DATASCI 207

Nedelina Teneva, PhD
nteneva@berkeley.edu

School of Information, UC Berkeley

Announcements

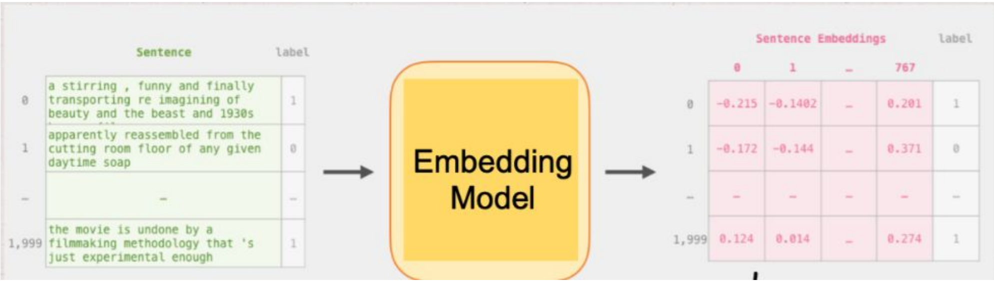
- **Course Evaluations:**
    - https://course-evaluations.berkeley.edu/
- Project presentations next week!
    - Details on the class website
    - **Reminders**: Put slides in the github repo; Remember to include a description of the items each member contributed to.

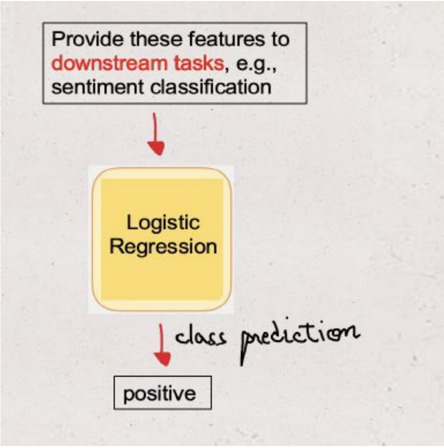
Today's topics

- Advanced topics: RNN, Transformers, BERT
- Applications on drug review classification

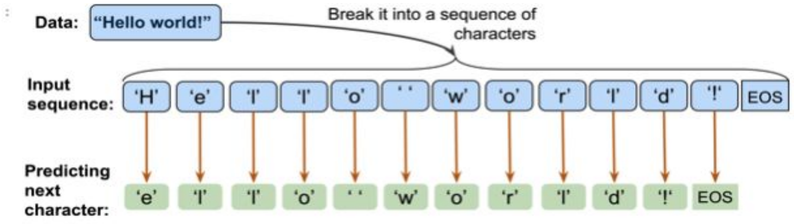
Sequential Data

- Sequential data (starting with embedding in hw9)
  - Context independent (FNN, CNN in week 10 demo)
  - Context dependent (so far we saw RNN, LSTM in week 11)





Recall Week 12: RNN/LSTM

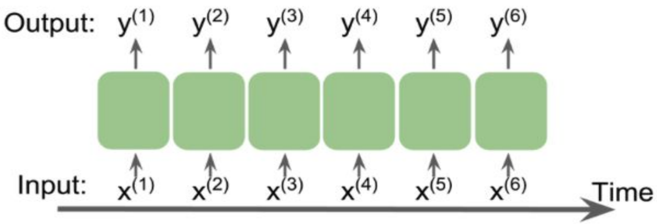


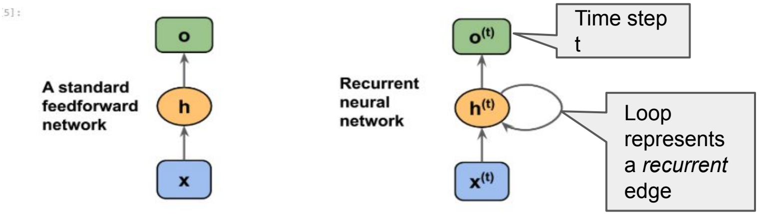
Can be modified to predict the next token (word)

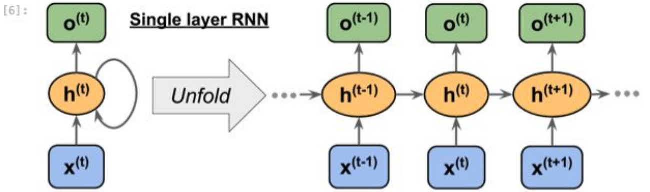
Demo (RM Chap 16): [Ch16_part1.ipynb]()

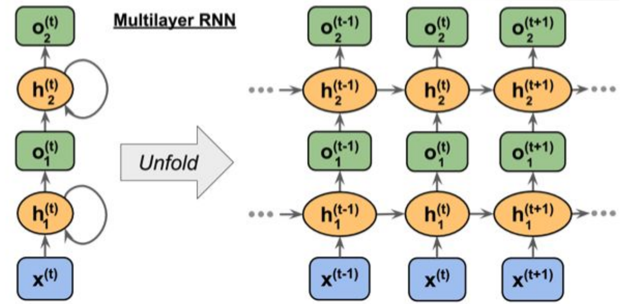
Representing sequences

```
Image(filename='images/16_01.png', width=700)
```





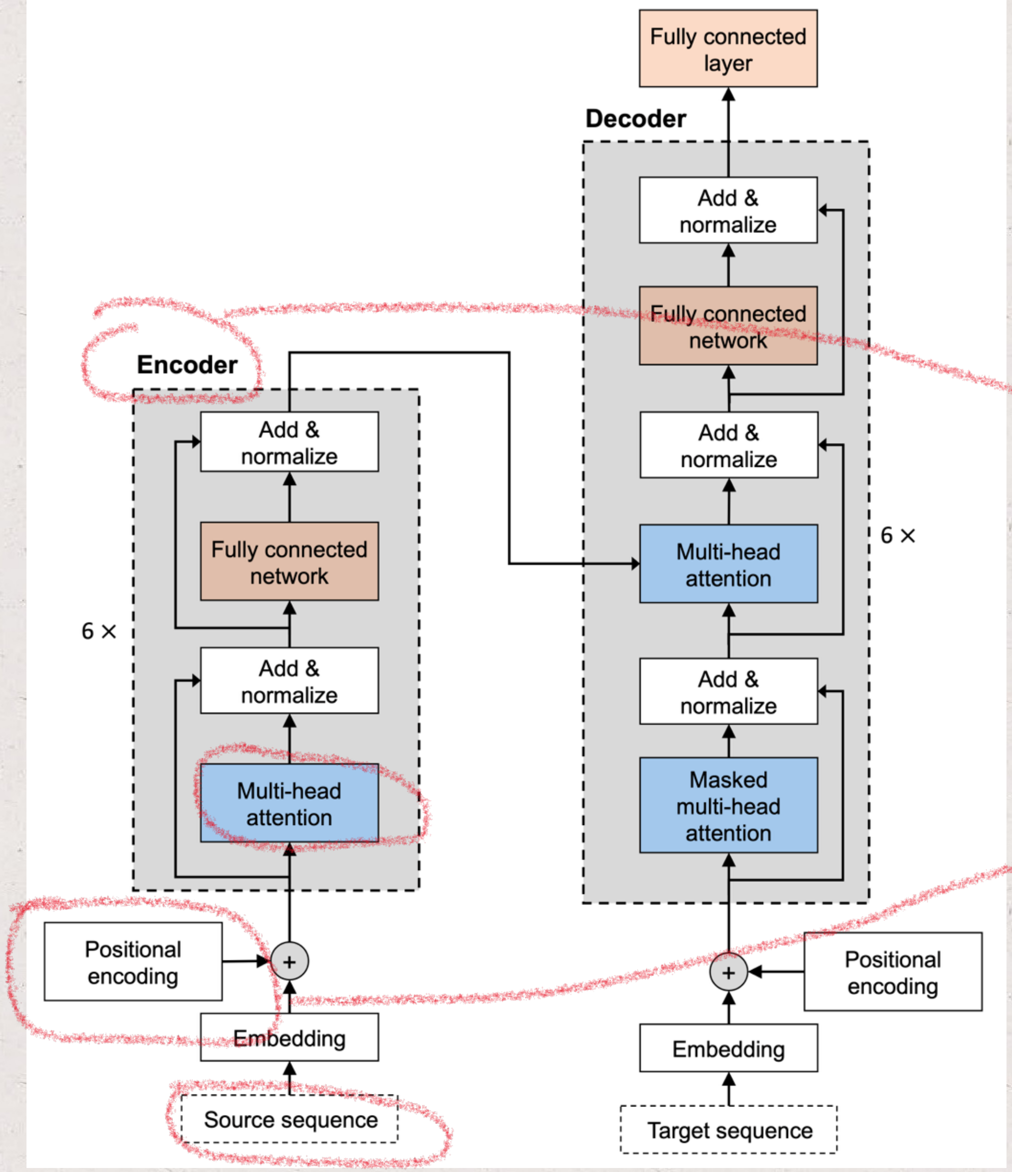




Transformer

- RNNs and LSTMs are powerful but they are computationally expensive and cannot parallelize well
- So since 2017 one model has ruled all of ML (including text, image, audio, tabular data…)
    - Transformer…(Attention is All you Need)
        - More powerful
        - Easy to parallelize
        - **Do not rely on recurrent layers, instead uses multi-head attention**
        - Impressive applications
        - But also require even more data

Attention is All you Need (NeurIPS 2017)



https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

impressive applications

even more powerful

deletes recurrent layers

Transformers

Can we do better ?

easy to parallelize

even more data hungry than other DL architectures

context aware (self-attention)

Learns a context-aware embedding vector (due to trainable self-attention weights)

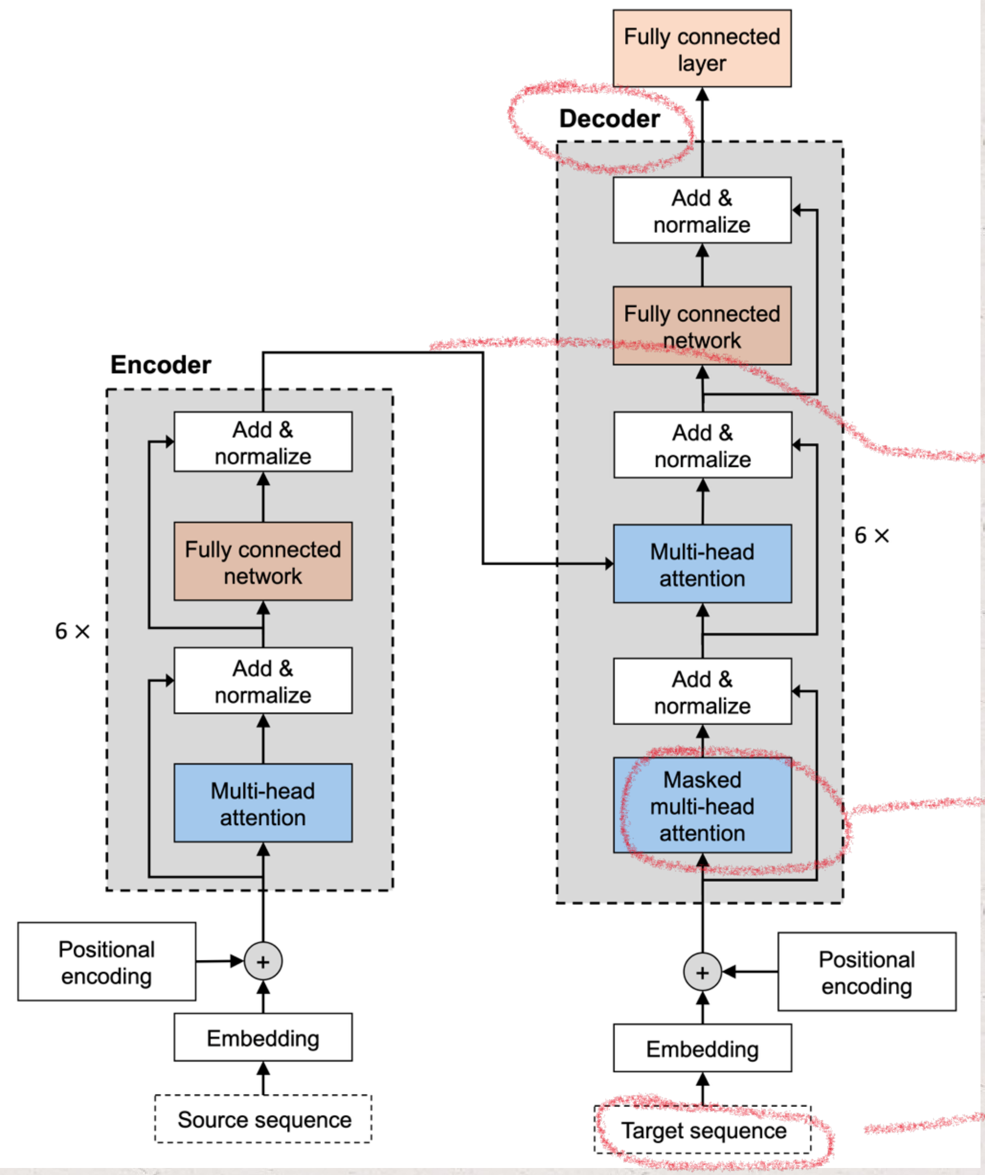
Captures information about the input sequence ordering (remember the architecture is not recurrent)

Mary gives John a flower

John gives Mary a flower

order is important

Attention is All you Need (NeurIPS 2017)



https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

impressive applications

even more powerful

deletes recurrent layers

Transformers

Can we do better ?

easy to parallelize

even more data hungry than other DL architectures

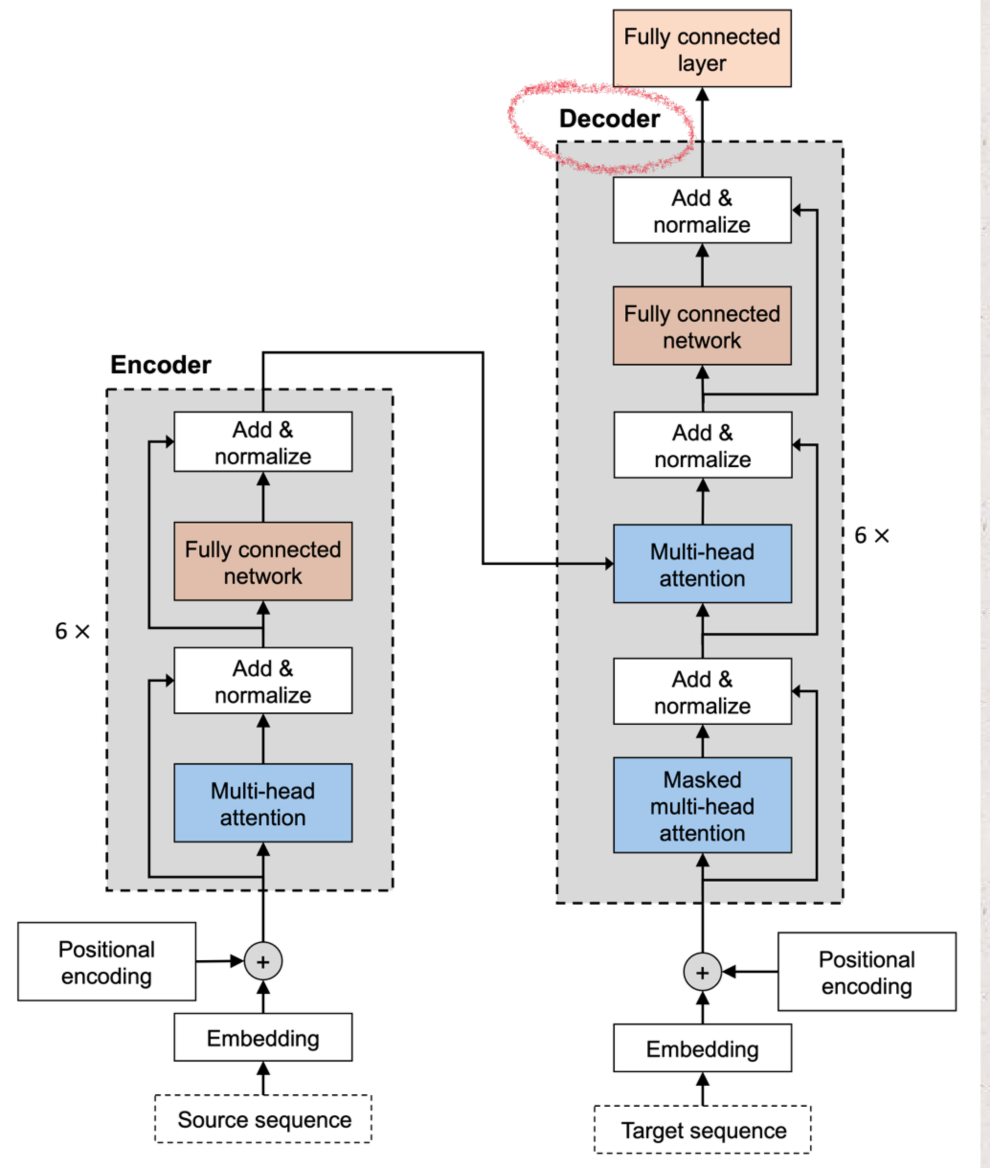
context aware (self-attention)

Receives encoded inputs from the Encoder block

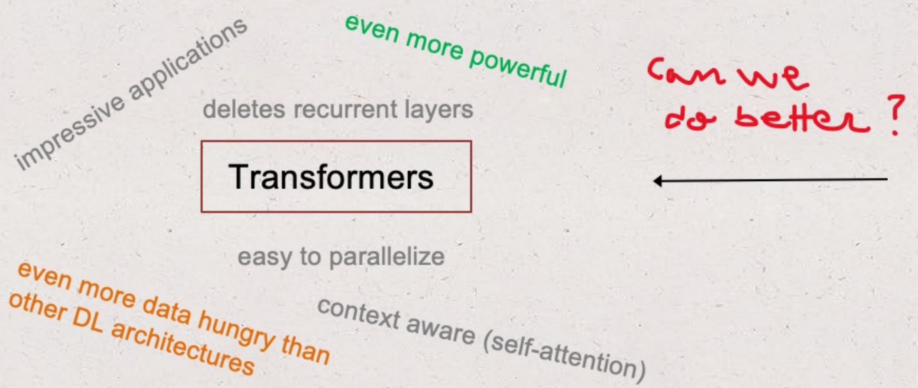
Masks certain number of tokens

Focuses on the output sequence

Attention is All you Need (NeurIPS 2017)



https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html



Can we build large-scale language models by leveraging unlabeled data and the transformer architecture?

↓

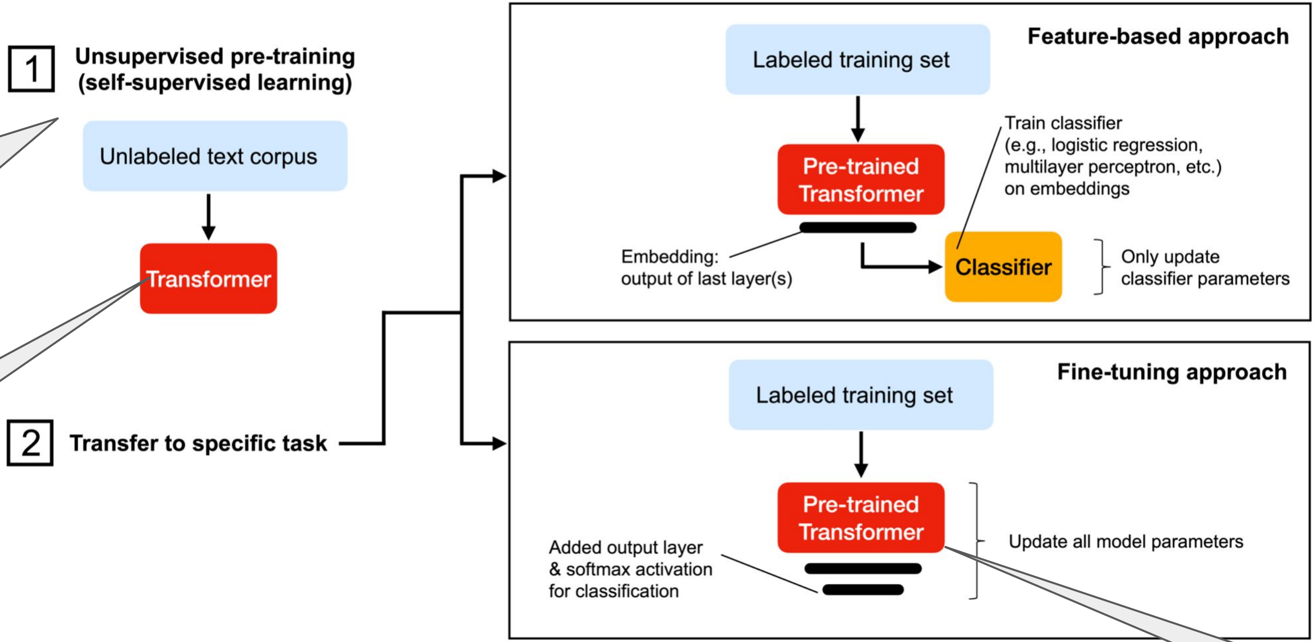
YES! (e.g., BERT)

↓

Pre-train on large corpuses of data (e.g., Wikipedia) and then fine-tune!

Transfer Learning - reuse knowledge from one model to another

Pretraining using e.g., ***BERT****: Pre-training of Deep Bidirectional Transformers for Language Understanding* *https://arxiv.org/pdf/1810.04805.pdf*



No need to train your own. Many models are open sourced.

"freeze" the embedding (BERT) layers

https://github.com/rasbt/machine-learning-book/blob/main/ch16/figures/16_10.png

Application: Sentiment Classification

Application using the drug review dataset (download link)

https://colab.research.google.com/drive/1Fc9R2cVnenRat7DZvGmIPkCaCxOQ2W1Q#scrollTo=beautiful-attendance