A BERT \cite{Bert\_2018} model maga egy olyan nyelvi reprezentációra használt model, amely előre tanított úgynevezett "pretrained" model, ami azt jelenti, hogy már nem kell tanuló adatokkal betanítani, hogy a lehető legjobb eredményt kapjuk vissza az azután neki badott bemenetekre, hanem ezt elvégezték már, ezután már csak a kész modelt kell letölteni, hogy használni lehessen utána tetszőleges feladathoz. Az előre betanított modelek sokszor már működtek a múltban is, ezekre a \cite{Bert\_2018} cikkben említenek példákat. Az előtte lévő népszerű modelekhez képest annyi eddig kevésbé használt módszert alkalmaztak, hogy kétirányú ("bidirectional"), vagyis egy adott egységet melyekre éppen felosztódik a bemenet, legyen ez akár szó, szórész (mint a legelterjedtebb BERT model esetén), vagy akár karakter, úgy reprezentál, hogy nem csak az előtte következő bemenetből, hanem az utána lévő bemenetből is meríti a kontextust, melybe az adott egység tartozik. A vizsgált modelek szórészekre bontást alkalmaztak, mely azt jelenti, hogy a tokenizáció során az előforduló szórészek közül azokat használja fel, melyeknek a legnagyobb a "likelihood"-ja/valószínűsége, és ezek segítségével osztja részekre a szöveget, az egyes szavakat, vagy akár egyes szórészeket, ha nincs megfelelő elem a gyakran használtak között. Ennek elvégzését a Transformers alapok segítik. A Transformers modeleket csak néhány éve használják a mély tanulásban. Lényegük, hogy az RNN és LSTM mélytanuló modelektől eltérően, nem egyesével kell neki a bemenetén beadni az adatokat, hanem lehet egyszerre az egész szöveget betáplálni, a párhuzamosítást pedig megoldja a model magától, és visszadja a kódolt, részekre osztott inputot a kimenetén. Minden egyes egység a szövegben ka egy úgynevezett "embedding"-et is, ami lényegében a kontextust adja hozzá. Tehát, ha egy szó más-más szövegkörnyezetben is megjelenik egyazon bemeneten, az "embeddig"-je különböző lesz mindegyik esetben, mert más a kontextus amiben szerepel az adott szó. Az, hogy ezeket a kontextusokat a konkrét model hogyan vegye figyelembe, figyelem vagy angolul "attention" vektorokkal éri el, mely a modellel tudatja, hogy az adott egység a szövegben mely más egységekkel van kapcsolatban, tehát ennek kódolása során mely más részeit érdemes figyelembe venni a szövegnek. Ennek köszönhetően a bemenet jelentős részét beadhatjuk egyszerre a modelnek, hiszen a később megjelenő szavakat a figyelem vektor már nem veszi be, minden szó/szórész kódolásánál vagy épp dekódolásánál csak annyi egyéb szavat/szórészt vesz figyelembe, amennyi az éppen kódolt szónál/szórésznél lényeges, lehet párhuzamosan is kódolni szavakat/szórészeket melyek nem kapcsolódnak.

A Transformers \cite{2019huggingface} egy Google által kidolgozott technológia, amelyet sok más mély tanuló model alkalmazott sikeresen a közelmúltban, beleértve a BERT szöveg reprezentáló modelt. A Transformers legfontosabb újítása az volt, hogy az azelőtti megoldásokban használt RNN és LSTM hálókat teljes mértékben lecserélte az attention vagy magyarul „figyelem” mechanizmusra, és így is megfelelő teljesítményt ért el. A model egy kódoló-dekódoló modelként épül fel, mind a kódoló, mind a dekódoló felében a modelnek az attention mechanizmus és egy fully connected Feed Forward network segítségével történik meg a bemenet átalakítása, de alkalmaznak úgynevezett residual kapcsolatokat is, amelyek egy-egy részét a hálónak „kikerülik” ezzel a benne rejlő információkat nem károsítva.

kép

Az attention mechanizmus lényege az, hogy az egyes részeihez a bemenetnek vektorokat rendel, amelyeknek értékei attól függ, hogy az adott szövegrész mely más szövegben szereplő elemekkel van kapcsolatban, pl. egy mutatónévmás és a szó, amire utal, akkor az egyik a másik attention vektorában magasabb értékeket kap. Ezt multi head attention módszerrel teszi, aminek az a lényege, hogy nem egyszer végzi el a mechanizmust, hanem többször párhuzamosan és végül ezek eredményét konkatenálja, hogy végül a megfelelő attention-ök legyenek hozzárendelve minden egyes szöveg részlethez, és így továbbítódjanak a model következő komponense felé. Végül a Transformers kódolója és dekódolója esetén is a fully connected Feed Forward network fogja a végső jelentősebb átalakításokat (még van egy softmax és egy linear layer utána dekódolás esetén) elvégezni a bemenet adott részein, mely egy egyszerű néhány rejtett réteggel rendelkező háló, más-más paraméterekkel és dimenziókkal rétegenként.

kép

A BERT vagyis Bidirectional Encoder Representations from Transformers a nevéből is adódóan az előbb bemutatott Transformers model felhasználásával végzi el a szöveg reprezentáló feladatát. Ennek következtében a BERT is kódoló-dekódoló modelt használ, attention mechanizmussal, viszont mindkét irányban elvégzi ezt, a bal és jobboldali kontextusra is a bemeneti szekvencián. A bemenetet egy tokenizáló segítségével szórészekre (WordPiece embedding) bontja, oly módon, hogy a szótárában szereplő szórészek közül azokra bontja szét a bemeneti szekvenciában szereplő mondatokat, amelyek a legvalószínűbben fordulnak elő a BERT-nek betanított szekvenciákban. A tokenizáló a bemenete elé és mögé különleges tokeneket helyet el, ez a CLS és a SEP tokenek. A tokenizált bemenetből a BERT kinyeri a kontextuális információkat a rejtett rétegei segítségével, melyek száma 12. A dolgozatban vizsgált BERT modelek már előre betanítottak, tehát a rejtett rétegeinek a súlyai már bizonyos tanuló adatok alapján elő lettek készítve ismeretlen inputok számára.