RUHR UNIVERSITÄT BOCHUM RUB

HYRJE NË INTELIGJENCËN ARTIFICIALE BASHKËKOHORE

Dr. Erion Çano

13 nëntor, 2025

www.trusthlt.org
Trustworthy Human Language Technologies Group (TrustHLT)
Ruhr University Bochum & Research Center Trustworthy Data Science and Security

TrustHLT 1001100110101

CENTER FOR TRUSTWORTHY DATA SCIENCE AND SECURITY

1 Vështrim historik

2 Analiza e figurave

3 Paraqitja e tekstit

4 Arkitektura Transformer

5 LLM-të koduese

6 LLM-të shkoduese

Vështrim historik

Vera e parë (1957 - 1970)

1950 Testi i Turingut

A. M. Turing (Oct. 1950). **"I.—COMPUTING MACHINERY AND INTELLIGENCE".** en. In: *Mind* LIX.236. piv-0, pp. 433–460

1952 Programi i lojës së tavës nga Artur Samuel (*Arthur Samuel*, 1901-1990)

1957 Rrjeti Perceptron nga Frank Rosenblat (*Frank Rosenblatt*, 1928-1971)

1958 Gjuha LISP nga Xhon Me-Karti (*John McCarthy*, 1927-2011)

1959 Emërtimi "Machine Learning" nga Artur Samuel

F. Rosenblatt (Jan. 1957). **The perceptron: A perceiving and recognizing automaton.** Report. Project PARA, Cornell Aeronautical Laboratory

1961 Roboti i parë industrial *Unimate* nis punën te linja e montimit të *General Motors*

1965 Zhvillohet ELIZA, program ndërveprues për dialogim

Dimri i parë (fund ’60 - fund ’70)

Shkaqet:

- Premtimet e tepruara dhe mosarritja e objektivave
- Raporti “Lighthill” i vitit 1973
- Kufizimet e sistemeve ekspert

Pasojat:

- Reduktime të ndjeshme të financimeve
- Rënie e interesit nga praktikuesit dhe industria

Vera e dytë (fund ’70 - fund ’80)

Zhvillimi i sistemeve ekspert që:

- Emulojnë vendimmarrjen e ekspertëve njerëzorë
  - Sistemi MYCIN i diagnozave të sëmundjeve
  - Sistemi XCON për konfigurim të porosive
- Përdorin të dhëna dhe rregulla që zbatohen ndaj tyre

Rritja e investimeve nga qeveritë dhe industria

Zhvillimi i harduerit dhe softuerit të specializuar

- Makinat LISP
- Gjuha Prolog

Dimri i dytë (fund ’80 - fund ’90)

Shkaqet:

- Rrënia e tregut të sistemeve ekspert
  - Përsëri mospërmbushje e pritshmërive
  - Kosto e lartë e zhvillimit dhe mirëmbajtjes
- Rënia e interesit për harduer të specializuar

Pasojat:

- Shumë firma falimentuan
- Skepticizëm dhe rënie e financimit për punë kërkimore në fushën e IA-së

Vera e tretë ('00 - sot)

Premisat:

- Disponueshmëria e sasive të mëdha të të dhënave
- Rritja e kapaciteteve përllogaritëse nëpërmjet GPU-ve
- Mundësia për huazim të infrastrukturës
- Përparimi i teknikave të mësimit të thelluar

Analiza e figurave

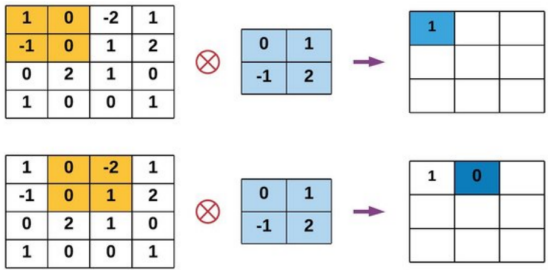
Çfarë përfshin Analiza e figurave

Bashkësi teknikash dhe teknologjishë që shërbejnë për të:

- dalluar kategori të ndryshme objektesh në imazhe
- identifikuar njerëz bazuar te pamja e fytyrës apo shenjat e gishtave
- dalluar the kthyer në tekst shkrimin nëpër figura
- dalluar e klasifikuar deformime apo lloje të ndryshme diagnozash në radiografi, ekografi dhe imazhe të ndryshme mjekësore
- dalluar objektet në lëvizje ose nga një mjet në lëvizje
- ...etj.

Veprimi i thurjes

Veprim matematikor: $(f * g)(t) = \int_{-\infty}^{\infty} f(\tau)g(t-\tau)\,d\tau$



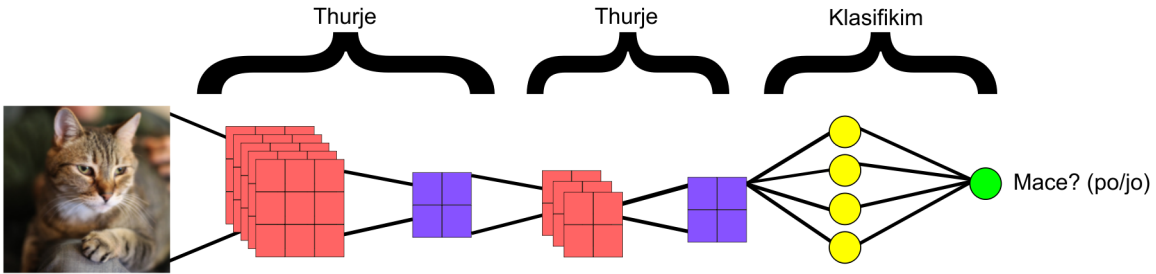
Burimi: `https://www.superdatascience.com`

Filtra të ndryshëm

Mprehje	$\begin{bmatrix} 0 & -1 & 0 \\ -1 & 5 & -1 \\ 0 & -1 & 0 \end{bmatrix}$	
Turbullim	$\frac{1}{9}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{bmatrix}$	

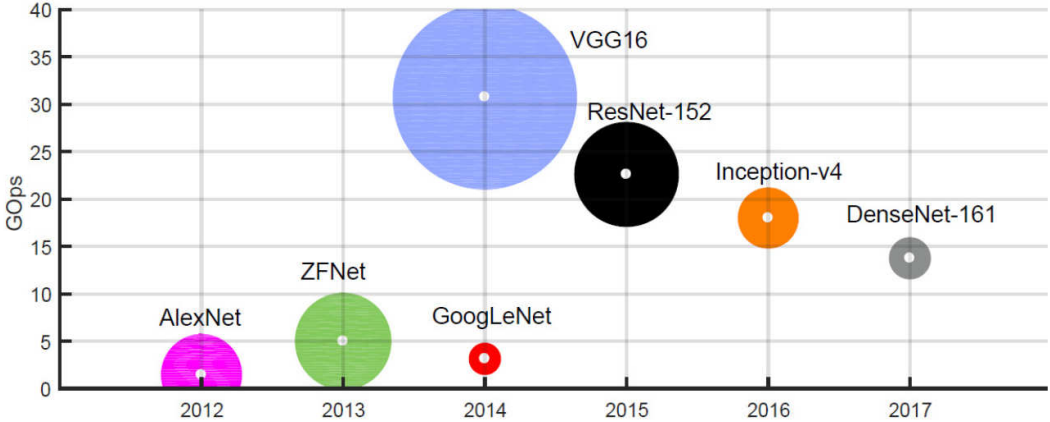
Burimi: `https://www.superdatascience.com`

Dallimi i figurave



Burimi: `https://www.superdatascience.com`

Arkitektura me CNN



Burimi: `https://www.imperial.ac.uk`

Shembull programimi

Dallimi (klasifikimi) i imazheve të shifrave 0 - 9 të shkruara me shkrim dore

...

Paraqitja e tekstit

Vektorizimi sipas shpeshtësisë

	text
0	Eddard Stark is a king in the north.
1	A king but one king : kings are everywhere.
2	Hodor was different : he was not a king .
3	But the North could not change without him.

	king	was	the	not	But	him	one	north	kings	is	in	he	Eddard	everywhere	different	could	change	but	are	Stark	North	Hodor	without
0	1	0	1	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	0	1	0	0	0
1	2	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	1	1	0	0	0	0
2	1	2	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	1	0
3	0	0	1	1	1	1	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1

Burimi: `https://knowledge.dataiku.com`

Vektorizimi TF-IDF

	text
0	Eddard Stark is a king in the north.
1	A king but one king : kings are everywhere.
2	Hodor was different : he was not a king .
3	But the North could not change without him.

	king	was	the	not	a		he	one	north	kings	is	in	him	everywhere	A	different	could	change	but	are	Stark	North	Hodor	Eddard
0	0.333333	0.0	0.5	0.0	0.5	0.0	0.0	0.0	1.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	1.0
1	0.666667	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	0.0
2	0.333333	2.0	0.0	0.5	0.5	1.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0
3	0.000000	0.0	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	1.0	1.0	0.0	0.0	0.0	1.0	0.0	0.0

Burimi: `https://knowledge.dataiku.com`

Mangësi të paraqitjeve matricore

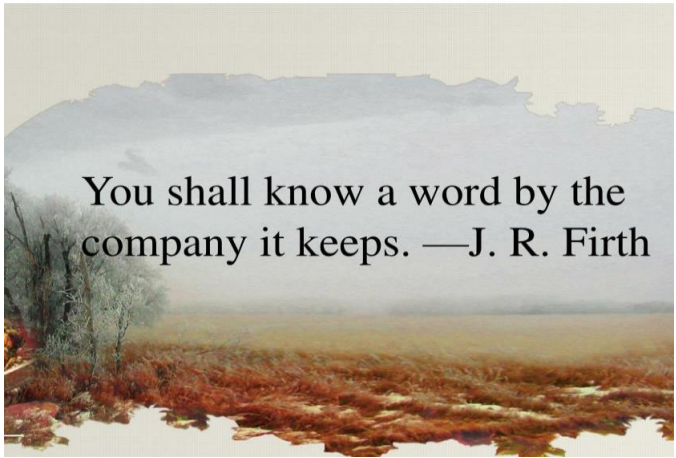
- Humbje e rendit të fjalëve
  - Humbje e saktësisë sintaktore / gramatikore
  - Humbje e semantikës dhe kontekstit
  - Humbje e shprehjeve dhe fjalëve të përbëra
- Strukturë me përmasa shumë të mëdha
  - Përmasa rritet përpjestimisht me fjalorin
  - Kërkon kapacitet të lartë kujtese
- Strukturë shumë e rradhë në përmbajtje
  - Shumica e njësive mbajnë vlerën 0
- Papërputhshmëri me rrjetet nervoere

Vektorët e fjalëve

Pritshmëritë:

- Përmasa të reduktuara
- Strukturë dhe paraqitje e dendur
- Shfrytëzim i frytshëm i kujtesës
- Përdorim në rrjetet nerovore

Rëndësia e kontekstit



Burimi: `https://www.slideserve.com`

Modelimi probabilistik i gjuhëve natyrore

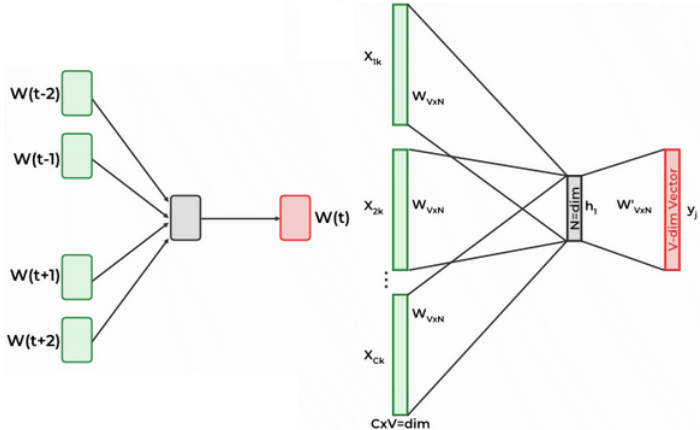
Probabiliteti që të hasim vargun e fjalëve: $w_1^n = w_1, \ldots, w_n$

$P(w_1^n) = P(w_1)P(w_2|w_1)P(w_3|w_1^2), \ldots, P(w_n|w_1^{n-1}) =$
$\prod_{k=1}^{n} P(w_k|w_1^{k-1})$

Përafrimi 2-fjalësh: $P(w_1^n) = \prod_{k=1}^{n} P(w_k|w_{k-1})$

Përafrimi N-fjalësh: $P(w_1^n) = \prod_{k=1}^{n} P(w_k|w_{k-N+1}^{k-1})$

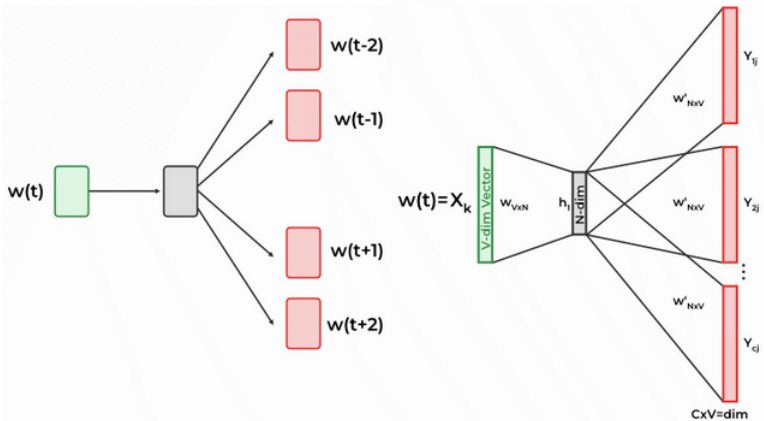
Arkitektura CBOW



T. Mikolov, K. Chen, G. Corrado, and J. Dean (Sept. 2013). **Efficient Estimation of Word Representations in Vector Space.**

Burimi: `https://www.geeksforgeeks.org`

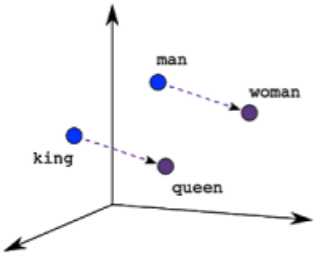
Arkitektura Skip-Gram

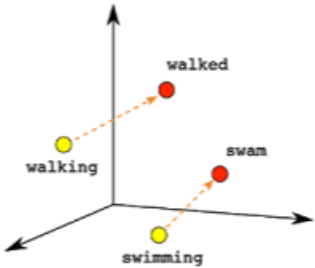


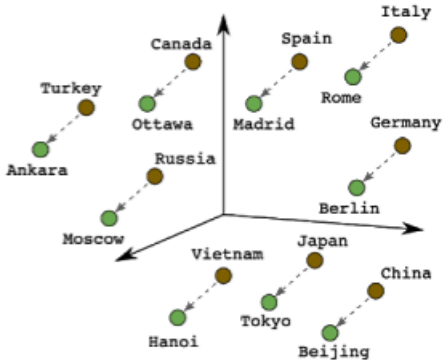
T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). **"Distributed Representations of Words and Phrases and their Compositionality".** In: *Advances in Neural Information Processing Systems*. Ed. by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger. Vol. 26. Curran Associates, Inc.

Burimi: `https://www.geeksforgeeks.org`

Analogjitë midis fjalëve

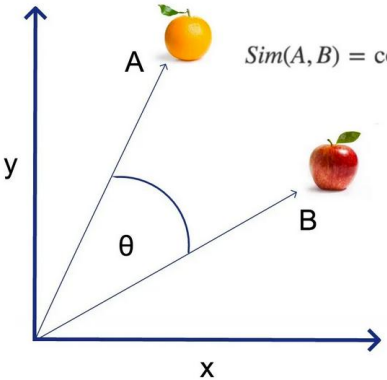




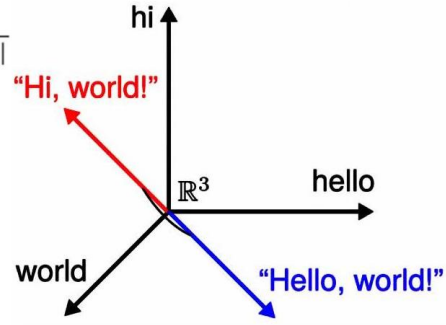


Burimi: `https://tonia.ai/dl-nlp.html`

Ngjashmëria e kosinusit



$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \, \|B\|}$$



Burimi: `https://www.engati.ai`

Shembull programimi

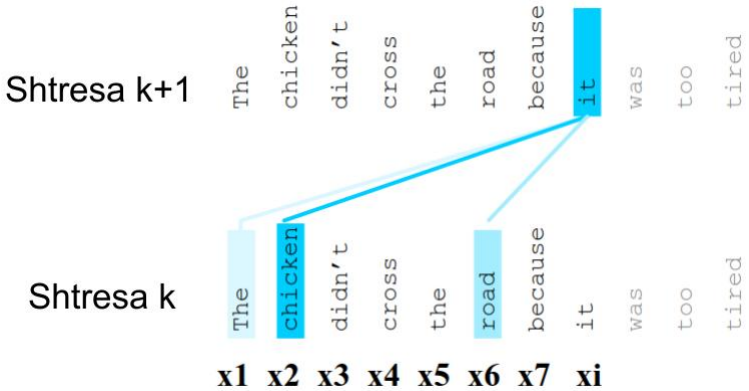
Trajnimi i vektorëve të fjalëve nëpërmjet teksteve të “Gamme of Thrones”

...

Arkitektura Transformer

1 Vështrim historik
2 Analiza e figurave
3 Paraqitja e tekstit
4 Arkitektura Transformer
5 LLM-të koduese
6 LLM-të shkoduese

Koncepti i vetvëmendjes

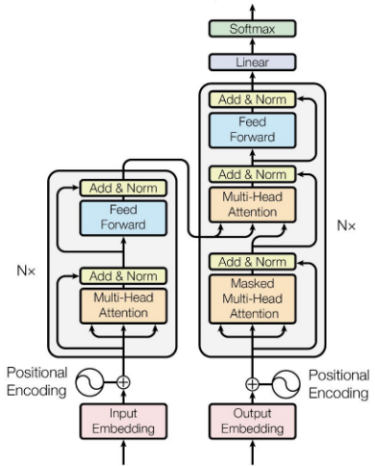


Burimi: https://web.stanford.edu

Koncepti i vetvëmendjes

- Çdo fjalë duhet të paraqitet nëpërmjet vektorëve të ndryshëm sipas ndeshjes së asaj fjale në kontekste të ndryshme
- Vetvëmendja pasuron paraqitjen e fjalëve me informacion nga konteksti
- Për çdo fjalë merren disa vektorë që përputhen me kuptimet e ndryshme të asaj fjale
- Kompleksiteti përllogaritës rritet me zgjerimin e dritares

Arkitektura Transformer



Burimi: `https://llm-class.github.io`

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). **"Attention is All you Need".** In: *Advances in Neural Information Processing Systems.* Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Vol. 30. Curran Associates, Inc.

LLM-të koduese

Kategoritë e LLM-ve

Kodues

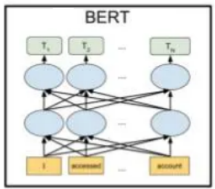
- BERT
- RoBerta
- Reformer
- FlauBERT
- CamemBERT
- Electra
- MobileBERT
- Longformer

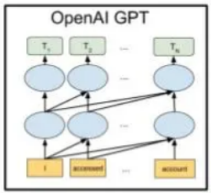
Shkodues

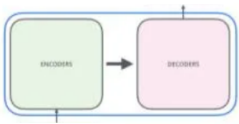
- Transformer-XL
- XLNet
- GPT series
- DialoGPT

Kodues-Shkodues

- Transformer
- XLM
- T5
- BART
- XLM-RoBerta
- Pegasus
- mBART







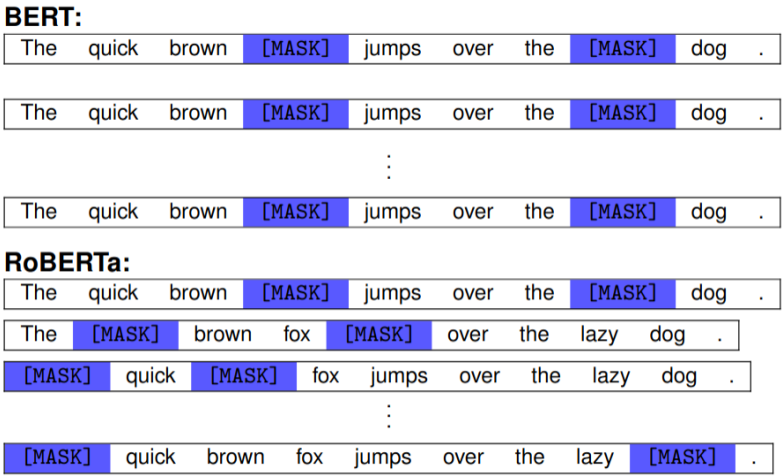
Burimi: `https://www.gabormelli.com`

Modeli BERT

- Trajnim koduesi për zbulim fjalësh të maskuara
- Dy versione:
  1. BERT_{Base} me 12 blloqe kodimi, 12 koka vëmendjeje, vektorë fjalësh me përmasë 768, 110M parametra në total
  2. BERT_{Large} me 24 blloqe kodimi, 16 koka vëmendjeje, vektorë fjalësh me përmasë 1024, 340M parametra në total
- Fjalëndarje në nivel morfemash, me fjalor 30 mijë njësi

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova (June 2019). **"BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding".** In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by J. Burstein, C. Doran, and T. Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186

Maskimi i fjalëve



Y. Liu et al. (2019). **"RoBERTa: A Robustly Optimized BERT Pretraining Approach".** In: *CoRR* abs/1907.11692. arXiv: 1907.11692

Modeli ModernBERT

- Zgjerim i gjatësisë së kontekstit nga 512 në 8192
- Optimizim i shpejtësisë së veprimit të vëmendjes
- Funksion aktivizimi GeGLU, në vend të GELU
- Paketim i vargjeve me gjatësi të ndryshme
- Të dhëna trajnimi më të larmishme

B. Warner et al. (July 2025). **"Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference".** In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 2526–2547

Shembull programimi

Dallimi (klasifikimi) i gjinive të librave bazuar te përmbledhja e tyre.

...

LLM-të shkoduese

Modeli GPT

- Shkodues me 12 shtresa
- Përmasë të shtresës së fshehur 768
- Dritare konteksti prej 512
- Fjalor prej 40 mijë njësishë
- Paratrajnim me korpus librash (cilësi e lartë)

A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever (2018). **"Improving language understanding by generative pre-training".** In

Modeli GPT-2

- Dritare konteksti nga 512 në 1024
- Paratrajnim me 8M dokumente, 40GB tekste
  - Një pjesë e tyre të vjela nga forume
- Numri i shtresave të shkodimit: 12, 24, 36, 48 shtresa shkodimi
  - Përmasat e shtresës së fshehur: 768, 1024, 1280, 1600
- Numri i parametrave: 117M, 345M, 762M, 1542M

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever (2019). **"Language Models are Unsupervised Multitask Learners".** In: *OpenAI*. Accessed: 2024-11-15

Modeli GPT-3

- Dritare konteksti 1024 e sipër
- Numri i shtresave të shkodimit: 12 - 96
- Përmasat e shtresës së fshehur: 768 - 12288
- Paratrajnim me 500B fjalë
  - Wikipedia, CommonCrawl, Books1, Books2, WebText2
- Numri i parametrave: 117M - 175B

T. Brown et al. (2020). **"Language Models are Few-Shot Learners".** In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., pp. 1877–1901

Modelet GPT-4 dhe GPT-5

...???

Pyetje...?

FALEMINDERIT...!