Mély nyelvmodellek, soknyelvű modellek

Számítógépes nyelvészet

2022.04.25.

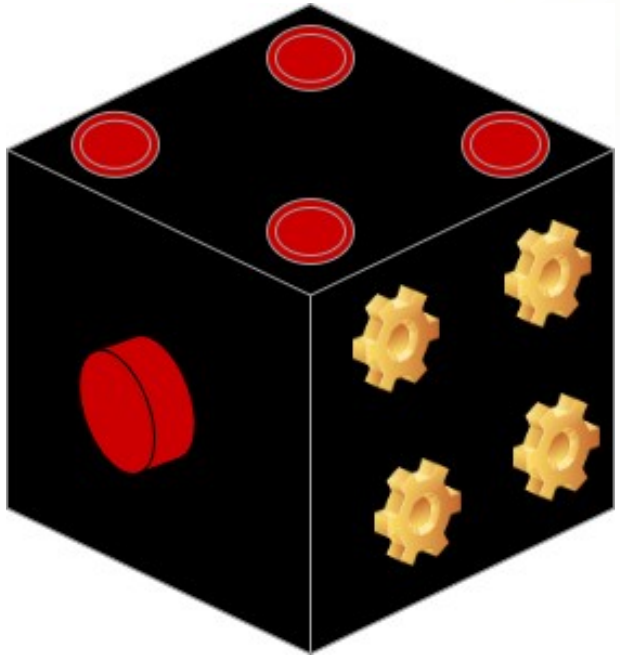
Nyéki Bence
Nyelvtudományi Kutatóközpont

Felügyelt tanulás: Mi a feladat?

- Van egy adathalmazunk: $\{\langle x_1, y_1\rangle, \langle x_2, y_2\rangle, \ldots, \langle x_n, y_n\rangle\}$.
- Olyan szabályokat (matematikai műveletek sorozatát) keresünk, amelyek megadnak egy f függvényt, amelyre $f(x_i) \approx y_i$, $i \in \{1,\ldots,n\}$.
- A mély neurális hálók univerzális approximátorok, azaz képesek ilyen szabályokat találni.
- Azt szeretnénk, hogy f olyan x adatpontra is „helyes” értéket adjon, amely nem szerepelt az adathalmazban.
- Ezt nehéz garantálni, de a gyakorlatban működik...

Mi a neurális háló?

„Fekete doboz gombokkal és karokkal”: ha más a gombok és karok helyzete, máshogy működik a neurális háló



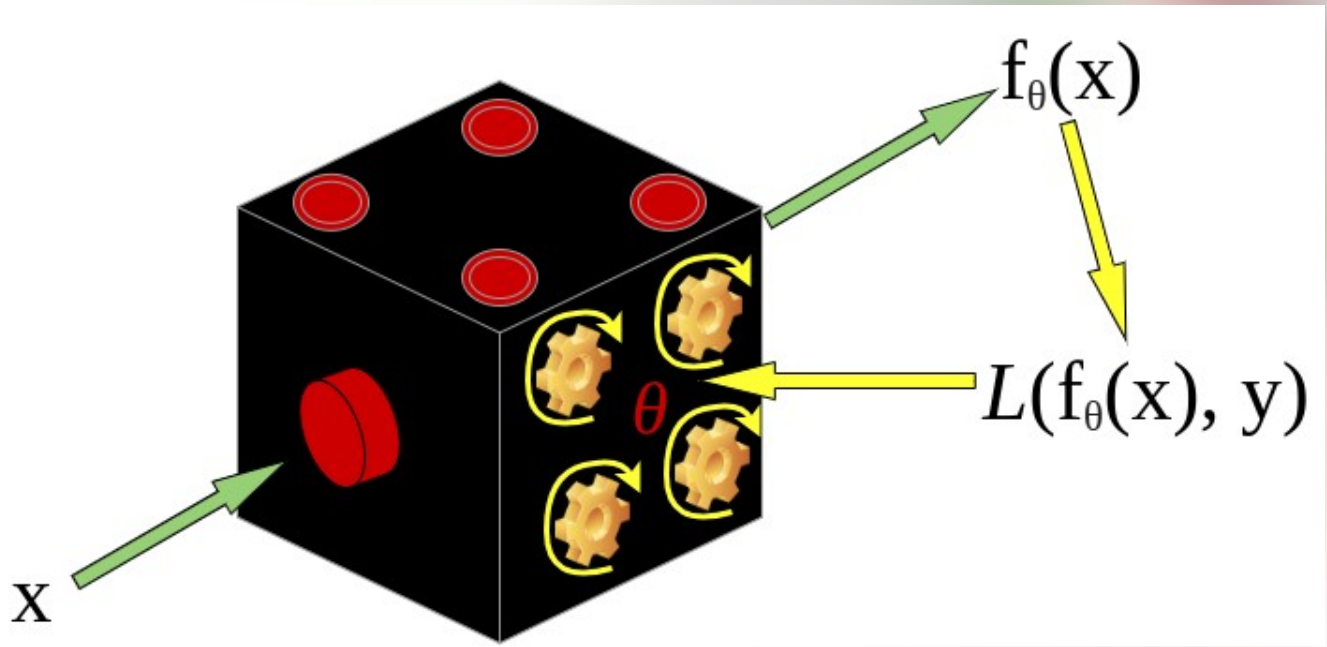
Hogyan zajlik a tanítás?

Van egy x bemenet és egy $f_\theta(x)$ kimenet, f_θ-t a neurális háló θ paraméterei (a gombok és karok) határozzák meg.

Ismerjük a helyes választ (elvárt kimenetet) is, ez y.

Az L függvény megmutatja, mennyire különbözik $f_\theta(x)$ y-tól.

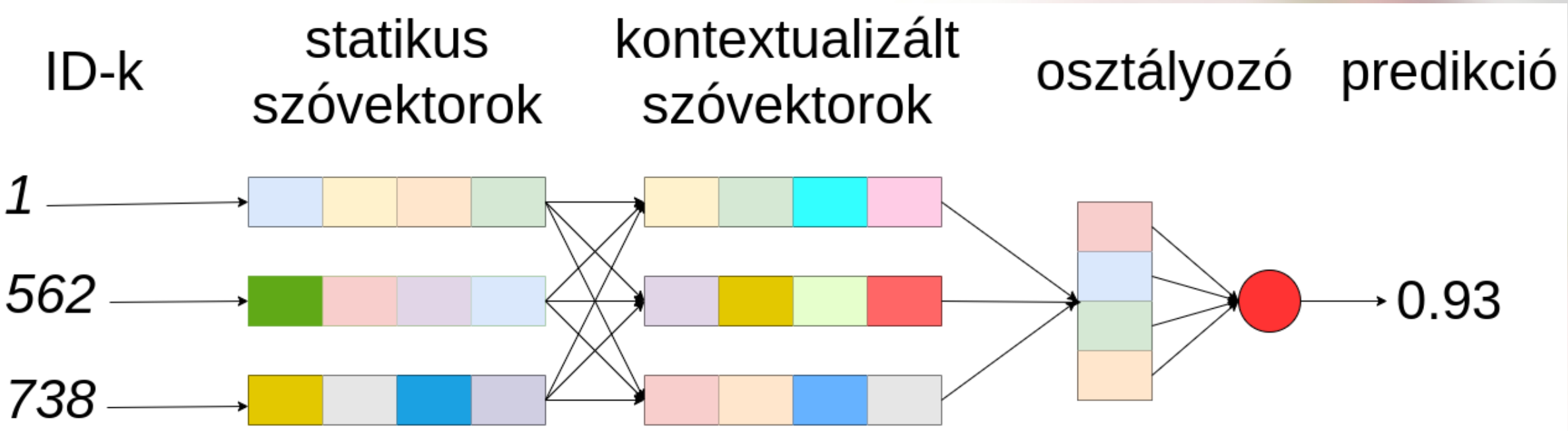
A különbség (hiba) alapján frissítjük θ-t.



Mi köze ennek a természetes nyelvekhez?

- A szótár szavait sorba rendezhetjük és megszámozhatjuk.
- pl. {*a*: 1, …, *kutya*: 562, …, *ugat*: 738, …}
- Ekkor az „*a kutya ugat*” mondat felírható így: [1, 562, 738].
- A mondatokhoz különböző címkéket rendelhetünk, pl. jólformált mondat (1), nem jólformált mondat (0).
- Már van is egy adatpontunk! <[1, 562, 738], 1>
- Ha sok adatponton tanítunk be egy neurális hálót, megtanulhatja a címkék és a bemenetek közötti összefüggést, pl. azt, hogy melyik mondat jólformált és melyik nem.

A modern modellek motorházteteje alatt



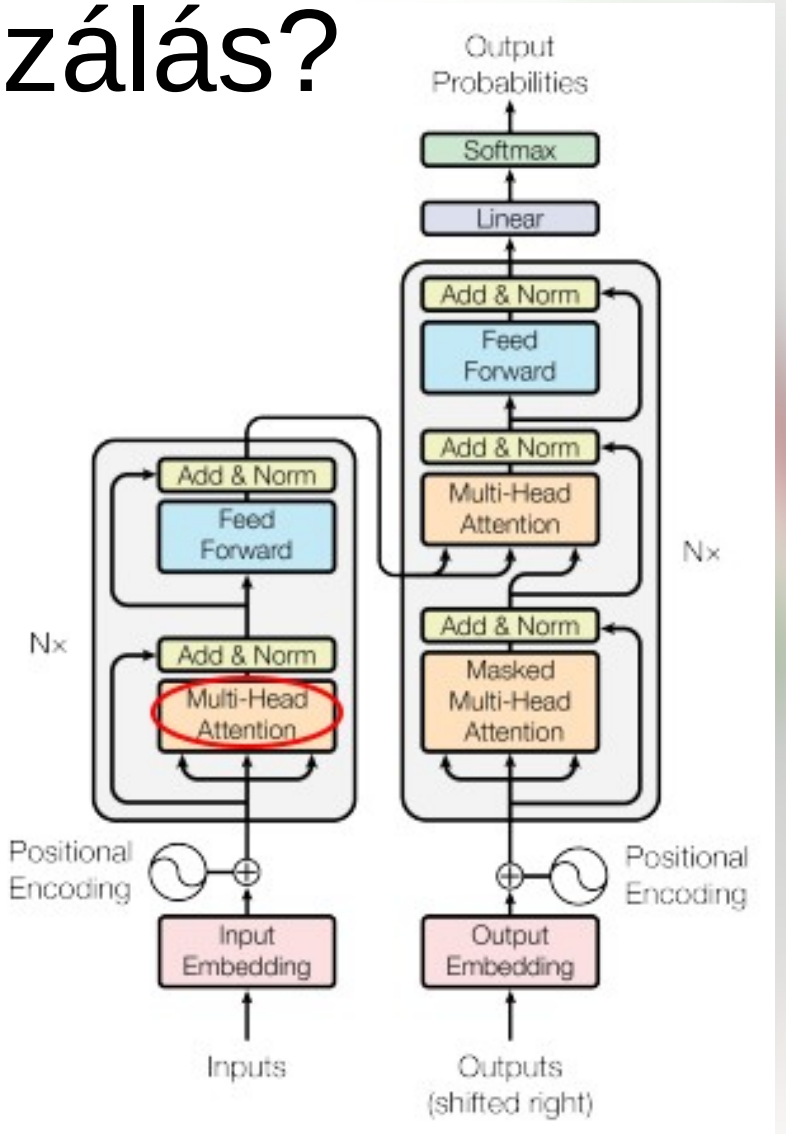
Miért van szükség kontextualizált beágyazásokra?

- „*Ő nekem igazi jó barát.*” vs „*Minden jó barát imádkozik.*”
- Poliszém szavak konkrét jelentésétől függhet pl. a mondat helyes címkéje.
- A kontextus (többnyire) feloldja a poliszémiát.
- A modellek nem különbözetnek meg diszkrét jelentéseket: ahány kontextusban fordul elő egy szó, annyi különböző beágyazás tartozhat hozzá.

Hogyan zajlik a kontextualizálás?

A kulcs az ***attention*** (figyelmi mechanizmus).

Már rekurrens neurális hálókban is használták egyes változatait, de a Transformer architektúra részeként vált igazán elterjedté.



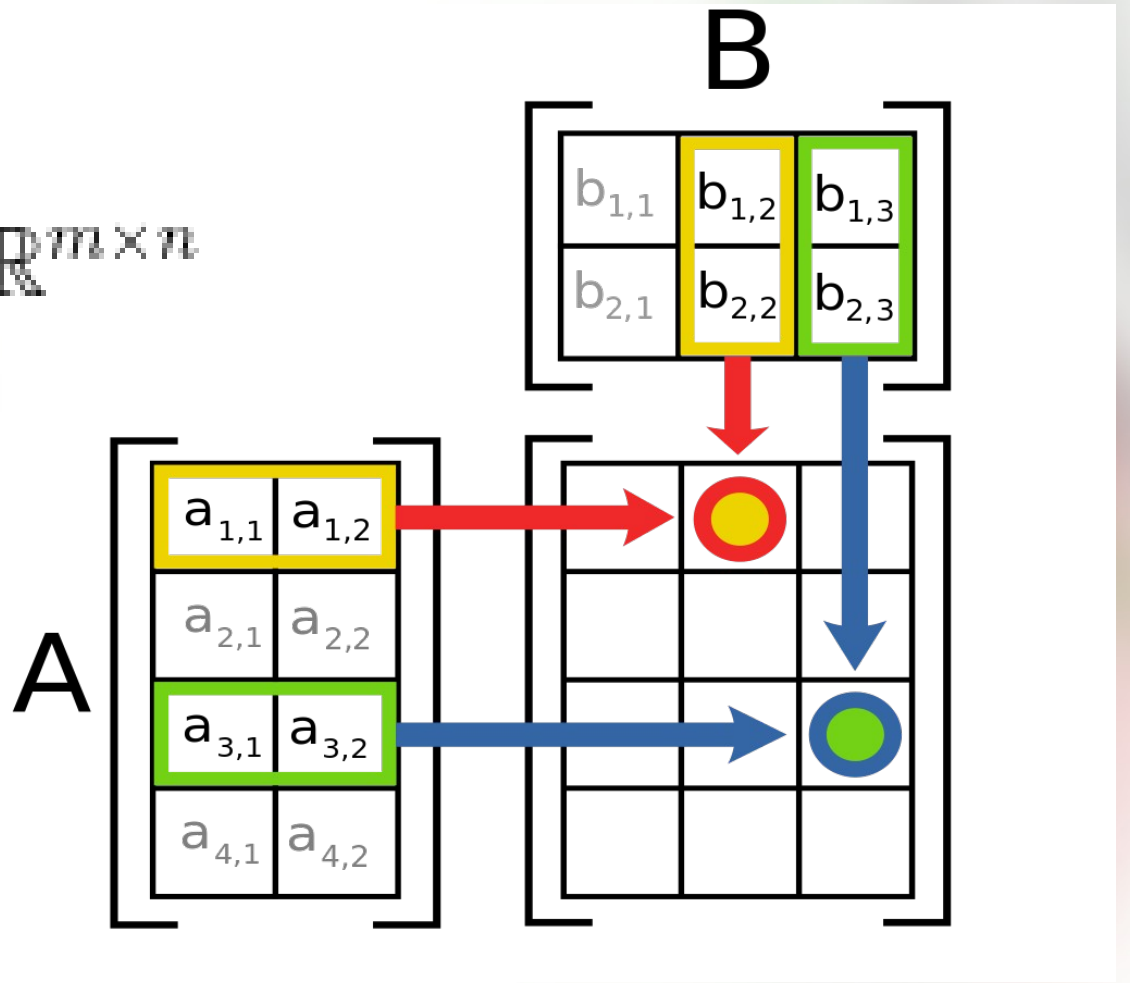
Az eredeti kép forrása: Vaswani et al. (2017)

Mátrixszorzás

$$\mathbf{AB} = \mathbf{C}$$

$$\mathbf{A} \in \mathbb{R}^{m \times l},\ \mathbf{B} \in \mathbb{R}^{l \times n},\ \mathbf{C} \in \mathbb{R}^{m \times n}$$

$$c_{ij} = \sum_{k=1}^{l} a_{ik} b_{kj}$$



A kép forrása: https://en.wikipedia.org/wiki/Matrix_multiplication

Softmax

$$\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_j}},\ i \in \{1, \ldots, n\},\ \mathbf{z} \in \mathbb{R}^n$$

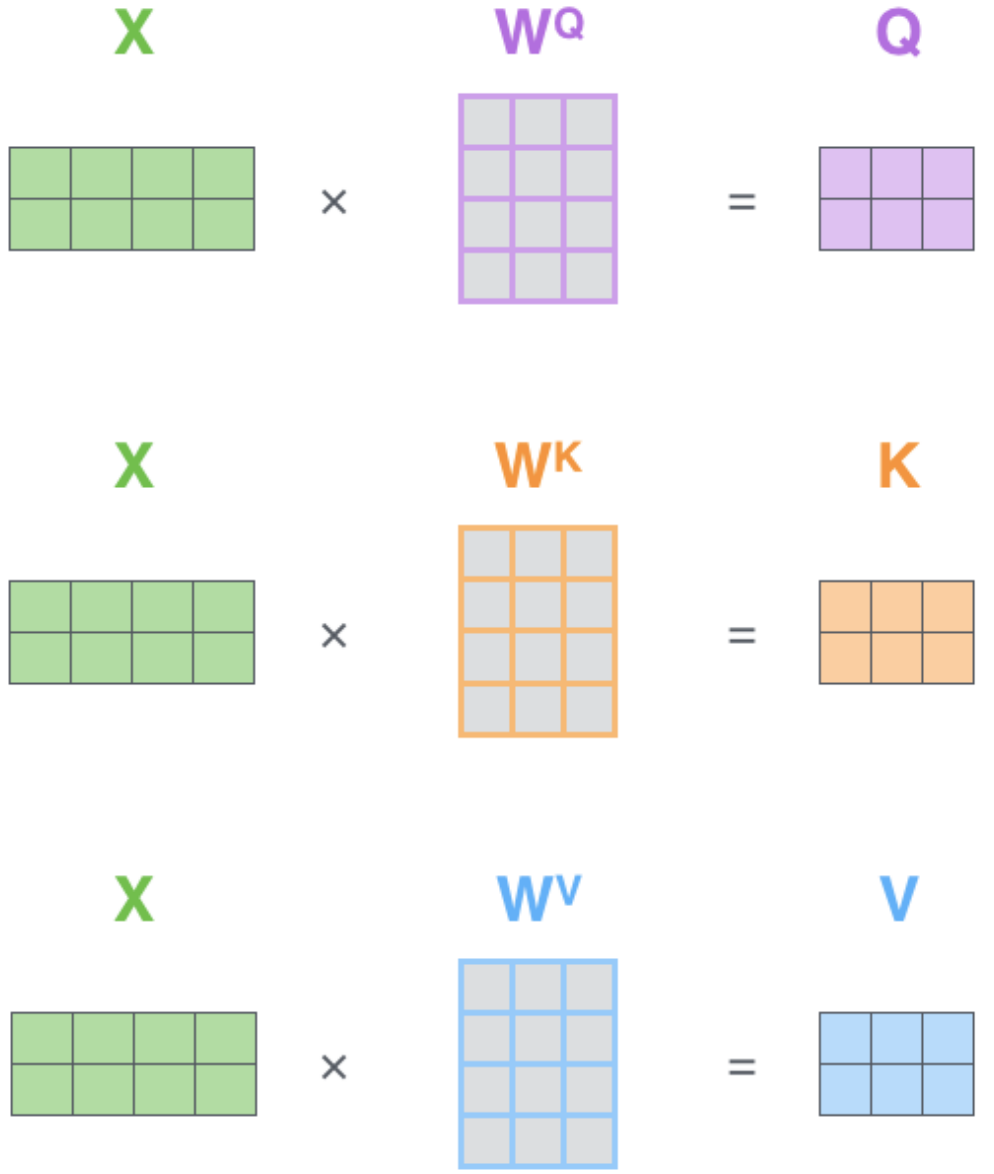
Példa:

$$\mathbf{z} = (0.5,\ 1.5,\ 1)$$

$$\sigma(\mathbf{z}) = (0.1863,\ 0.5065,\ 0.3072)$$

Q, K, V mátrixok

- X: statikus szóbeágyzások
- W^Q, W^K, W^V: súlymátrixok
- Q, K, V: *query*, *key*, *value* mátrixok
- Az X, Q, K és V mátrixoknak annyi soruk van, amennyi szóból áll a bemeneti szöveg.

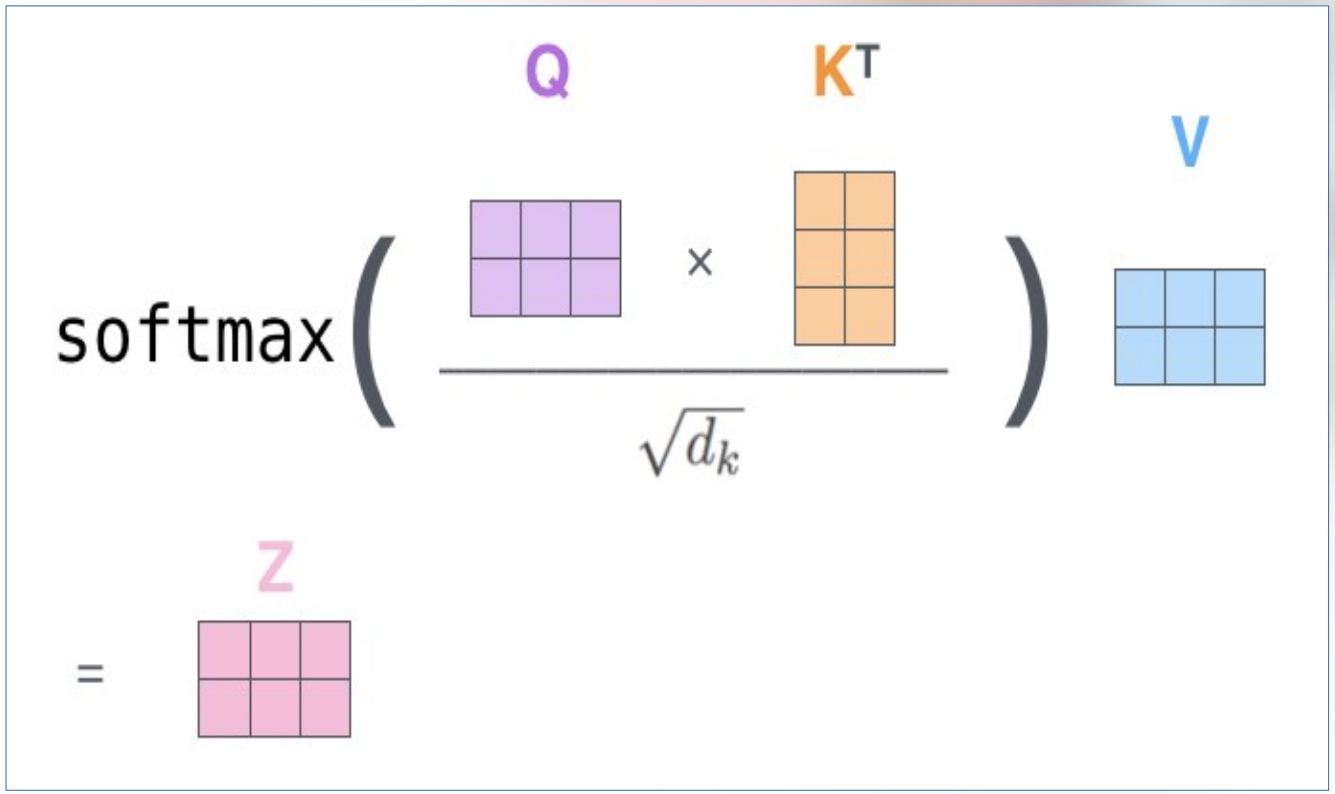


A kép forrása: Jay Alammar: The Illustrated Transformer

Attention

Keressük azokat a súlyokat, amelyek megmutatják, milyen mértékben vegyük figyelembe a környezetet az egyes szavak beágyazásainak módosításához.

- Q: szóreprezentációk a kereséshez, *amiket összehasonlítunk*.
- K: reprezentációk a Q-val való összehasonlításhoz, *amikkel összehasonlítunk*.
- V: azok az értékek, *amiket frissítünk*.



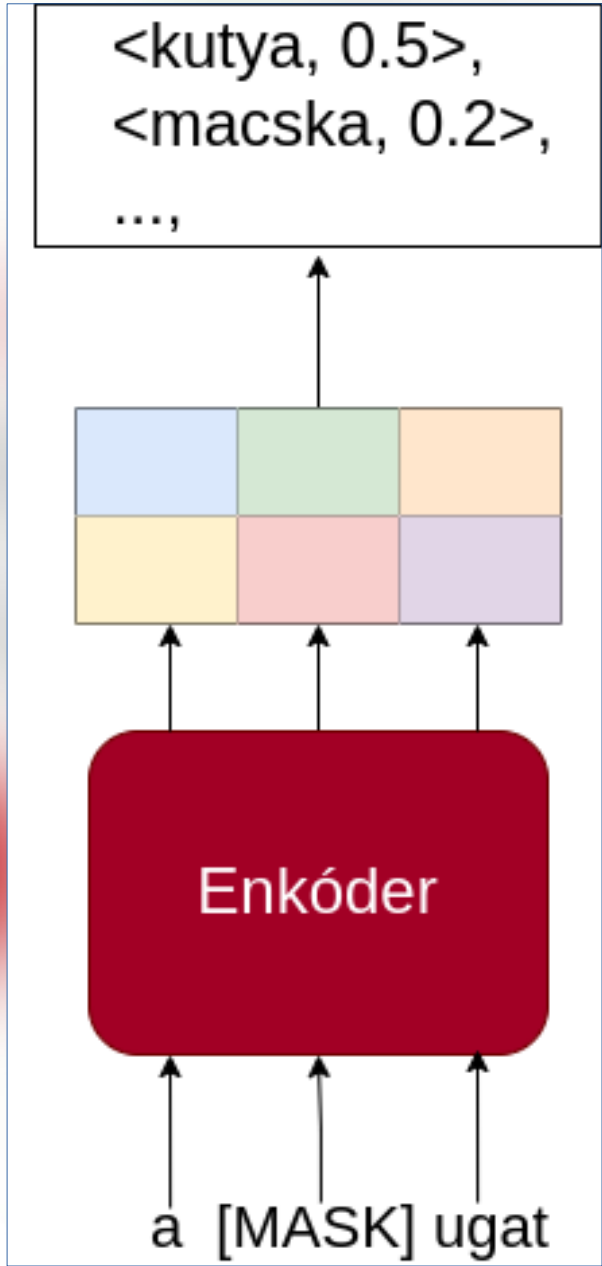
A kép forrása: Jay Alammar: The Illustrated Transformer

Transfer learning

- Az egyes feladatokhoz (pl. véleményelemzés) jellemzően kevés a tanítóanyag.
- Azonban nagy mennyiségű, annotálást nem igénylő adaton *előtaníthatjuk* a modelleket, hogy általános jellemezőket tanuljanak a nyelvről.
- *Finomhangoláskor* továbbtanul a modell, hogy egy konkrét feladatot meg tudjon oldani.

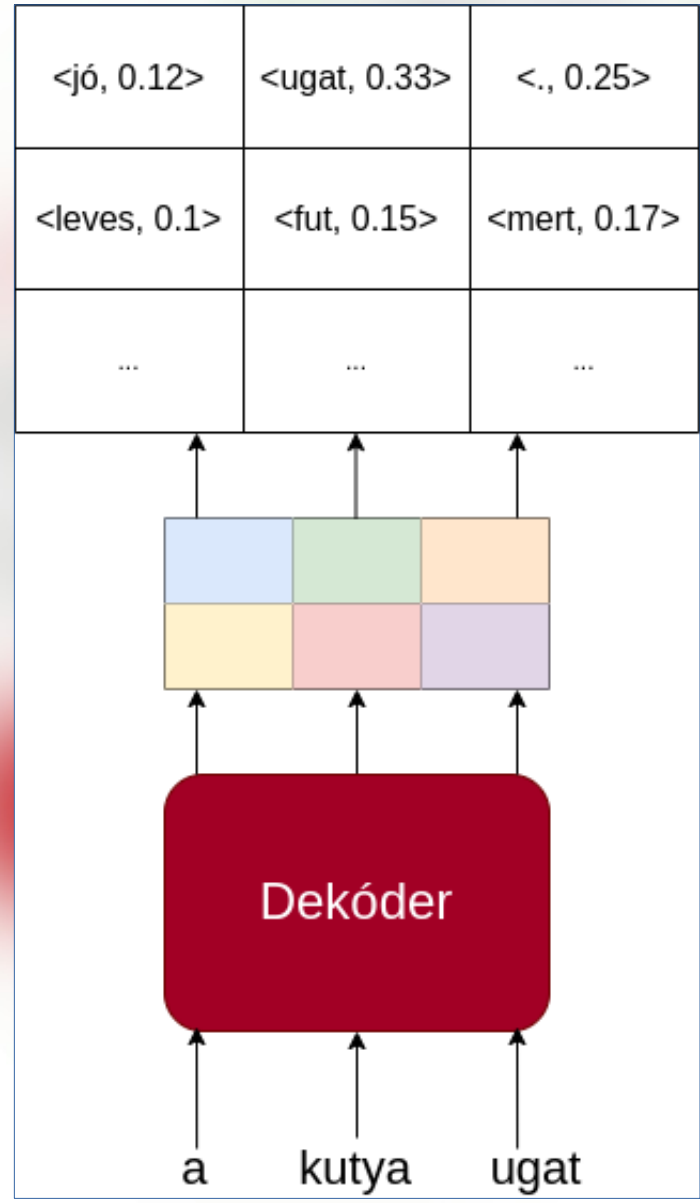
Enkóderek

- A cél reprezentációk előállítása.
- Az előtanítás általában maszkolt nyelvmodellezéssel történik: valószínűségi eloszlásokat várunk letakart szavakra.
- Nincs szükség kézi annotálásra, *self-supervised learning.*
- Ha a modell már jó kontextualizált szóbeágyazásokat állít elő, akkor egy klasszifikációs feladat megoldásához már csak egy osztályozó fejet (néhány új réteget) kell hozzáadni a modellhez.



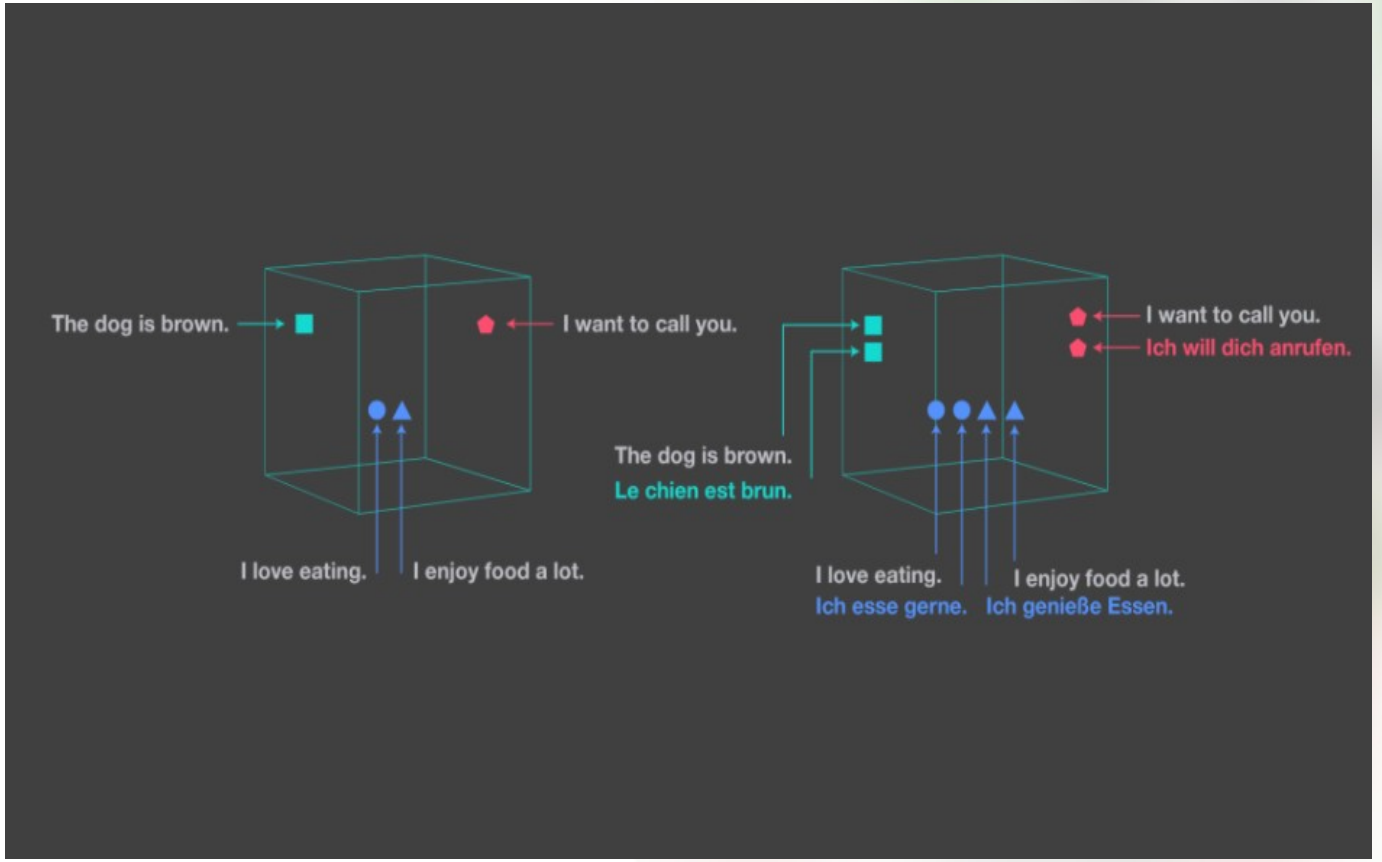
Dekóderek

- Eredetileg szöveggenerátorok, bemenetük valamilyen szövegreprezentáció.
- Ma gyakran enkóder nélkül használják őket, megkezdett szövegeket folytatnak.
- Előtanítás nyelvmodellezéssel: a következő token (szó) megjóslása.



Többnyelvű modellek

- Elv: hasonló jelentésű szavak/mondatok nyelvtől függetlenül legyenek közel egymáshoz a szemantikai tervben.
- Előtanítás egyetlen soknyelvű szótárral soknyelvű (nem feltétlenül párhuzamos) szöveganyagon.
- Ha egy nyelven finomhangolunk, az érinti a többi nyelv feldolgozását is.

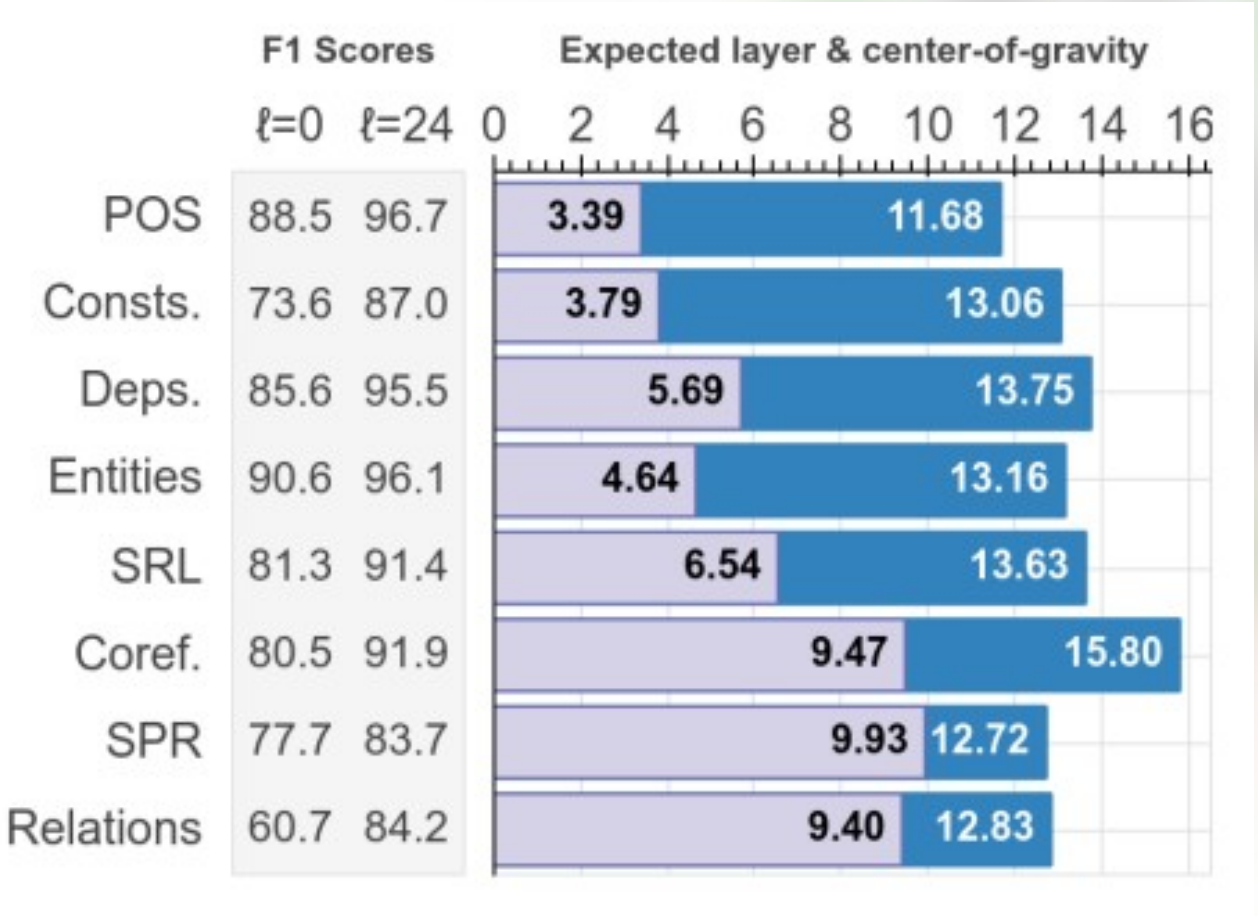


A kép forrása: Holger Schwenk:
Zero-shot transfer across 93 languages: Open-sourcing enhanced LASER library.

Feldolgozólánc a BERT-ben

A BERT Transformer-alapú enkóder-modell rétegei tükrözik az klasszikus szövegelemző-lánc moduljait.

A diagramon az látható, várhatólag melyik réteg lesz képes először megoldani egy feladatot (lila), illetve várhatólag melyik réteg tartalmazza a legtöbb hasznos információt a feladat megoldásához (kék).



A kép forrása: Tenney et al. (2019):
BERT Rediscovers the Classical NLP Pipeline

Cikkek:

Mikel Artetxe, Holger Schwenk (2019): *Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond*. Link

Ian Tenney, Dipanjan Das, Ellie Pavlick (2019): *BERT Rediscovers the Classical NLP Pipeline*. Link

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin (2017): *Attention is all you need*. Link

Egyéb linkek:

Jay Alammar: *The Illustrated Transformer*. Link

Holger Schwenk: *Zero-shot transfer across 93 languages: Open-sourcing enhanced LASER library*. Link

Matrix multiplication (Wikipédia). Link