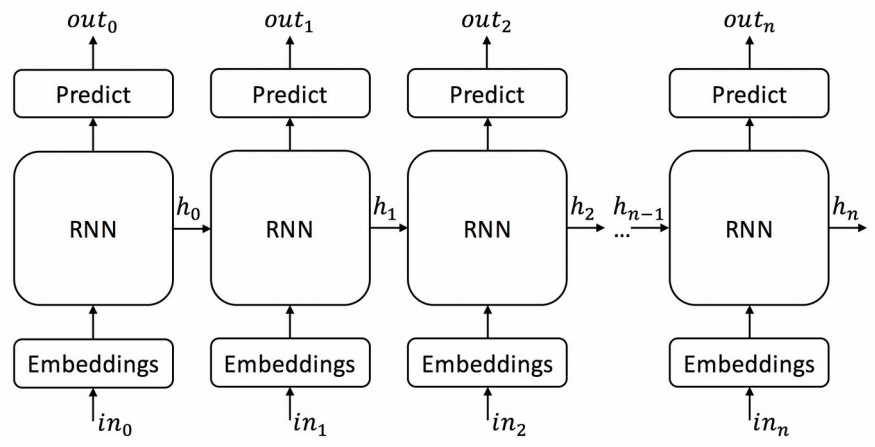
Kontextuális reprezentációk

Neurális hálók

- Neurális hálókkal bonyolult függvényeket írhatunk le
  - Cél: (input, output) párok alapján az inputokból az elvárt outputokba képező túlparaméterezett függvény tanulása
    - Mindahányszor egy tanítóinputra nem az elvárt outputot adja a modellünk válaszul, módosítsuk a modellünk paramétereit

Rekurrens neurális hálók (RNN)

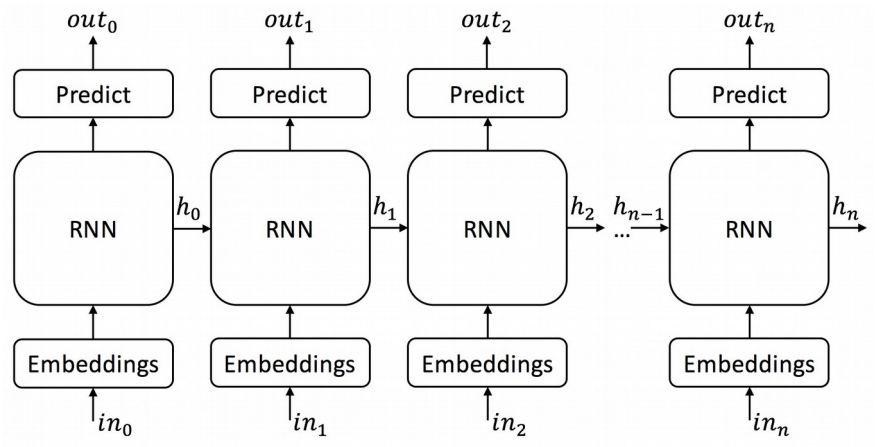
- Szekvenciális input esetében szeretnénk az előzményeket is beépíteni a modellbe
  - „Átvettem a **levelet** a _____” vö. „Fölvettem a **levelet** a _____”
- Egy sematikus RNN

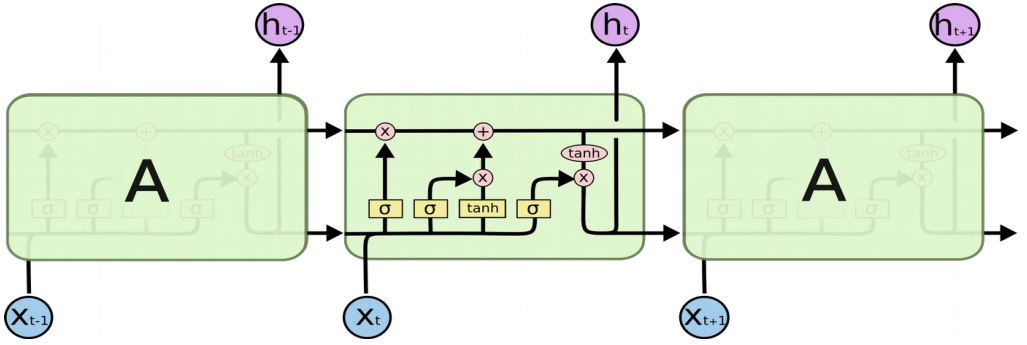


Ábrák forrása: Tal Baumel

Rekurrens neurális hálók (RNN)

- Szekvenciális input esetében szeretnénk az előzményeket is beépíteni a modellbe
  - „Átvettem a **levelet** a _____” vö. „Fölvettem a **levelet** a _____”
- Egy sematikus RNN (sok változata közül az LSTM az egyik legnépszerűbb)





Ábrák forrása: Tal Baumel, Chris Olah

Vanilla RNN

- Az adott pillanatra vonatkozó **h** rejtett állapot:

  h = np.tanh(W_hh @ **h** + W_xh @ **x**)

  - A korábbi **h** egy transzformáltjának (W_hh @ **h**) és az aktuális **x** input (szó)vektor egy transzformáltjának (W_xh @ **x**) összegzésével nyert vektoron elemenként vett nemlinearitás (tanh)

Vanilla RNN

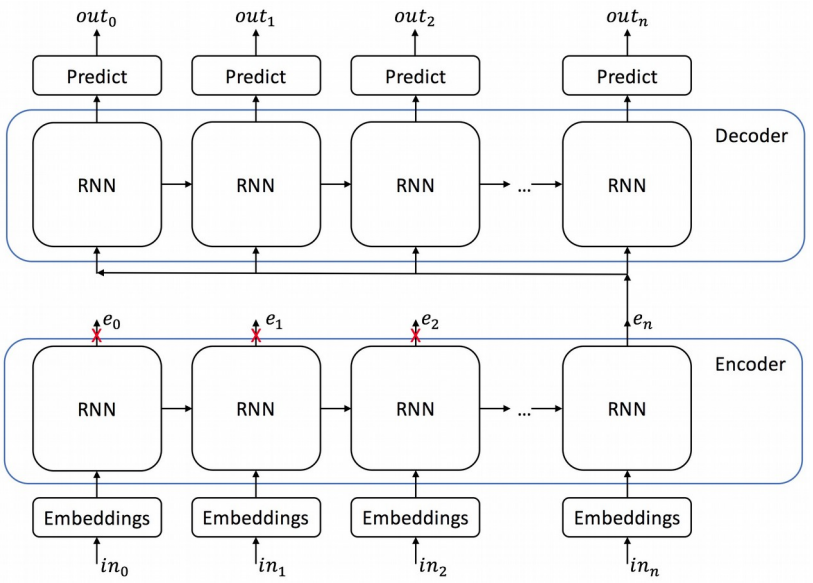
- Az adott pillanatra vonatkozó **h** rejtett állapot:

  h = np.tanh(W_hh @ **h** + W_xh @ **x**)

  - A korábbi **h** egy transzformáltjának (W_hh @ **h**) és az aktuális **x** input (szó)vektor egy transzformáltjának (W_xh @ **x**) összegzésével nyert vektoron elemenként vett nemlinearitás (tanh)
- **h** az adott pozícióig bezárólag egy aggregált rejtett reprezentációja az inputnak
  - Segítségével (és egy W_hy paramétermátrix bevezetésével) egy prediktív modellt hozhatunk létre (amit SGD-vel tanítani tudunk)

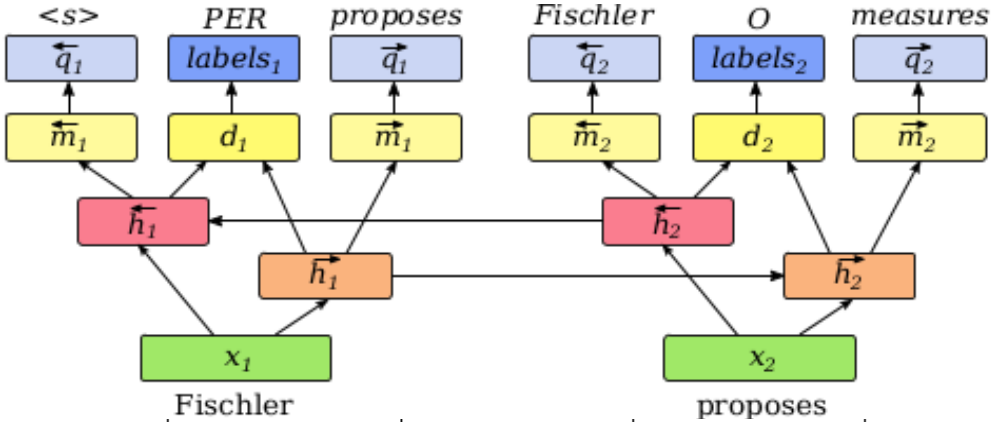
Enkóder–dekóder architektúrák

- Az RNN-ek jól működnek nyelvi modellezésre, seq2seq feladatokra azoban az enkóder–dekóder architektúra jobb
  - seq2seq probléma alatt valamilyen (lazán értelmezett) fordítási feladatot értünk (pl. gépi fordítás vagy beszélgető ágens)



Multi–task learning

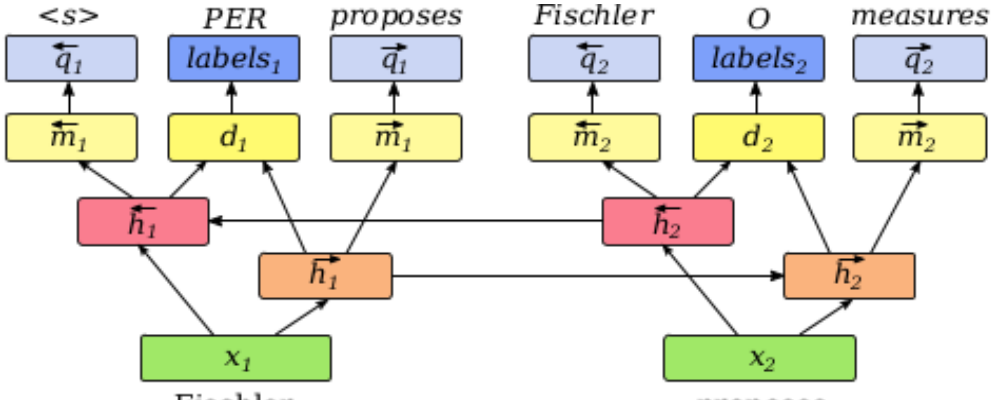
- Egy végső feladat megtanulásához hozzásegíthet bennünket kísérőfeladatok tanulása: A háló ne csak egy dologhoz értsen!
  - Semi-supervised Multitask Learning for Sequence Labeling (Rei, 2017)



	FCE		CoNLL-14		CoNLL-03		CHEMDNER		CoNLL-00		PTB-POS		UD-ES		UD-FI	
	DEV	TEST	TEST1	TEST2	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
Baseline	48.78	44.56	15.80	23.62	90.85	85.63	83.63	84.51	92.92	92.67	97.23	97.24	96.38	95.99	95.02	94.80
+ dropout	48.68	42.65	14.71	21.91	91.14	86.00	84.78	85.67	93.40	93.15	97.36	97.30	96.51	96.16	95.88	95.60

Multi–task learning

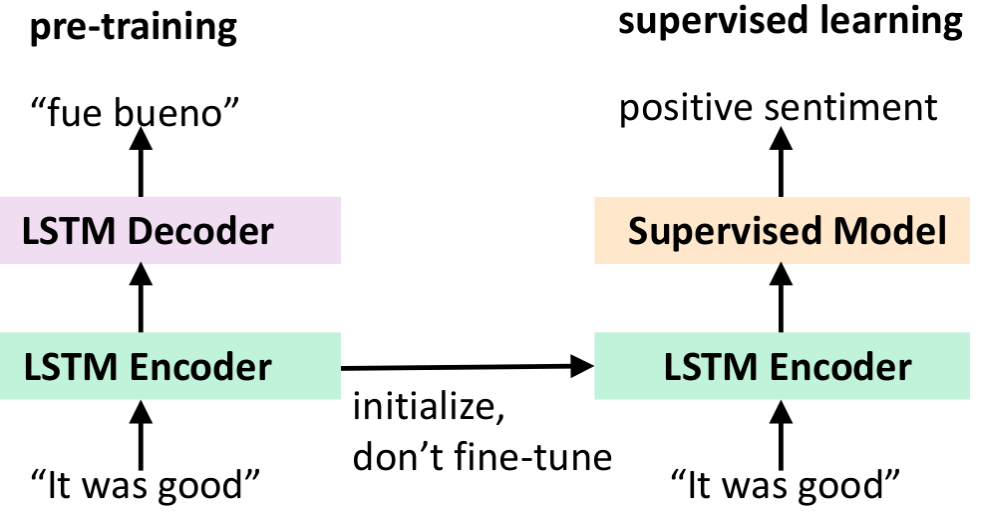
- Egy végső feladat megtanulásához hozzásegíthet bennünket kísérőfeladatok tanulása: A háló ne csak egy dologhoz értsen!
  - Semi-supervised Multitask Learning for Sequence Labeling (Rei, 2017)



	FCE		CoNLL-14		CoNLL-03		CHEMDNER		CoNLL-00		PTB-POS		UD-ES		UD-FI	
	DEV	TEST	TEST1	TEST2	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST	DEV	TEST
Baseline	48.78	44.56	15.80	23.62	90.85	85.63	83.63	84.51	92.92	92.67	97.23	97.24	96.38	95.99	95.02	94.80
+ dropout	48.68	42.65	14.71	21.91	91.14	86.00	84.78	85.67	93.40	93.15	97.36	97.30	96.51	96.16	95.88	95.60
+ LMcost	**53.17**	**48.48**	**17.86**	**25.88**	**91.48**	**86.26**	**85.45**	**86.27**	**94.22**	**93.88**	**97.48**	**97.43**	**96.62**	**96.21**	**96.14**	**95.88**

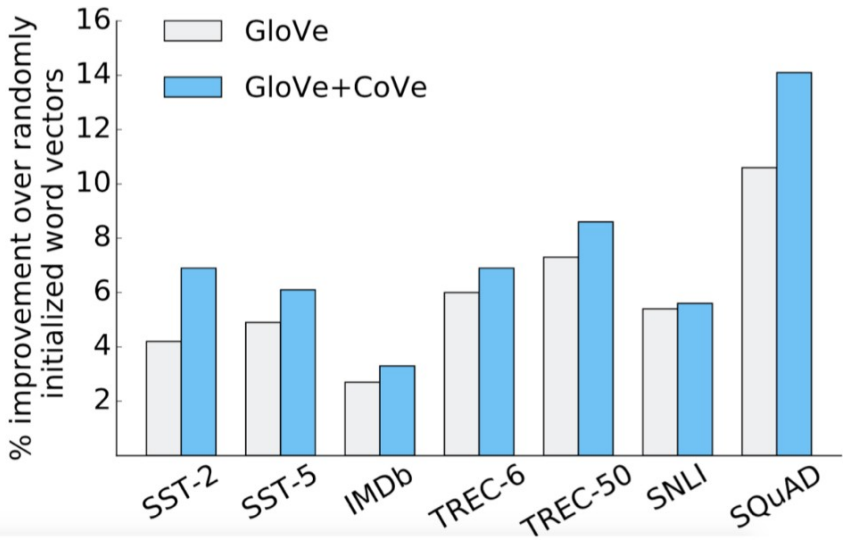
Környzetalapú reprezentációk

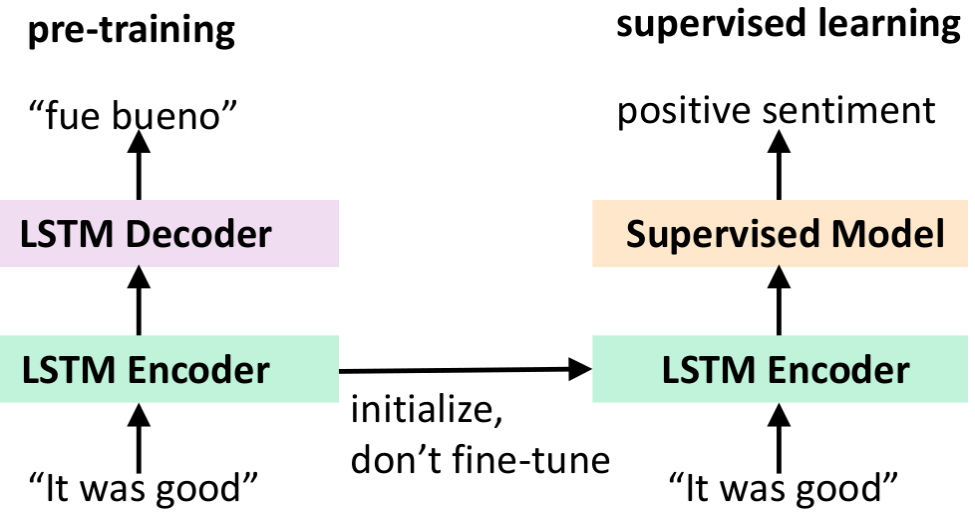
- Motiváció: egy szó pontos jelentése előfordulásonként eltérő
- Multi–task és transzfer learning alapok
  - CoVE – Contextualized Word Vectors (McCann et al.,2017): gépi fordító modell újrahasznosítása



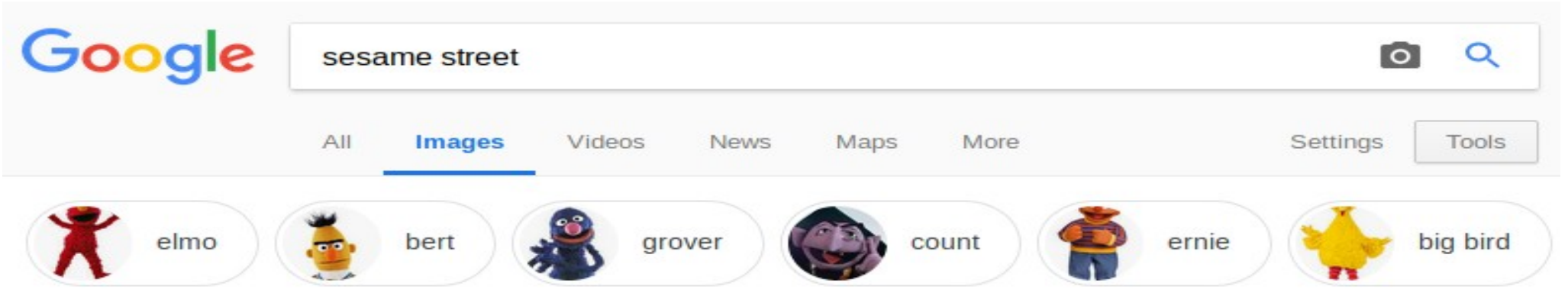
Környezetalapú reprezentációk

- Motiváció: egy szó pontos jelentése előfordulásonként eltérő
- Multi–task és transzfer learning alapok
  - CoVE – Contextualized Word Vectors (McCann et al.,2017): gépi fordító modell újrahasznosítása



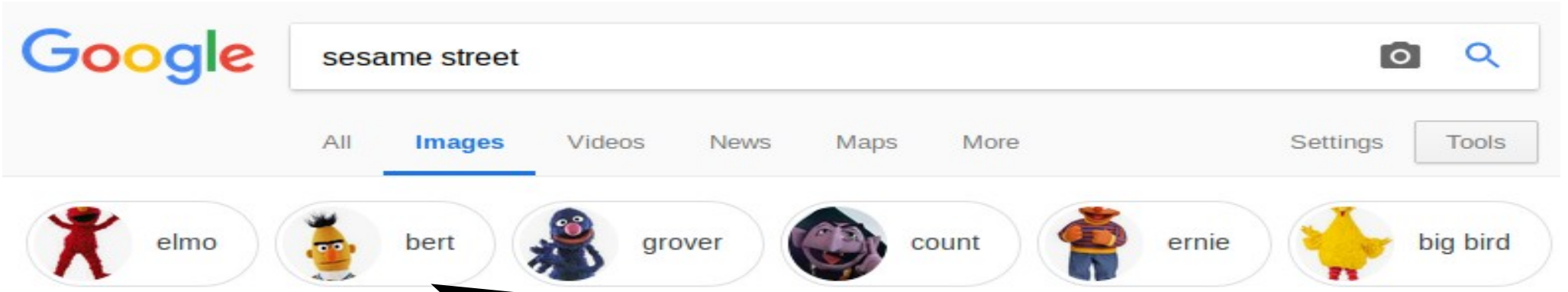


A Szezám utca térhódítása



2018. február
(arxiv.org/abs/1802.05365)
AI2

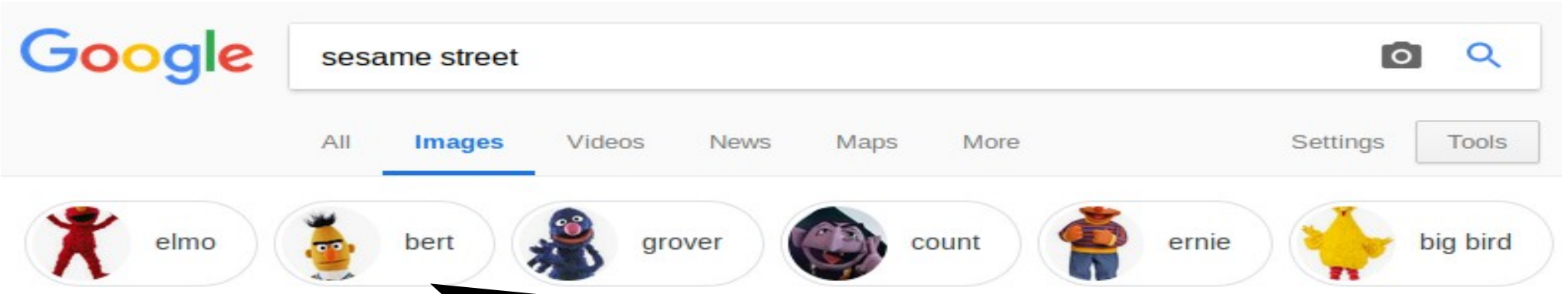
A Szezám utca térhódítása



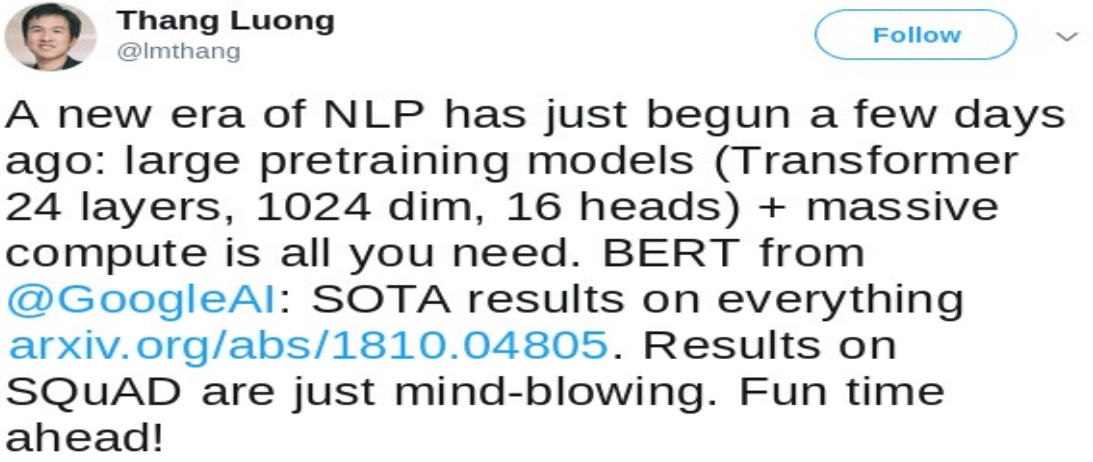
2018. február
(arxiv.org/abs/1802.05365)
AI2

2018. október
(arxiv.org/abs/1810.04805)
Google

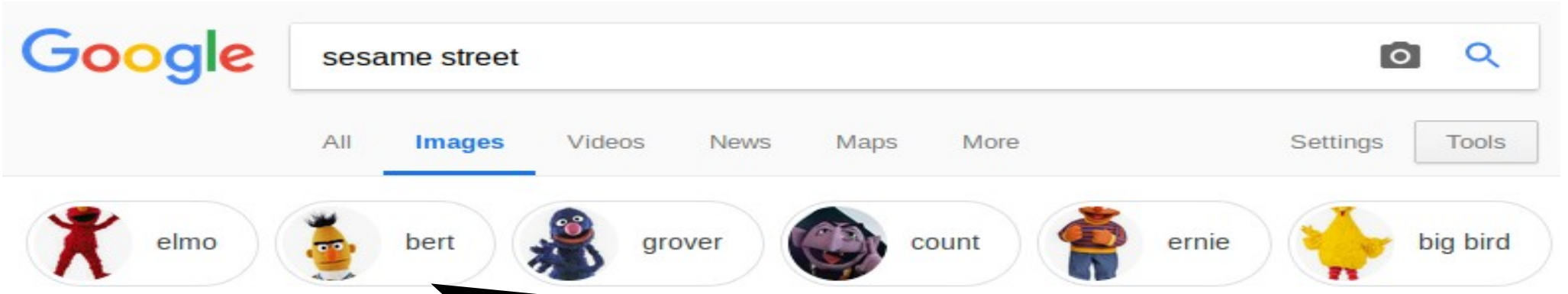
A Szezám utca térhódítása



2018. február
(arxiv.org/abs/1802.05365)
AI2

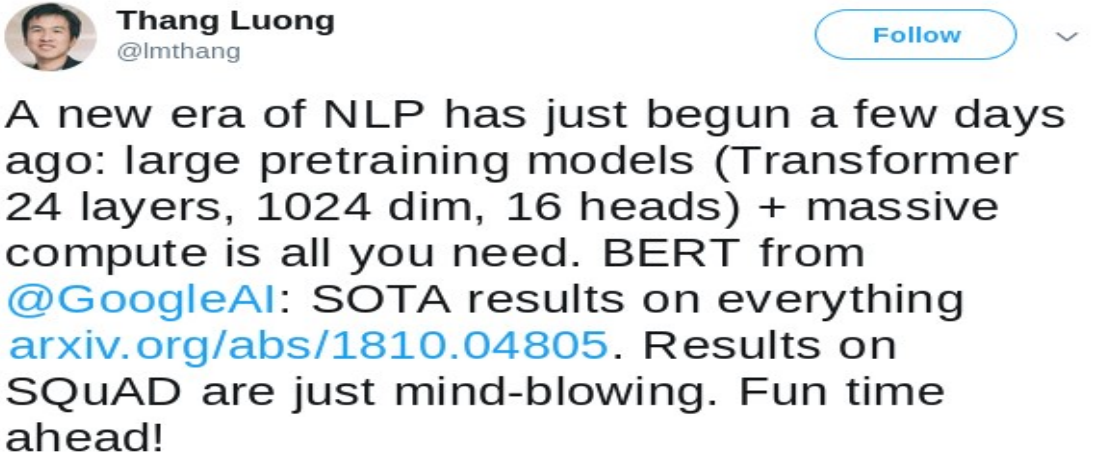


A Szezám utca térhódítása



2018. február
(arxiv.org/abs/1802.05365)
AI2

Demo

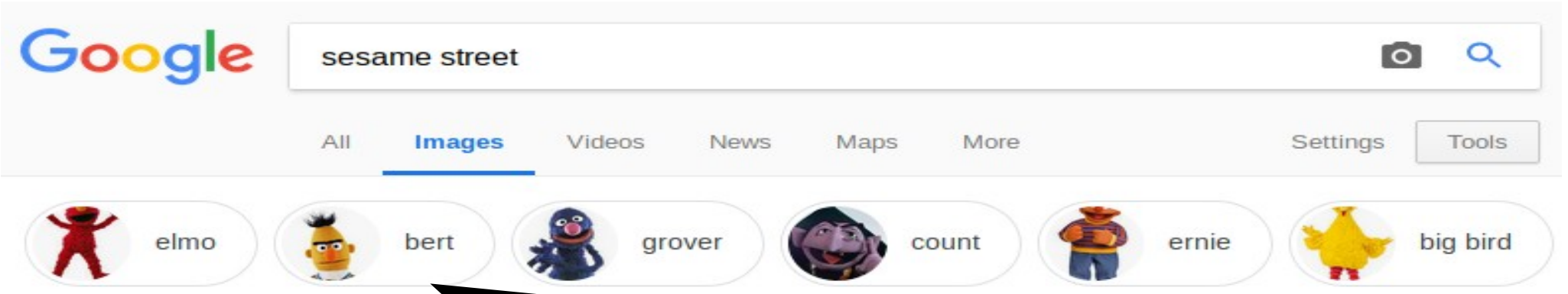


Thang Luong
@lmthang

Follow

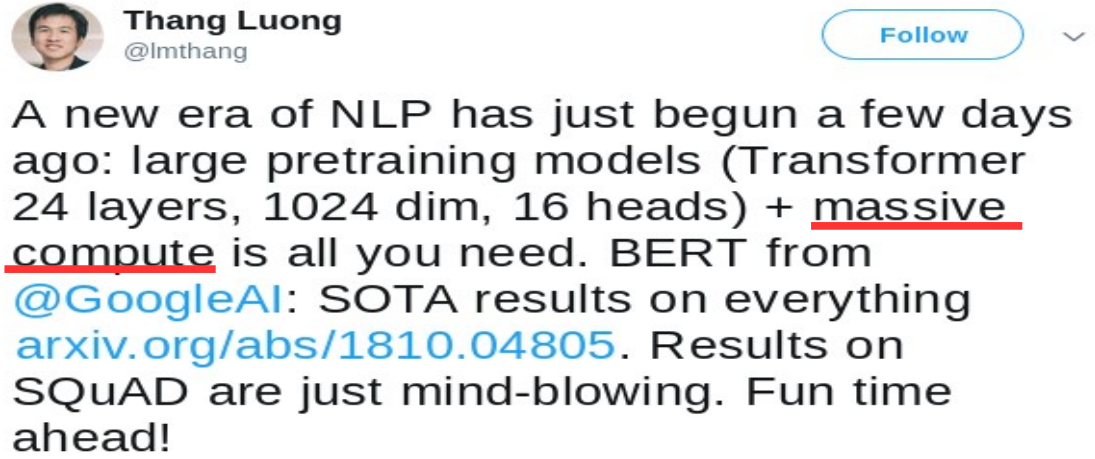
A new era of NLP has just begun a few days ago: large pretraining models (Transformer 24 layers, 1024 dim, 16 heads) + massive compute is all you need. BERT from @GoogleAI: SOTA results on everything arxiv.org/abs/1810.04805. Results on SQuAD are just mind-blowing. Fun time ahead!

A Szezám utca térhódítása



2018. február
(arxiv.org/abs/1802.05365)
AI2

Demo



Thang Luong
@lmthang

Follow

A new era of NLP has just begun a few days ago: large pretraining models (Transformer 24 layers, 1024 dim, 16 heads) + massive compute is all you need. BERT from @GoogleAI: SOTA results on everything arxiv.org/abs/1810.04805. Results on SQuAD are just mind-blowing. Fun time ahead!

Óriási előtanított hálók

- “The total compute used to train this model was 0.96 petaflop days.”

  https://blog.openai.com/language-unsupervised/

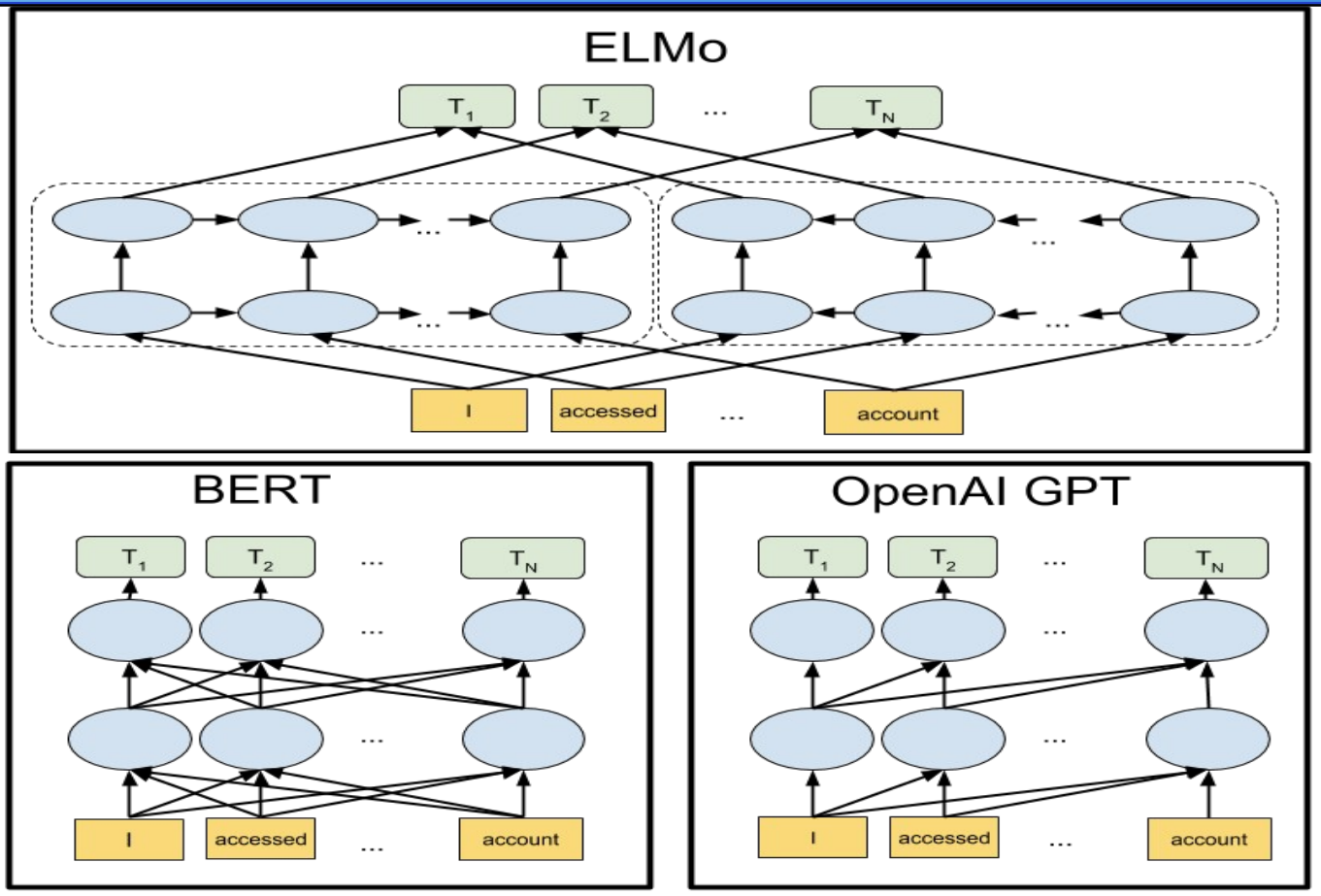
Óriási előtanított hálók

- “The total compute used to train this model was 0.96 petaflop days.”

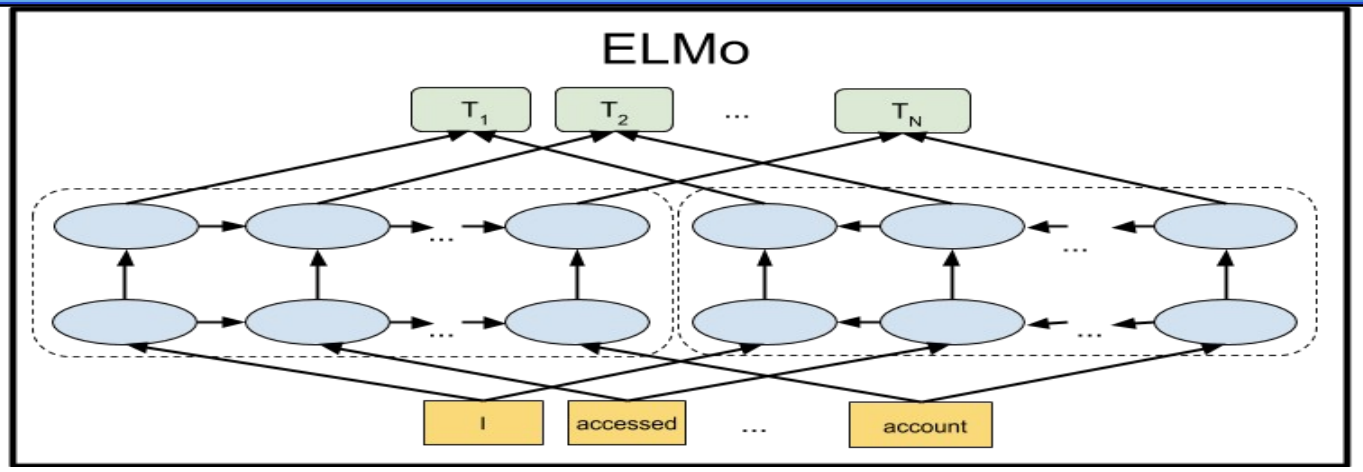
  https://blog.openai.com/language-unsupervised/

- “actually training the model is the piece of work that no one wants to replicate themselves (as it's **~$30k just to train the thing once**)”
  https://github.com/allenai/allennlp/issues/1901

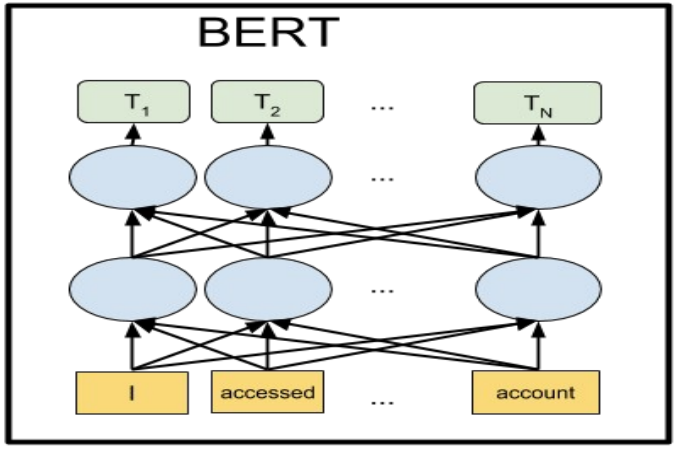
Áttekintés

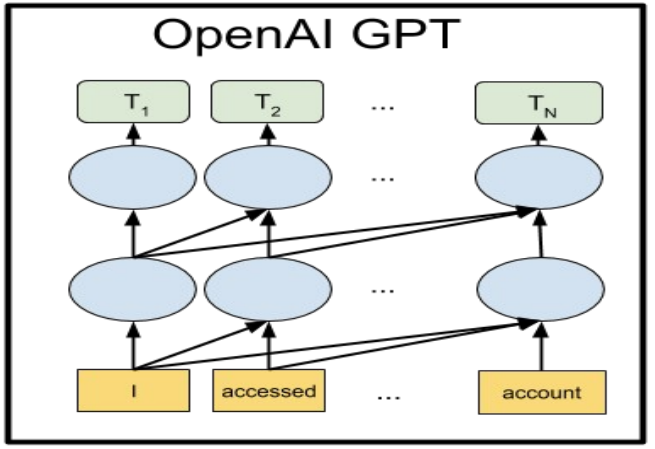

ELMo
T1
T2
...
TN
I
accessed
...
account
BERT
T1
T2
...
TN
I
accessed
...
account
OpenAI GPT
T1
T2
...
TN
I
accessed
...
account

Áttekintés



(bi)LSTM-et használ





Attention alapú Transformer modellre építenek

ELMo előnyös tulajdonságai

- Kontextuális
  - A szavak reprezentációi a környezetükben állók függvényei is

ELMo előnyös tulajdonságai

- Kontextuális
  - A szavak reprezentációi a környzetükben állók függvényei is
- Mély
  - Több réteg reprezentációjának **egyidejű** használata

ELMo előnyös tulajdonságai

- Kontextuális
  - A szavak reprezentációi a környezetükben állók függvényei is
- Mély
  - Több réteg reprezentációjának **egyidejű** használata
- Karakteralapúság (szóalakok helyett)
  - Morfológia hatékonyabb kezelése és OOV

Specializálódó rétegek

- Alacsonyabb: szintaxis

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

POS tagging eredmények (PennTB)

Specializálódó rétegek

- Alacsonyabb: szintaxis ↔ magasabb: szemantika

Model	Acc.
Collobert et al. (2011)	97.3
Ma and Hovy (2016)	97.6
Ling et al. (2015)	**97.8**
CoVe, First Layer	*93.3*
CoVe, Second Layer	92.8
biLM, First Layer	*97.3*
biLM, Second Layer	96.8

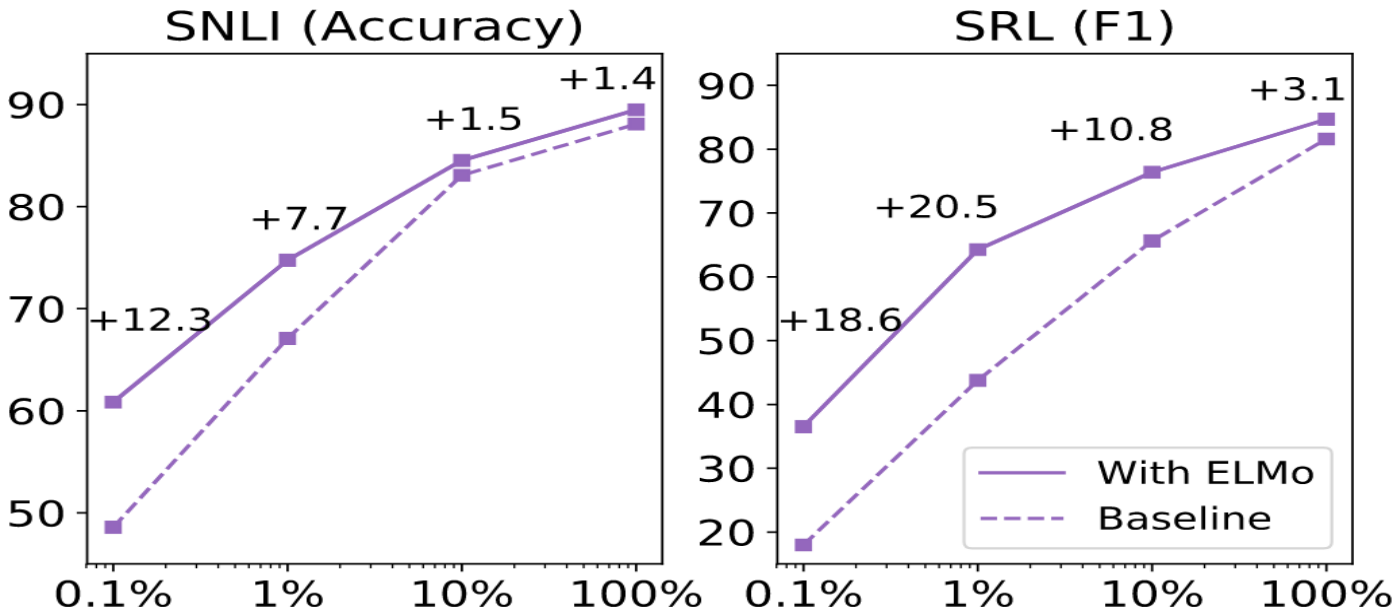
POS tagging eredmények (PennTB)

Model	F_1
WordNet 1st Sense Baseline	65.9
Raganato et al. (2017a)	69.9
Iacobacci et al. (2016)	**70.1**
CoVe, First Layer	59.4
CoVe, Second Layer	*64.7*
biLM, First layer	67.4
biLM, Second layer	*69.0*

WSD eredmények (SemCor 3.0)

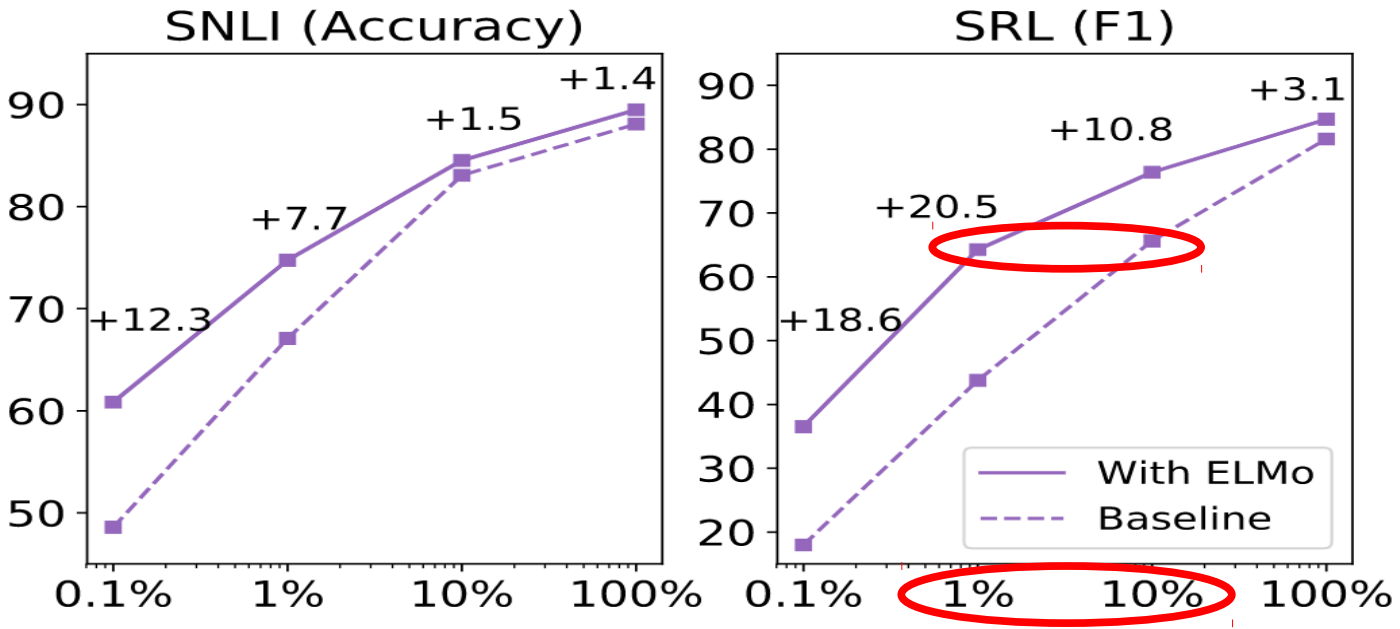
Transfer learning

- Célfeladatra vonatkozó adat jobban hasznosul
  - Konvergencia 486 helyett 10 epoch után SRL-en

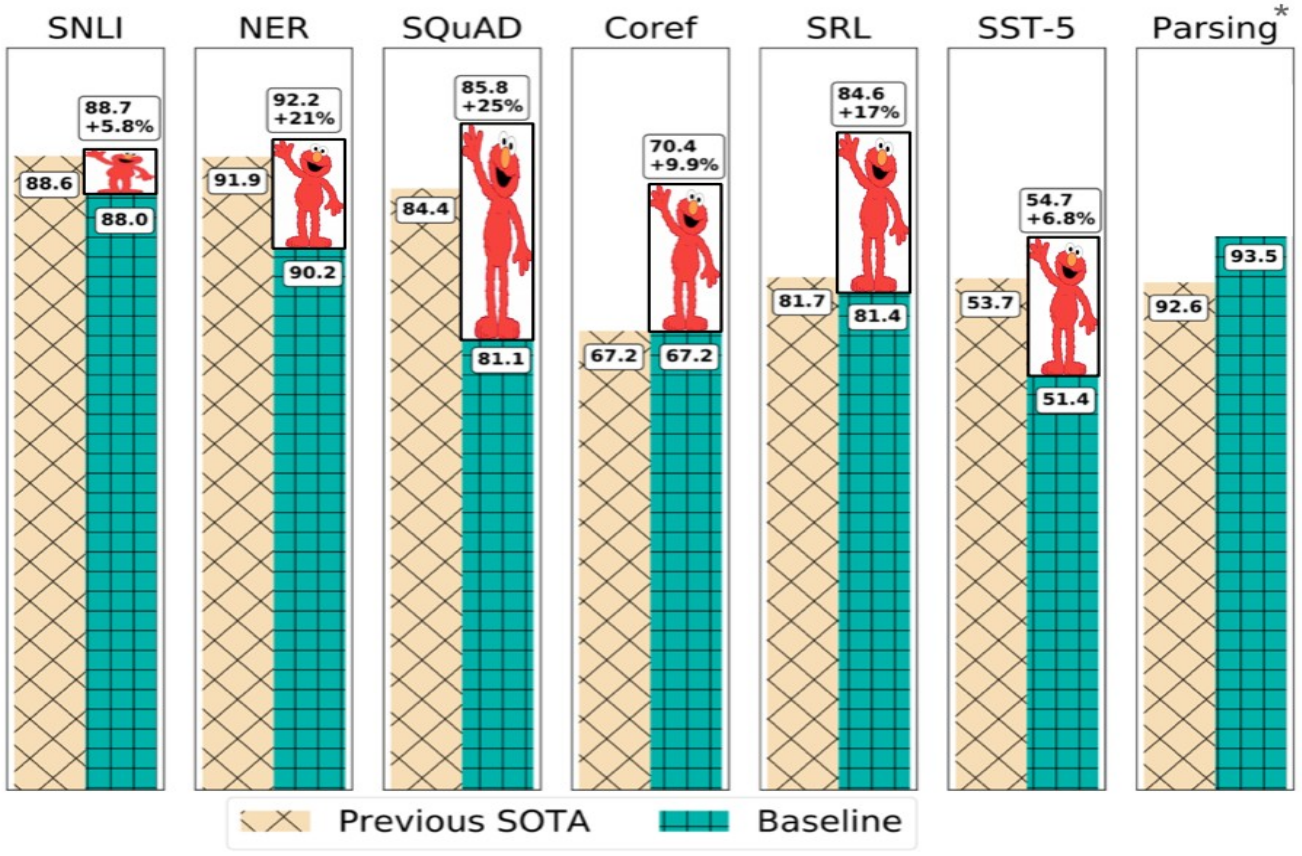


Transfer learning

- Célfeladatra vonatkozó adat jobban hasznosul
  - Konvergencia 486 helyett 10 epoch után SRL-en



ELMo-val elért eredmények

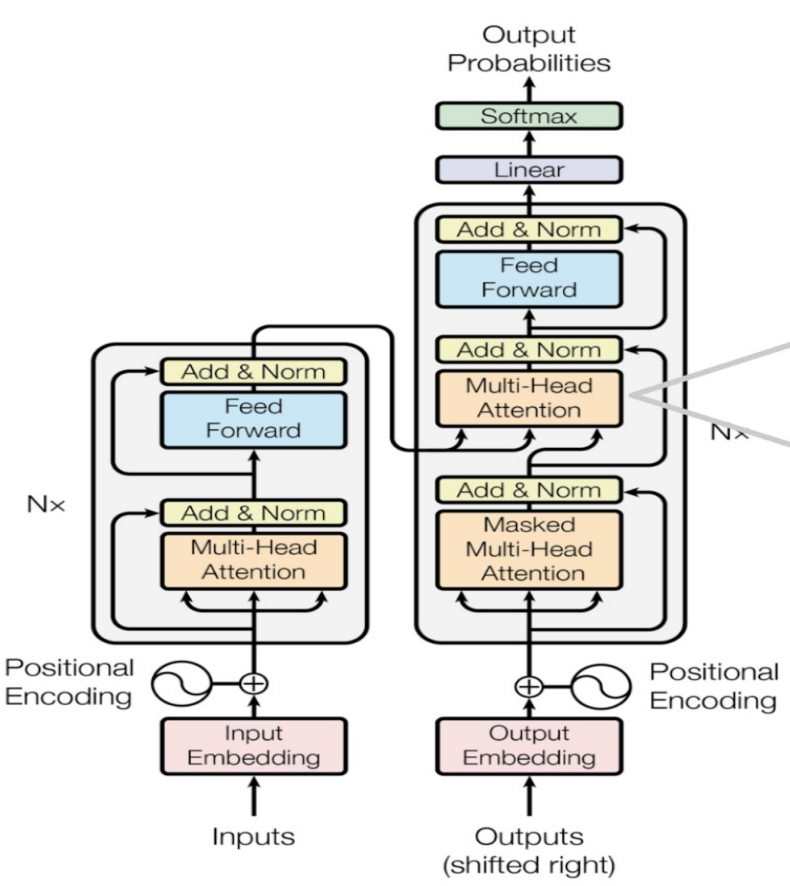


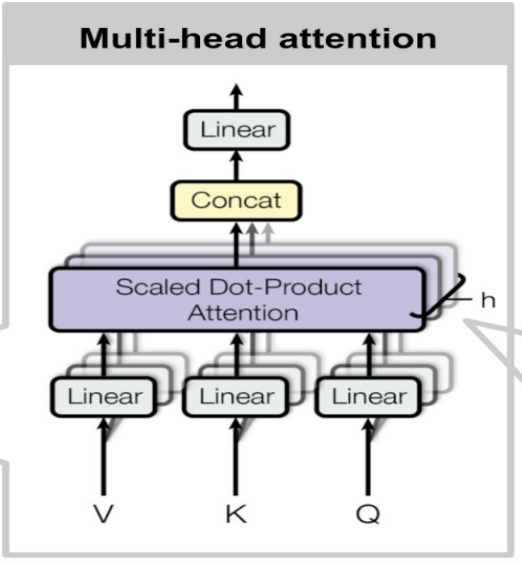
*Kitaev and Klein, ACL 2018 (see also Joshi et al., ACL 2018)

Bidirectional Encoder Representations from Transformers

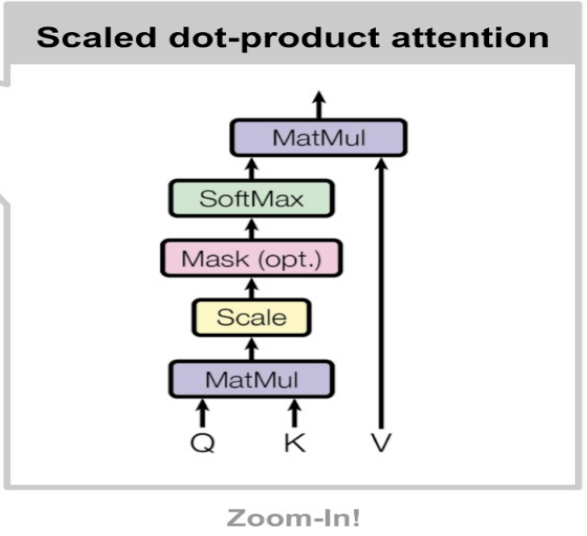
- Kétirányú transzformer (24 blokk, 340M paraméter)
- 40 epoch 3,3mrd szövegszavon tanítva
- WordPiece embeddingek használata
  - Átmenet a szó-és karakteralapúság között (gyakori karakterkombinációkon alapul)
- Maszkolt nyelvi modellezés
- Következő mondat előrejelzése mint segédfeladat
- (bi)LSTM helyett transzformer használata

Transformer



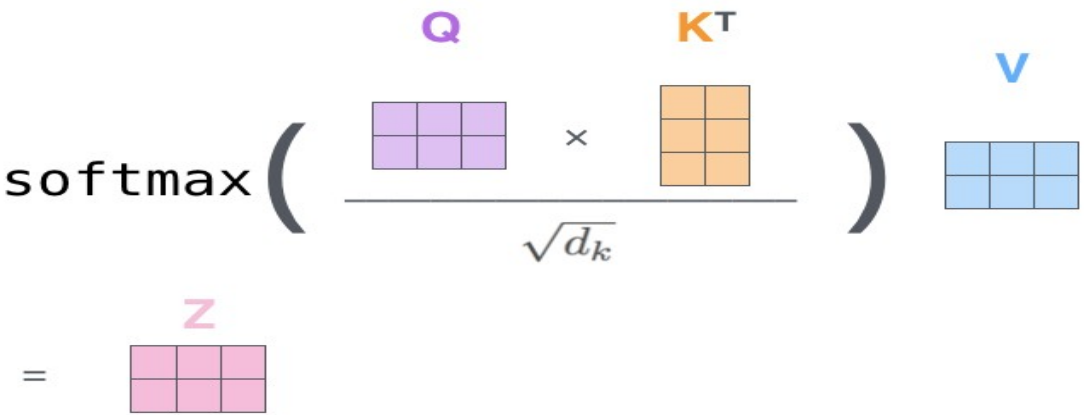


Eredetileg gépi fordításban használt (encoder–decoder) modell

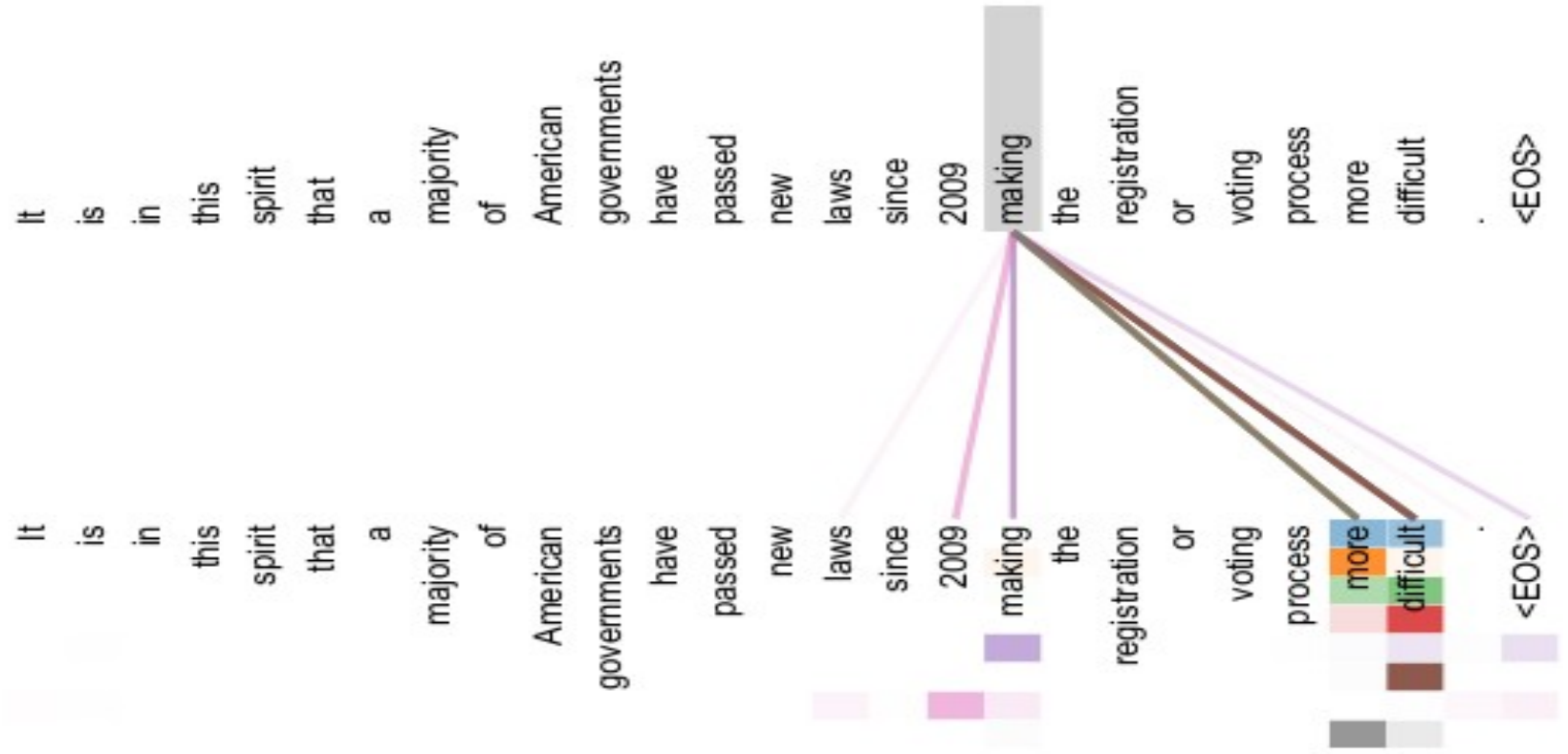


Scaled dot product attention

- Az inputból „query”, „key” és „value” vektorok létrehozása
  - Q, K és V mátrixokba szervezhetők
- A query és key vektorok közötti pontszorzatok a szavak közötti kompatibilitást (figyelem mértékét) adják meg
  - A jobban viselkedő gradiens érdekében sqrt(dim)-mel való osztás



Az attention mechanizmus



Position-wise Feed-Forward Networks

- Minden pozícióra azonosan és függetlenül hajtjuk végre
  - $FFN(x)=max(0,xW1+b1)W2+b2$
- A paraméterek rétegenként eltérnek
- 2048 dimenziós belső reprezentáció (a végeredménye ugyanúgy 512 dim) → kb. 25M paraméter

Bert – Maszkolt nyelvi modellezés

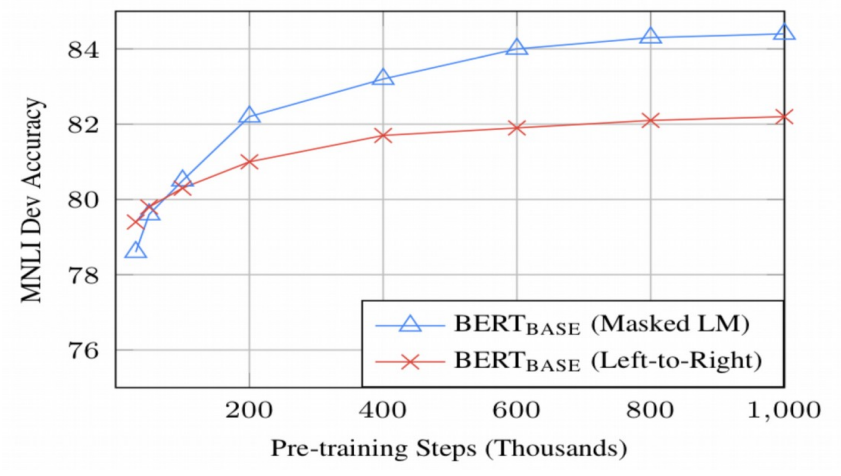
- A mondat “szavainak” 15%-ának véletlenszerű kitakarása után a kitakart szavak visszaállítása
  - És még néhány további trükk

Bert – Maszkolt nyelvi modellezés

- A mondat “szavainak” 15%-ának véletlenszerű kitakarása után a kitakart szavak visszaállítása
  - És még néhány további trükk
- Pazarló az adattal (15% alapján tanul)

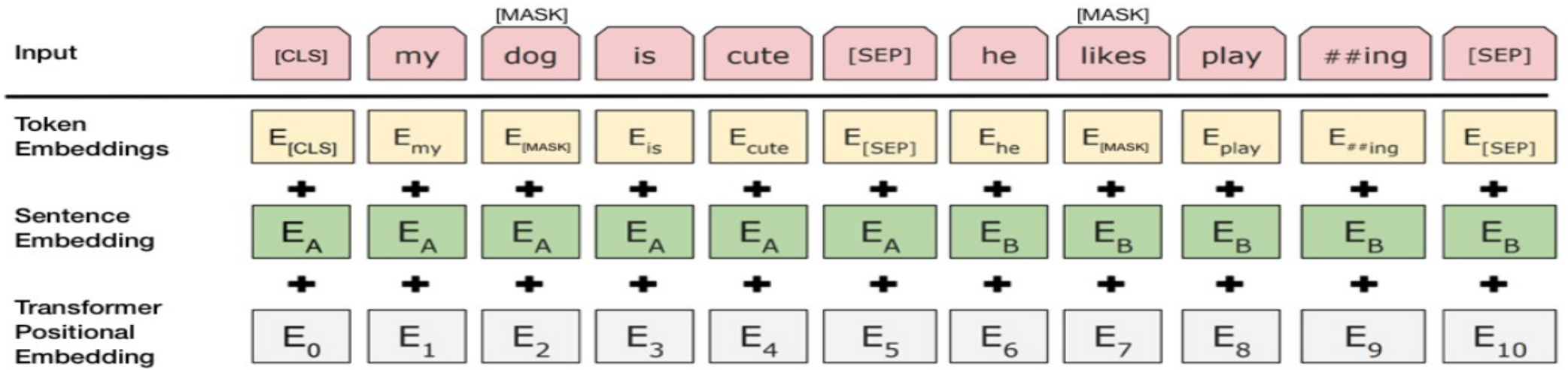
Bert – Maszkolt nyelvi modellezés

- A mondat “szavainak” 15%-ának véletlenszerű kitakarása után a kitakart szavak visszaállítása
  - És még néhány további trükk
- Pazarló az adattal (15% alapján tanul), de megéri



Bert – Mondatpárok osztályozása

- A tanítás során (A, B) mondatpárok jönnek
  - A cél annak eldöntése, hogy B valódi folytatása-e A-nak
  - A tanult modell ~97%-os pontosságú



Bert ablációs eredmények

- NSP: next sentence prediction
- LTR: left-to-right transzformer

Tasks	Dev Set MNLI-m (Acc)	QNLI (Acc)	MRPC (Acc)	SST-2 (Acc)	SQuAD (F1)
$\text{BERT}_{\text{BASE}}$	84.4	88.4	86.7	92.7	88.5
No NSP	83.9	84.9	86.5	92.6	87.9
LTR & No NSP	82.1	84.3	77.5	92.1	77.8
+ BiLSTM	82.1	84.1	75.7	91.6	84.9

Összegzés

- Kontextuális szövegreprezentációk használatával state-of-the-art eredmények
- Az előtanítás ára magas, de szerencsére több nyelvre is elérhetők már előtanított modellek
  - A feladatspecifikus finomhangolás már nem (annyira) költséges

Hivatkozások

- CoVE: arxiv.org/abs/1708.00107
- The annotated transformer: nlp.seas.harvard.edu/2018/04/03/attention.html
- Illustrated transformer: http://jalammar.github.io/illustrated-transformer/
- Cross-View Training: arxiv.org/abs/1809.08370
- NLP’s ImageNet moment: ruder.io/nlp-imagenet/