



BERT

Presentation

Kalle Prorok

2020-10-01



Vad är BERT?

- **Bidirectional Encoder Representations from Transformers**
- En språkmodell/language model
- Natural Language Processing (NLP), texthantering
- To better understand user searches; meningar istf ord
- Context-free models such as [word2vec](#) or [GloVe](#) generate a single word embedding representation for each word in the vocabulary, where BERT takes into account the context for each occurrence of a given word – orden i sitt sammanhang
- Jacob Devlin and his colleagues from Google 2018

Träning av BERT

- Tittar på meningar där vissa ord är bortmaskade (gömda)
 - Gissa det saknade ordet
- It has been pre-trained on a lot of words – and on the whole of the English Wikipedia 2,500 million words
- 72 språk 9 dec 2019
- Ny hårdvara: Cloud TPU – Krävande (då)
- Använder delar av ord, ”stavelser”
- Typisk storlek: Max 512 stavelser i rad.

Transformers



- Traditionella RNN har svårt att klara längre sekvenser av ord
- In each step, it applies a self-attention mechanism which directly models relationships between all words in a sentence, regardless of their respective position.
- for a given word - “bank” for example - the Transformer compares it to every other word in the sentence. The result of these comparisons is an attention score for every other word in the sentence. These attention scores determine how much each of the other words should contribute to the next representation of “bank”.

Attention is all you need



- The attention scores are then used as weights for a weighted average of all words' representations which is fed into a fully-connected network to generate a new representation for “bank”, reflecting that the sentence is talking about a river bank.

(från <https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>)

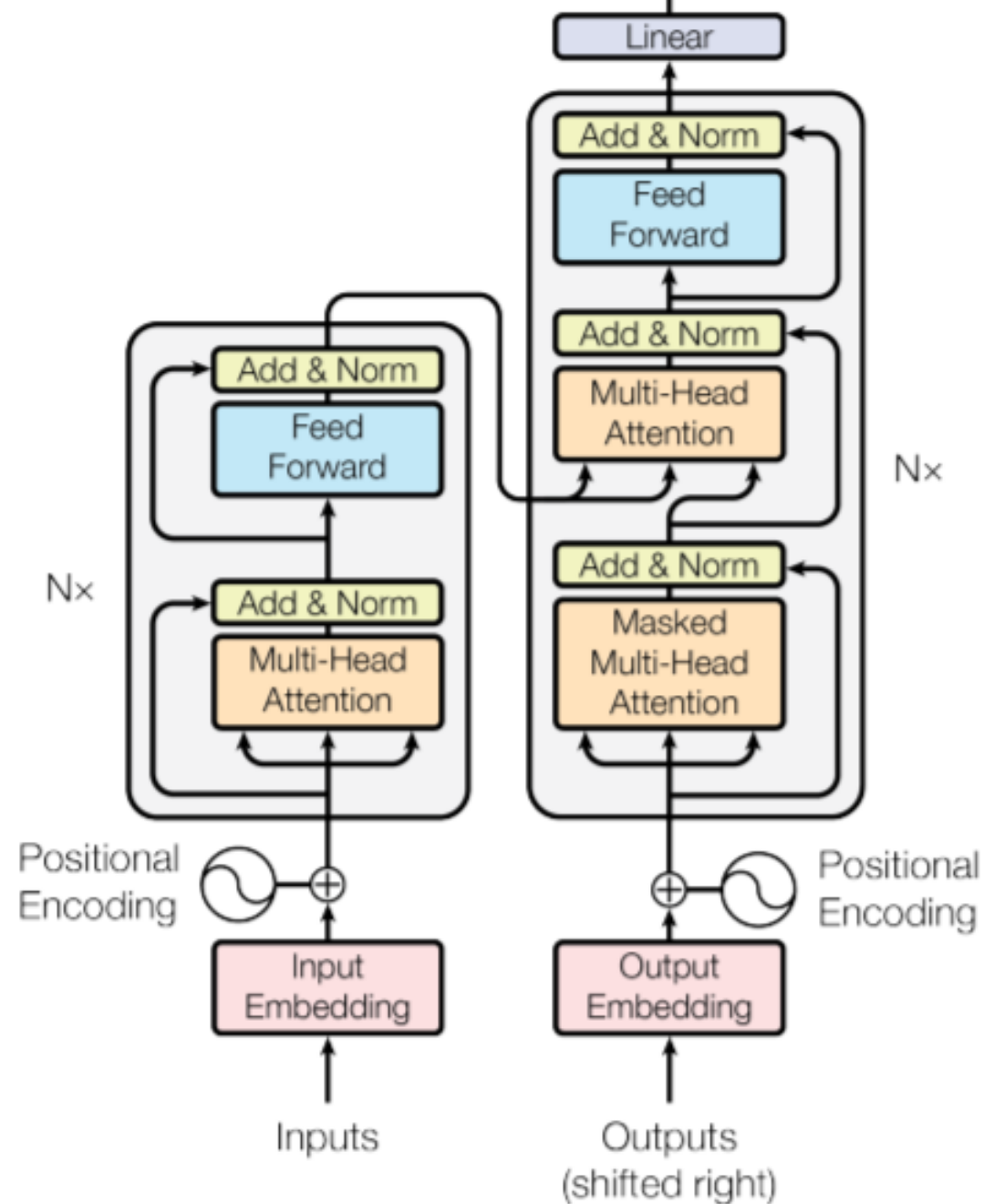


Figure 1: The Transformer - model architecture.



Vad är BERT bra på?

- Named entity determination.
- Textual entailment next sentence prediction.
- Coreference resolution.
- Question answering.
- Word sense disambiguation.
- Automatic summarization.
- Polysemy resolution.

What is BERT good at?

- Namngivna enhetsbestämning.
- Textförutsägelse nästa mening förutsägelse.
- Coreference-upplösning.
- Frågeställning.
- Förtydligande av ordkänsla.
- Automatisk sammanfattning.
- Upplösning av polysemi.

Named Entity Recognition (NER)

- Justera BERT/Fine-tune BERT
- NER är baserat på taggade ord/stavelser (WordPiece)
- “B” or “I”, which stand for “beginning” and “inside.” Example: “Charlene Chambliss” would be tagged as (B-PER, I-PER).
- Exempel: Platser, Tider, Organisationer, Personer, Mått, ..
- https://en.wikipedia.org/wiki/Named-entity_recognition



Datasets, Verktyg och artiklar för/om NER

- Exempel på dataset
 - <https://github.com/JohnSnowLabs/spark-nlp/tree/master/src/test/resources/conll2003>
- Spark NLP
 - <https://nlp.johnsnowlabs.com/>
- <https://github.com/docco/docco>
- <https://github.com/DataTurks>
- På Svenska,
 - KTH, <http://kth.diva-portal.org/smash/get/diva2:1451804/FULLTEXT01.pdf>
 - Kungliga biblioteket, <https://huggingface.co/KB/bert-base-swedish-cased-ner>

Huggingface



- <https://huggingface.co/>
- " advance and democratize NLP for everyone. Along the way, we contribute to the development of technology for the better."
- Transformers
- Vårdar för språkmodeller



KB/Sics/Peltarion

- Kungliga Biblioteket
- Arbetsförmedlingen
- Sics
- Peltarion
- Samarbete
- Lab: Starta colab, colab.research.google.com
 - välj Arkiv, Öppna anteckningsbok, välj Github och ange nutte2 och välj sedan filer från BERT. Välj filen KBBERT.ipynb, ange GPU och Kör alla.
 - Långsamt med långa dokument

Sammanfattningar i BERT

- works by first embedding the sentences, then running a clustering algorithm, finding the sentences that are closest to the cluster's centroids. This library also uses coreference techniques, utilizing the <https://github.com/huggingface/neuralcoref> library to resolve words in summaries that need more context.
- SpaCy, neuralcoref
- <https://pypi.org/project/bert-extractive-summarizer/>



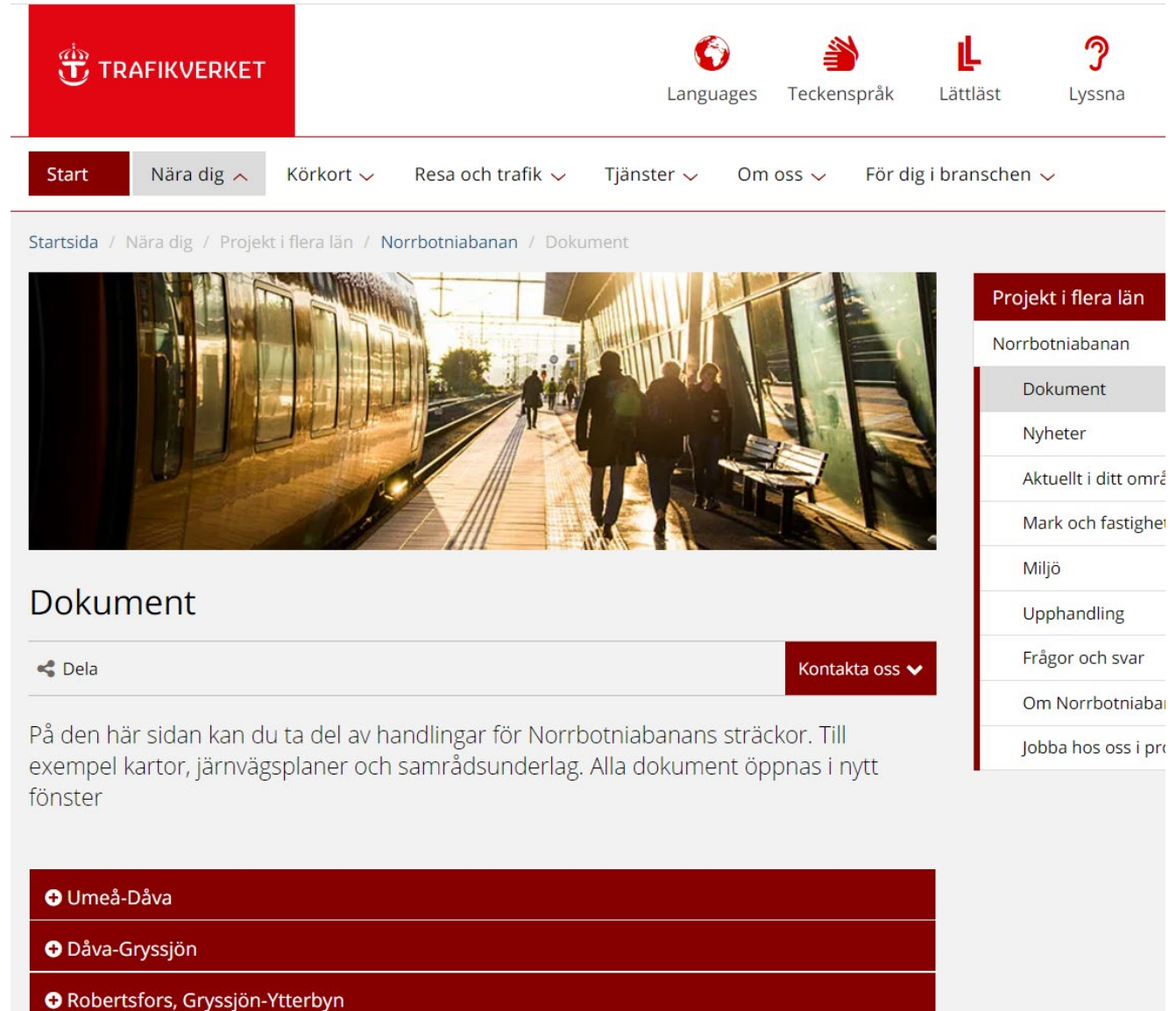
Tensorflow/PyTorch

- Numera en tuffare konkurrens
 - PyTorch växer i popularitet
- Tensorflow från Google
 - Keras
 - Enkelt för vanliga AI/DL-uppgifter
 - Krångligt om man vill göra något annat
 - <https://www.tensorflow.org/>
- PyTorch från Facebook (Microsoft)
 - Man får göra en del själv, tröskel
 - Lättare att skapa specialare
 - <https://pytorch.org/>



Läsa in filer/hämta från Web

- Mycket tidsbesparande om det är många filer
- Kan arbeta med olika delmängder
 - Texter, kartor, planer
- Urllib3
- Bs4 (BeautifulSoup)
- Ofta strul med sidor, rubriker, sidnummer etc
- <https://www.trafikverket.se/nara-dig/projekt-i-flera-lan/Norrbotniabanan/Dokument/>



The screenshot shows the Trafikverket website. The header includes the Trafikverket logo and navigation links for Languages, Teckenspråk, Lättläst, and Lyssna. The main navigation bar has links for Start, Nära dig, Körkort, Resa och trafik, Tjänster, Om oss, and För dig i branschen. The breadcrumb trail reads: Startside / Nära dig / Projekt i flera län / Norrbotniabanan / Dokument. A large image of a train at a station platform is displayed. Below the image, the title 'Dokument' is shown, followed by a 'Dela' button and a 'Kontakta oss' button. The main text states: 'På den här sidan kan du ta del av handlingar för Norrbotniabanans sträckor. Till exempel kartor, järnvägsplaner och samrådsunderlag. Alla dokument öppnas i nytt fönster'. A sidebar on the right lists various project categories: Projekt i flera län, Norrbotniabanan, Dokument, Nyheter, Aktuellt i ditt område, Mark och fastigheter, Miljö, Upphandling, Frågor och svar, Om Norrbotniabanan, and Jobba hos oss i projekt. At the bottom, a list of project segments is shown: Umeå-Dåva, Dåva-Gryssjön, and Robertsfors, Gryssjön-Ytterbyn.

Anropa Google Translate



```
!pip install googletrans
```



```
from googletrans import Translator
```



```
translator = Translator()
```



```
oversatt=translator.translate(text,dest='en')
```



```
print(oversatt)
```



```
print(oversatt.text)
```



Inte samiska (men via Norska)

.pdf/.doc

- Krångliga filer med bilder, tabeller, figurer etc
- Finns många olika bibliotek, dessa är de "bästa" jag hittade
- PyMuPDF/fitz
- docx2python
- Lab: getpdf, getpdfandshow
- Lab: ladda ner TanklabStyrRegler, kör readdoc och läs upp filen (eller en annan) till colab

Rensning

- För att kunna hantera Sammanfattningar, Frågor&svar etc
- Rubriker
- Sidnummer
- Punktlister
- Indelning i block, lämpliga avbrottpunkter
 - Avsnitt
 - Meningar
 - Lämplig storlek (512 för BERT, 384 för Pavlov)
- Tips på verktyg: Tika <https://tika.apache.org/>

Besvara Frågor

- Att ha en lång text och sen kunna ställa frågor på den
 - Enkla frågor – matcha ord
 - Kluriga frågor – behöver ”förstå”/associera
 - Frågor utan svar
- Stanford SQuAD v1.1 Dataset
 - 100k crowd source Question Answer Pairs. A data point contains a question and a passage from wikipedia which contains the answer.
 - Mest på Engelska, finns (ännu) inte så stora på Svenska
- Deep Pavlov
 - Moskvagrupp som använder BERT (multilingual)
 - PyTorch

Pavlov

- Lab
 - Tyvärr fungerar pavlovQA (nog) inte i Colab
 - Behöver äldre varianter av saker
 - Fungerar på min dator hemma
 - Torrsim men fungerar ändå inte så bra
- <http://docs.deeppavlov.ai/en/0.2.0/intro/tutorials.html>
- Exempelfil: 180704_planbeskrivning
- Några frågor och svar: pavlov_result_28sept

Finns många varianter på BERT

- Olika språk/ämnesdomäner
 - Anpassningar till olika språk, ämnesdomäner mm
 - Speciella, multilingual
- Olika storlekar
 - 90% av prestanda till 10% av HW
 - Ny Mobilversion förra veckan av Tensorflow Lite
- <https://camembert-model.fr/publication/camembert/>

Generering av text

- Skapa en längre realistisk text från några enskilda ord eller meningar
- Text GPT-2
- <https://app.inferkit.com/demo>
- "Jonas jobbar med järnvägar på Trafikverket och vi tittar på hur Norrbottniabanan ska byggas vidare från Umeå och hur miljöhänsyn kan tas. Frank är min handledare."
 - Genererade :
<https://github.com/nutte2/BERT/blob/master/gpt2text.txt>

Andra språkmodeller

- XLM-R (RoBERTa, facebook)
 - Lovande och troligen bättre än BERT på frågor&svar
 - Examensarbeten på gång
- GPT-3 (Open AI, fd Elon Musk)
 - Jättestor – online på förfrågan/betalning
 - GPT-2 finns tillgänglig
 - [Recension](#) på svenska
- [T-NLG](#) (Microsoft, azure)
 - Fler neuron än i hjärnan

Deployment

- Att se till att användarna får fram något att köra på
- Lokalt installerat
 - Nvidia-kort (t ex nya 3090), Python, massa bibliotek, PyCharm
 - Gui: PySide2 (gratisversion av PyQt5)
- Server
 - Python + Django
 - [REST](#) (REpresentational State Transfer)
 - Flask
 - Nvidia EULA – förbjudet att köra konsumentkort i server
 - Molntjänst; Azure, Amazon, Google, [Heroku](#), ..

Slutlaboration

- Trafikverket – min introfilm (preliminär version, 7 min)
- <https://youtu.be/QUUwKh8odWU>
- Plocka ihop valda delar av kodexemplen och kör på dina egna (text)filer!
- Frågor?
 - Kalle.prorok@gmail.com
- Tack!