CS419M

Lecture 21: Deep Learning & Transformers

10 February 2023

Lecturer: Abir De *Scribe: Shubhhi*

1 Deep Learning in Natural Language Processing

In the context of NLP, deep learning algorithms are used to teach machines to understand language by analyzing large amounts of text data. This enables machines to identify patterns and relationships within the data, which can then be used to make predictions or generate new language.
S = We are students of IIT Bombay
Then for modelling this statement we find the relationship between the words.
P(S) = P('we') * P('are' | 'we') ..... * P('Bombay' | ...)
To create word relations, a machine learning algorithm is trained on a large set of text data. During training, the algorithm learns to predict the context in which each word appears, based on the words that appear around it. The resulting word vectors capture the meaning of each word in a high-dimensional space, based on its relationship to other words in the text.

2 Vectorization

Vectorization is jargon for a classic approach of converting input data from its raw format (i.e. text) into vectors of real numbers which is the format that ML models support. This is used in NLP.

$$W_{t+1} = F(W_t, h_t)$$

W_i are the word in vectorized form and h_t summarizes the previous state.

We have,

$$h_{t+1} = F(W_t, h_t) \quad W_{t+1} = G(h_{t+1})$$

We can attach positional encoding to each word

0	1	2	3	4	5	6	
W_1	W_2	W_3	W_4	W_5	W_6	W_7	...

This allows us to show sequential dependence using positional encoding. Word Embeddings or Word vectorization is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which used to find word predictions, word similarities/semantics.

$$h_{w_i} = \sum^{j} A_\theta(w_i, w_j) V_\phi(w_j) \tag{1}$$

h_{w_i} is contextual vector
(w_i, w_j) is the dependence
V_ϕ is value of each word

3 Scope

All this is indeed the backbone of Google search and ChatGPT. Relevant discussion was held during lecture regarding the various limitations of a google search.