioned above) word embedding techniques. Then it dives into how to use this library. Maybe I should try

to use it on the data set from fastai.

23.03.18 ∼ Continuous Bag of Words
I learned a bit about the **continuous bag of word** model is used which differs from the skip-gram model (presented in this page that I had already read) as explained in this page from Quora:
Quora: What are the continous bag of words and skip-gram architectures?

02.04.18 ∼ Constituency parser, Conditional Random Fields (CRS), Viterbi algorithm
I learned a bit about **constituency parser** here:
Stackoverflow: Difference between constituency parser and dependency parser
I read this introduction to **Conditional Random Fields (CRF)**:
`http://blog.echen.me/2012/01/03/introduction-to-conditional-random-fields/`
But this page doesn't mention any neural network.
Edit (28.10.19): The first thing to understand is that the goal is to assign a probability to a sequence of tags for a sequence of items (typically a sequence of words), and its strength comes from the fact that it tries to compute the **joined** probability of these items (in the specific given order).
I read the description of the **Viterbi algorithm** on wikipedia:
Wiki: Viterbi algorithm
From what I understood, it is just a kind of "dynamic programming" algorithm where instead of computing the probability of every possible sequence, we first find for each possible stae the most likely sequence of length 2 finishing by this state, then use it in order to find the most likely sequence of length 3 finishing by every possible state and so on.
Edit (30.04.18): In this paper it is explained (unfortunately not in a detail fashion) how to combine a (Bi)-LSTM network with a CRF:
`https://arxiv.org/pdf/1508.01991.pdf`
I also read this page which gives **an intuition of the Viterbi algorithm**:
Quora: What is an intuitive explanation of the Viterbi algorithm?
Edit (02.05.18): I think I understood how one can mix the LSTMs and the CRFs. It is kind of explained in this tutorial:
`https://pytorch.org/tutorials/beginner/nlp/advanced_tutorial.html`
Here are the main points: let say we want to use LSTMs without anything else to predict what should be the labels/tags of each word in a sentence. For each input (each word) the LSTM would output a distribution over the possible labels, and we would take its argmax as the predicted label (for a given word). Let say the words are $x_1, x_2, \dots$ and the labels can be taken from a set $S = \{s_1, \dots, s_N\}$. Now, we can see each entry of this vector of probability (of length N) as one of the weighted "feature function" of this explanation of CRF. For an word/input x_i, entry 2 of the "hidden state" (I don't really like this terminology because with LSTM, we never know if we are talking of h_i or C_i

(following the terminology introduced in this tutorial) but here I mean h_i) outputed by the LSTM is seen as a feature function giving the likelihood of seeing x_i if the tag/label is s_2. Then we add some "transition scores" which are additional feature functions independent from the LSTM and which allows to use the power of usual CRF models.
Edit (03.05.18): Actually, the idea goes beyond superposing the two processes. As I realized in the pytorch tutorial mentioned above, the transition scores/probabilities, are themselves some tensor which are optimized over, when doing backpropagation.
Edit (12.10.18): I guess that both the hidden states produced by the lstm and the transition functions/scores/probabilities can be seen as feature functions, but (and this is where my guess is) the hidden states make use of the information regarding the position in the sentence, whereas the transition functions look at the neighbouring words of the given word. But since the transition functions are also learned, I guess that these special "CRF" are merely an LSTM with an additional layer of weights on top of it.

30.04.18 $\sim$ Chat bots: Tai, Xiaoice, Named entity recognition, A nice blog about NLP
It seems that some fairly advanced chatbots already exist:
Wiki: Xiaoice
Wiki: Tay
I learned a bit about **Named Entity Recognition (NER)** here:
Wiki: Named Entity Recognition
I found this **nice blog (mainly) about NLP**:
`https://www.depends-on-the-definition.com/about/`

01.05.18 $\sim$ TF-IBF, and Latent-Dirichelet-Allocation
Roman mentioned **tf-idf**, **Latent-Dirichelet-Allocation**, HMM, as a classical alternative to NN models, which should be used as benchmarks to measure the success of the NN models.
From the wikipédia page for tf-idf:
Wiki: tf-idf
this tf-idf seems to be used in order to find among a given set of documents, which ones are relevant for a specific query ("the brown cow") for instance. It gives a score to each document.
Edit (25.05.18): Actually TF-IDF is quite well explained in the first week of the NLP coursera course. It's simply a way to improve a bag-of-word model, where instead of entering just 0 or 1 in each column to indicate if a word is present or not, one puts a score which indicates both how much frequent the word is in the given document/sample, how rare the word is in the total corpus of documents/samples.

The **Latent Dirichelet Allocation** is a bit more complicated. From the wikipedia page:

"In natural language processing, latent Dirichlet allocation (LDA) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. For example, if observations are words collected into documents, it posits that each document is a mixture of a small number of topics and that each word's creation is attributable to one of the document's topics." The math behind it is explained on the wikipedia page. From what I saw, it seems to be a bit similar to the bayesian nets studied in the course of Krauser in the sense that we make some (smart) assumptions on the distributions of the words in a corpus of documents, and then using algorithm to estimate the parameters of the distributions. But the wikipedia page is very long and technical. Maybe the coursera course on NLP would be better.

06.05.18 ~ Statistical Parsing
I read this wiki article about **Statistical Parsing**:
Wiki: Statistical Parsing

25.05.18 ~ Useful metrics for binary and multi-label classification, ROC curve
In the notebook of the exercise of the first week, I (re)discovered the notion of **F1-score**:
Wiki: F1 score which is a useful metric for binary or multi-label classification.
I also came back on the notion of **ROC curve**. I was confuse at first because, I had the impression that if we were doing binary classification (I guess we can generalize this discussion easily to multi-label classification) and we have just the predicted class of a model for a given set of data, there is only one True Positive Rate (TPR) and one False Positive Rate (FPR) to compute. But actually, we don't use the predicted classes but the *scores* output by the model, i.e. for each sample the probability to belong to each one of the two classes. Then we can consider for each $c \in [0, 1]$, the samples for which the probability to be in the second class is above c, to be "positive" and the other samples to be "negative", and then compare to the true labels. And this way we can compute this ROC curve.

29.05.18 ~ More about NLP
In the NLP coursera course, I was introduced to various models, some using Neural networks, some using classical algorithms.
In particular, I learned that the way to evaluate language models is based on the notion of **Perplexity** (c.f. journal NLP coursera for more infos).

30.05.18 ~ Dropouts for regularization of LSTMs
I read this article (I have it on my computer):
`https://arxiv.org/pdf/1409.2329.pdf`
explaining how one can use dropouts (i.e. randomly put some coefficients of the NN

to 0 during the different training phases) to **avoid over-fitting**. In the case of RNN, one has to be careful not to affect the coefficients which manage the transmission of the memory.

31.05.18 ∼ Evolutionary learning: an alternative to gradient descent
I read this article that Mike recommended to me:
`http://togelius.blogspot.com/2018/05/empiricism-and-limits-of-gradient.html`
It presents some alternative technique to gradient descent to come up with a model achieving good performance. The basic idea is to start with a bunch of different models, test them, throw away the one performing the worst, modify randomly the others and combine them, and then repeat the procedure. The advantages are that you don't need differentiability of the loss function, and according to the author it is more likely to learn "something which isn't in the data" like a mathematical formula. At the end of the article there is a list of recent paper on the topic by some AI labs like Uber AI, Sentient Technologies, DeepMind, and OpenAI.

01.06.18 ∼ Gradient clipping
I learned about **Gradient Clipping** there:
hackernoon: Gradient clipping
It seems to come down to putting a bound on the value of the norm of the gradient to avoid NaN appearing. It seems to be useful for RNN:
Stack Overflow: Why do we clip by global norm to obtain gradients while performing RNN?

03.06.18 ∼ Truncated Back Propagation Throw Time (TBPTT) for RNN
I read this article which explains some tricks used to change a bit the update scheme of the parameters when using RNN:
`https://machinelearningmastery.com/gentle-introduction-backpropagation-time/`
It is used in the RNN TF tutorial.

10.06.18 ∼ permutation importance: a criterion to assess importance of features
I read this article about **permutation importance**:
`http://parrt.cs.usfca.edu/doc/rf-importance/index.html`
which explains that there is a better criterion that the "impurity decrease" called "permutation importance" to **judge how much a feature is useful**. This is particularly easy to use in the random forest case (because we have some OOB samples), but it can always be used if one has a validation set. I think it is implemented in scikit learn random forest estimator.

14.06.18 ∼ Encoder-decoder

I read this Quora page about **encoder decoder Network**:
https://www.quora.com/What-is-an-Encoder-Decoder-in-Deep-Learning
because I wanted to understand something in the "lesson4-imdb" notebook of fastai (see entry of the 15.06.18 of the fastai journal, for details).
Edit (12.10.18): From what I understand an encoder-decoder is just a way to go from one initial representation which is the input of the encoder (for instance a sequence of bag of word vectors) to an intermediate one which is both the output of the encoder and the input of the decoder (for instance a vector) and then to a third one (which is for instance a sequence of words). Autoencoders are just NN which are typically stacks fully connected layers which have to predict their (cathegorical) input. By keeping only the first layers, we build a useful way to represent the data in a low dimensional space (and we build this way an encoder). One can have a CNN as encoder and RNN as decoder as explained in the link above.

15.06.18 ~ Simple but efficient models for sentiment analysis
I read this article:
https://www.aclweb.org/anthology/P12-2018
(I have it on my computer) which shows that some simple methods can be used efficiently for sentiment analysis on texts. It provides comparison tables for different common data set. I discovered it in the "nlp" dataset from fastai.

16.06.18 ~ Stop words
Stop words are words which are removed from a text before it is further processed. There are some lists of them (for instance in the "english" file of my NLTK package, or on this page) but they are not the same. Plus sometimes, it could actually harm the model to remove some common words (typically negations) which have a strong impact on the meaning of a sentence, if one wants to perform sentiment analysis. Some people suggest not to remove them at all:
Why sentiment words are in stopwords list?
Is there a stopword list specifically designed for sentiment analysis?

27.06.18 ~ Hyperparameter tuning for RNN models, measures against overfitting
There is an interesting section in this tutorial:
https://www.oreilly.com/learning/perform-sentiment-analysis-with-lstms-using-tensorflow
about how to **tune your hyperparameters** for a RNN.

In the same tutorial, the author explains how one should fight against overfitting: "The basic idea is that we train the model on our training set, while also measuring its performance on the test set every now and again. Once the test error stops its steady decrease and begins to increase instead, you'll know to stop training, since this is a sign

that the network has begun to overfit."

14.07.18 ~ Convolutional Neural Network for text classification
I read this paper about the use of **CNN for text classification**:
`https://arxiv.org/abs/1412.1058`
There are two different ways to do it. Either do the convolution directly on texts converted to sequence of one-hot-encoded vectors representing the words, or something a bit more subtle where each local region where a convolutional kernel would be applied is encoded as a bag of words.

16.04.18 ~ Difference Valdiation and Test set
I read this page which clarified what was the distinction between the **test set and the validation set**:
Stack Exchange: What is the difference between test and validation set

06.08.18 ~ Multi-label and multiclass classification
I learned a bit about multi-label and multiclass classification:
`https://en.wikipedia.org/wiki/Multiclass_classification#One-vs.-rest`
`https://en.wikipedia.org/wiki/Multi-label_classification`
which are not exactly the same thing. But the approaches are similar. Interestingly, the canonical methods for each one (One-vs-rest and binary relevance methods) are implemented in the same python class in Scikit-Learn: sklearn.multiclass.OneVsRestClassifier.

21.08.18 ~ Tuning a model
In order to tune a deep learning model, here are the recommendations from Roman:
"I start with a model config that I expect to overfit and slowly reduce nn complexity until validation performance stops improving. As for non nn parameters as learning rate, I pick them after I settle on network topology, I also pick them manually with some sort of coordinate gradient descent (picking them on by one) or using the defaults if I think the parameter is not important".

This link also gives some instructions to tune an XGBoost model:
`https://machinelearningmastery.com/xgboost-python-mini-course/`

12.10.2018 ~ StarSpace

This paper explains how **StarSpace** is working:
`https://arxiv.org/abs/1709.03856`
I had encountered it already when doing duplicate detection in the NLP coursera course. It seems to be a way to build some custom embeddings. I found again a reference to it on this page explaining how the TensorFlow embedding pipeline of RasaNLU works.

Edit (08.10.19): I looked at the paper again. The main idea is that depending on the task, we define a set of (discrete) features (for text classification, the features would be all words into a fixed English dictionary, and document labels), and then we train embedding for these features. Then we define "entities" as bag of features. A text to classify would have as features the words it contains, and the document labels would have only one feature each, the feature corresponding to the given labels. Then one trains the features embeddings using a special loss function based on similarity, and positive/negative examples. Later we can try for a new text to classify it by looking at the labels which have an embedding similar to the embedding of the new document (obtained by summing the embedings of the words it contains). One other application is to try to **predict relations in a graph**. I was thinking that it's kind of pointless since one can query the graph, but then I realized that it could be used to **guess relations which aren't explicitly encoded in the graph**, if the graph is big enough and contains some recurring structure.

21.10.18 ~ Parallel SGD computation with Hogwild and others

I read the beginning of this page concerning **asynchronous stochastic gradient descent algorithm** performed with **Hogwild**:
`https://srome.github.io/Async-SGD-in-Python-Implementing-Hogwild!/`
which offers also some sample code in python to implement it. At the end of the article, the author points out to other methods.

01.11.18 ~ Neural Turing Machines

I read this page about **Neural Turing Machines (NTM)**:
`https://blog.acolyer.org/2016/03/09/neural-turing-machines/`
and even if I didn't understand every detail, it was interesting. One has to see how far these Machines can learn. The article focuses on some very basic tasks like copying. This topic was also mentioned in the course about Synthesis and other topics of ETH. This page:
`https://medium.com/snips-ai/ntm-lasagne-a-library-for-neural-turing-machines-in-lasagne-`
gives much more insight.

13.02.19 ~ Levenshtein distance

I watched this video introducing the **Levenshtein distance** which gives a **distance** between two finite **sequences**:
`https://stackabuse.com/levenshtein-distance-and-text-similarity-in-python/`

17.03.19 ~ Tree LSTM

I read the beginning of this article about **tree LSTM**:
https://arxiv.org/pdf/1503.00075.pdf
which are a bit like LSTM, but the memory/state cell associated to element/word i of the sentence is computed using the memory/state cell and hidden states of the elements/words which "children" of i ($C(i)$). The set $C(i)$ is computed using a tree which is also given as input to the model (typically a consitutency or dependency tree). This seems to be used in the paper that Aurélien used for his master thesis:
https://arxiv.org/pdf/1609.06038.pdf

23.03.19 ∼ Capsule networks

This blog gives a good introduction to **capsule networks**:
https://medium.com/ai%C2%B3-theory-practice-business/understanding-hintons-capsule-netwo
Maybe I could use them to generate animated characters from static picture (front, side, back).

26.03.19 ∼ Multi-Task Deep Neural Network for NLP

Rolland presented a paper called **Multi-Task Deep Neural Network for NLP**:
https://arxiv.org/pdf/1901.11504.pdf

31.05.19 ∼ Sequence to sequence translation, Attention, Transformer, and Bert

I started reading this very well illustrated tutorials:

- **Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention):** https://jalammar.github.io/visualizing-neural-machine-translation-mechan
- **Transformer:** http://jalammar.github.io/illustrated-transformer/
- **BERT:** http://jalammar.github.io/illustrated-bert/

Since there is a lot to say abou each model, I will split this entry into sections that I will complete little by little.

3.1 Neural Machine Translation Model (Mechanics of Seq2seq Models With Attention)

- In the sequence to sequence translation with attention model, the input is a sequence (of words in English for instance) and the output is another sequence (of words in French for instance), not necessarily of the same length.

- In the tutorial about sequence to sequence translation with attention model, in the illustration called Neural Machine Translation, the pink output of the Dense layer after the first unit of the RNN (1) is not exactly the same as the input of the second unit of the RNN (2) even if they are depicted in the same way. Most likely (1) is a vector giving probabilities for each word in the vocabulary, whereas (2) is a (fixed) embedding for the word found using (1).
- In the tutorial about sequence to sequence translation with attention model, I guess that the attention coefficients are probably computed in a similar way as in the Transformer tutorial.
- **Edit (06.08.19):** I found this other page which explains with more details the attention mechanism:
`https://distill.pub/2016/augmented-rnns/#attentional-interfaces`

3.2 The Transformer

- In the tutorial about the Transformer, the **input** is a sequence and the **output** is a sequence (like in the sequence to sequence translation with attention tutorial).
- Unlike in the attention tutorial, the model of the Transformer, doesn't seem to use LSTMs. Instead the encoder layers receive a list of vectors of fixed length (this is a hyper parameter and I guess that padding is used) containing all the input as once, and outputs a list of vectors containing all its output at once. It has a component called **self-attention layer**, which doesn't seem (a priori) to process input with an order. The function which takes this list as input takes the whole list, i.e. it is not repeatedly applied to every vector in the list individually. Indeed this self-attention mechanism uses every element of the input list to compute each element of the output list. In order to help the model better learning to make use of the relative positioning of the elements of the input sequence, each initial input word/object vector is concatenating with another vector encoding only the position of the word/object.
- Looking at this Transformer encoder, one thing which wasn't clear to me at first was how we can take a sentence of length n and output a sentence of size different from n. When we use an LSTM like in the sequence to sequence translation with attention model, it is clear since we can let run the decoder as long as it hasn't predicted the "stop" character (which is added to the end of each sentence of the training data). In the case of the Transformer, I guess that **a sequence cannot be longer than the predefined fixed size** of the list of vector given as input to the self-attention layer, but that discarding the elements of the output list which will be predicted as padding elements, we can have shorter sentences.
- I think that the reason why some layers are considered to be part of the Encoder and some of the Decoder is because the one in the Decoder only receive the output of the last layer of the Encoder.

- In the Transfomer model, the Decoder had an architecture different from the Encoder. At the **decoder** level, there are **two attention mechanism** used: the first one is called **(masked) self-attention layer** where "all of the keys, values and queries come from the same place, in this case, the output of the previous layer in the (decoder)" (p.5 of "Attention is all you need"), and the second one **encoder-decoder attention layer**, where "the queries come from the previous decoder layer, and the memory keys and values come from the output of the encoder" (p.5 of "Attention is all you need"). For the encoder-decoder attention layer, I guess that the memory key, value, and query matrices are specific to the given layer (not shared with any other) but that the vectors used to generate the memory key and value vectors are the output of the encoder, whereas the vectors used to generate the query vectors are the output of the previous decoder layer. For the self-attention layer, they explain that the attention at position i is restricted to positions 1 to $i-1$ (to prevent the mechanism to simply learn to copy rather than predict). From what understand, if I use the notation of the tutorial $q_m * k_n$ is set to $-\infty$ if $n > m$. This implies than the the weight for this position is 0 after we take the soft-max. My question is: is that differentiable? But for the encoder-decoder attention layer, in order to compute the new output vector at position i, the output vectors of the previous layer for all positions can be used. I understand that if we train a model to do language modeling (guessing what is word at position i looking only at the outputs until position $i-1$), where we try to rebuild the input sentence, the model could otherwise simply learn to copy the input embeddings. But if we try to translate English to French, the reason is less obvious. The authors of the paper justify it by stating: "We need to prevent leftwardinformation flow in the decoder to preserve the auto-regressive property." I guess that it somehow forces the model to make smart embeddings.

- In the paper , there is a very intresting remark about **long-term dependency**: "The third is the **path length between long-range dependencies in the network**. Learning long-range dependencies is a key challenge in many sequence transduction tasks. One key factor affecting the ability to learn such dependencies is the length of the paths forward and backward signals have to traverse in the network. The shorter these paths between any combination of positions in the input and output sequences, the easier it is to learn long-range dependencies. Hence we also compare the maximum path length between any two input and output positions in networks composed of the different layer types."

- The Transformer model presented in the tutorial mentioned above appeared in the paper *Attention is all you need* here:
`https://arxiv.org/pdf/1706.03762.pdf`
where there is no language modeling task and where the goal is to do English-to-French translation, but the idea of using it as a basis for leveraging an **unsupervised-learning task (language modeling)** and then fine-tuning it for other specific tasks appear in a paper called *Improving language understanding by Generative*

pre-training (**GPT**) from OpenAI, here:
`https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervi`
`language_understanding_paper.pdf`
They actually explain in this paper, that their approach (training in an unsupervised manner a model for language modeling and fine-tuning in a supervised manner for a specific NLP task) is very close to ULMFit presented by Howard and Ruder, here:
`https://arxiv.org/pdf/1801.06146.pdf`
with the difference that they use the Transformer architecture instead of the LSTM. In the *Improving language understanding by Generative pre-training* article, the authors explain that they didn't use the original architecture of the transformer but a modified version called `Transformer Decoder` presented in the paper *Generating Wikipedia by summarizing long sequences* here: `https://arxiv.org/pdf/1801.10198.pdf`
(c.f. p.5). They explain there that they use only the decoder part. This makes me wonder what happens with the encoder-decoder attention sub-layer. Where are the queries computed from? Do we use the vectors of the previous decoder layer? Or do we remove the encoder-decoder attention sub-layers? If we remove them, why do we say that this is the decoder part and not the encoder part? Is it because we keep the mask thing (which is actually modified in this paper *Generating Wikipedia by summarizing long sequences* and replaced by two alternatives)? But coming back to using the transformer to get good embeddings, I don't understand why we would keep the masks if we want to generate embeddings. On the opposite, we would like that every token gets enriched by the whole context. Personally I would train the encoder together with the decoder, and then use only the encoder... It isn't clear at all in the *Improving language understanding by Generative pre-training* article where the explanation is very shallow. I guess I should look at the implementation. I could look at it on the website paper with code:
`https://github.com/openai/finetune-transformer-lm`
but unfortunately it is in tensorflow which is a pain in the ass.
Edit (25.09.19): Joram explained me that actually, what is used is indeed just the decoder part but the layers don't have the encoder-decoder attention sublayers. They keep the mask because it forces the model to have the "auto-regressive" property, so in a sense, to predict one word after the other.

- In the paper *Improving Language Understanding by Generative Pre-Training* (**GPT**), they present the idea of pretraining a transformer-like model on language modeling. I guess that for an input sequence there is an output sequence where output with label k is the predicted $k+1$ input token, and since there are mask layers, input $k+1$ and following cannot be used to compute the value of output k. But all tokens are predicted at the same time.

- From what I understand, unlike in ELMo, only the output of the last transfromer layer is used as input for the downstream task.

- One of the strength of the Transformer, is that the predictions for the different masks tokens can be done in **parallel**.

3.3 ELMo

ELMo is one of the ancestors of BERT. From what I understood, BERT is a kind of ELMo where the LSTM has been replaced by a Transformer. The paper is here:
`https://arxiv.org/pdf/1802.05365.pdf`
Edit (10.10.19): I read it again and the idea is simple: create contextual embeddings for each input token of a sentence by taking linear combination of the output of the different layers of a multi-layer bi-lstm pretrained on language modeling task.
The language modeling task goes like this: there are to LSTMs, a forward one and a backward one. In order to train the model to produce a vector embedding for token i, the first $i-1$ tokens are passed recursively to the L-layers LSTM, and parallely, tokens $i+1$, $i+2$, ..., n are passed recursively (starting by the last) to the L-layer backward LSTM. The output vecors of both are then fed to a feed forward network with last layer having output size equal to the size of the dictionary. Finally a soft-max layer is applied, and Cross-Entropy loss is used as a loss function to determine if the prediction was good or not.
The input is a sentence of token and the output is a sequence of vectors (embeddings). Actually all intermediary layers of the bi-lstm can be used to create the final embeddings using a learned linear combination of them. The weights of the linear combination depend on the downstream task and are chosen during fine-tuning.
I think that there won't be any problem of model learning to copy the input because the forward lstm and the backward lstm stay separate, and receive different inputs. The second layer of the forward lstm doesn't see the out put of the first layer of backward lstm.

3.4 BERT

- BERT seems to a priori be a sequence to sequence translation model in the sense that it takes a sequence as input and output a sequence of vectors. I guess that the length of both input and output sequence is fixed and that padding is used. In order to turn it into a classifier, the authors add a first [CLS] symbol at the beginning of the input sequence. The first vector in the output sequence is then used to predict the class (with an additional simple feed-forward network with soft-max).

- I think that like GPT, BERT uses only the output of its last layer for the downstream task (I think that this is what is meant by "BERT and OpenAI GPT are fine-tuning approaches, while ELMo is a feature-based approach".

- In the way it is used, BERT is similar to ELMo and GPT, but tries to take strengths from both models: it uses (decoder) transformer blocks like GPT, but doesn't stick to forward language modeling like GPT (so is closer to ELMo in this sense). In

GPT, the self-attention layers can use only input at positions $1, \ldots, i-1$ to predict the output at position i. If the self-attention layer was allowed to use the inputs of position $1, \ldots, i-1$ and $i+1, \ldots, n$ in order to predict the output for position i, then the next layer could indirectly "see" the input of the first layer at position i by looking at the output of the first layer at positions different from i. In order to avoid this problem, 15% of the input tokens are randomly replaced by a special MASK token, and the loss is only computed for the prediction of these tokens (the prediction should match the original token).

- Additionally to language modeling, BERT also has a "binarized next sentence prediction task" as pretraining task. There, it has to find if two sentences given in inputs and separated by a special token, are consecutive or not.

04.06.19 ~ Batch Normalization, Weight Normalization, and Layer Normalization

I read the two following posts about different types of normalization which are supposed to speed up the training of NN:
`http://mlexplained.com/2018/01/10/an-intuitive-explanation-of-why-batch-normalization-re`
`https://mlexplained.com/2018/01/13/weight-normalization-and-layer-normalization-explaine`
It seems to be based on the following papers:
`http://proceedings.mlr.press/v37/ioffe15.pdf`
`https://arxiv.org/pdf/1607.06450.pdf`

06.06.19 ~ Resiudal Connection

Here is a page explaining what **Residual Connections** are in NN architecture. From what I understand, it just means adding the input of a layer to its output (a bit like in a lstm):
`https://www.quora.com/How-does-deep-residual-learning-work`

28.06.19 ~ Inductive vs Transductive learning

I learned here the difference between **inductive learning** and **transductive learning**. It seems to me that inductive learning is the usual approach where we try to build a classification/regression function f using only the training set, whereas, for transductive learning, we already have the full test set and might use it (as a whole) to build an algorithm (which will of course also use the training set) which will give the right labels to the inputs of the test set.

19.07.19 ~ Ressources for semantic parsing

This paper (that I haven't read yet) gives a **good historical overview of semantic parsing**:
`https://arxiv.org/pdf/1812.10037.pdf`
Edit (06.11.19): I read the section 2, "Related Work". It's indeed quite detailed but I would need to read the papers mentioned there to really have an overview. and this paper gives a good introduction to **logic based formalism**, **Graph based formalism**, which can both be used for semantic parsing:
`https://openreview.net/pdf?id=HylaEWcTT7`

21.07.19 ~ Common sense with Concept NET

I read in the description of the COIN (EMNLP 19) workshop that **Concept NET** can be used to model **common sense**:
`http://conceptnet.io/`
the Read The Web project is also mentioned:
`http://rtw.ml.cmu.edu/rtw/`

30.07.19 ~ Lambda Calculus, Haskell and Category, Homotopy Type Theory

I read part of this wikipedia page on **Lambda Calculus**:
`https://en.wikipedia.org/wiki/Lambda_calculus`
what I get out of it is that you can write a sequence of operation on a given type of input using this formalism. The specific low-level operations that one performs can be arbitrary (I guess) as long as there are preliminary set. I think that this formalism can be used to write trees or grammar trees (Bilyana had explained me something like this with map-reduced).

I met a man called Declan, working for the UK government who told me that **Category Theory** can be used with Haskell a "purefly functional programming language". He explained me that a map function can be seen as a functor. Here is a stack overflow page about the link between category theory and haskell:
`https://stackoverflow.com/questions/57128213/what-concept-in-category-theory-can-be-used`
Declan also mentioned **homotopy type theory**:
`https://en.wikipedia.org/wiki/Homotopy_type_theory`
Edit (22.10.19): Sarvesh pointed me toward this book:
`https://github.com/hmemcpy/milewski-ctfp-pdf`

03.09.19 ~ Generating Logical Forms from Graph Representations of Text and Entities

Joram presented this paper:
https://arxiv.org/pdf/1905.08407.pdf which adapts the concept of Transformer to see it as a way to embed a fully connected graph and transforms, layer after layer, in the encoder the representation of it nodes. In the described case the inputs are the list of token in a natural language sentence, and a list of possible entities (with possible overlaps or wrong entities), and the output is a query for a graph data base (the graph is given). One of the key ideas is to compute differently the attention depending on the two types of nodes (for which the attention is computed).

17.09.19 ∼ Introduction to XLNet

I started reading this article which introduces **XLNet**:
https://mlexplained.com/2019/06/30/paper-dissected-xlnet-generalized-autoregressive-pret
That made me ask myself the following question: how are exactly concatenated the intermediate hidden outputs of the previous layers with the hidden outputs of the current layers? I guess that instead of having n different (customized) word embeddings, we have 2n, but that one computes new vectors only for the n new (the n old are only used to compute keys and values).
The key points I take out of this are:

1. XLNet is based on **Transformer XL**. This uses a recurring structure to **get rid of the limitation on the size of the input** of the vanilla Transformer. It also replaces the mask mechanism by **Permutation Language Modeling**. The ability to parallelize the computation of the prediction for each input token is lost but some problems (masks do not appear at testing time, predictions for two masks tokens are independent) are resolved. It is not absolutely clear to me if the authors of XLNet are using Transformer XL as such or only take inspiration from it.

10.10.19 ∼ Loss function for classification: Cross Entropy

I came back to the topic of the loss function since it's a crucial point when it comes to the training of these NN (using SGD). I read this interesting post:
https://rdipietro.github.io/friendly-intro-to-cross-entropy-loss/
which explains that cross-entropy is related to the idea of **compressing information** and also to the idea of **likelihood**, which makes it a good loss function. If there are n examples the loss will be

$$-\sum_{i=1}^{n} \log(\mathbb{P}\left(\{\text{prediction for sample } i = \text{ground truth for sample } i\}\right)$$

15.10.19 ∼ ERNIE and ERNIE 2.0

I started reading the paper about **ERNIE 2.0** coming from Baidu:
`https://arxiv.org/abs/1907.12412`
For this I had to look at the first version of ERNIE:
`https://arxiv.org/pdf/1904.09223.pdf`
What I get from ERNIE is that replace the masking procedure of BERT by more sophisticated ones. First, there are **several pretraining-task**, all **similar to word prediction**, that they regroup under the name of **knowledge integration**. The first one (**Basic-Level Masking**) is similar to what's done in BERT. The second (**Phrase-Level Masking**) consists in masking (logically coherent) parts of sentences. And the third one (**Entity-Level Masking**) consists in masking entire entities. But then one obviously needs more than simple text. One needs tagged text for entities, and a tool to cut (in a good way) parts of sentences. Let's note that all these tasks were performed on English and **Chinese** as well.
ERNIE 2.0 generalize the idea by adding new pretraining tasks. They are split in 3 categories:

1. **Word-aware**
    - Knowledge Masking Task (this covers all 3 original tasks of ERNIE 1.0)
    - Capitalization Prediction Task
    - Token-Document Relation Prediction Task
2. **Structure-aware**
    - Sentence Reordering Task
    - Sentence Distance Task
3. **Semantic-aware**
    - Discourse Relation Task
    - Information Retrieval (IR) Relevance Task

The Information Retrieval Relevance Task is based on a dataset created by Baidu search engine.