Pregled GPT knjižnic in implementacija lastnega GPT modela

Lena Trnovec

Fakulteta za računalništvo in informatiko

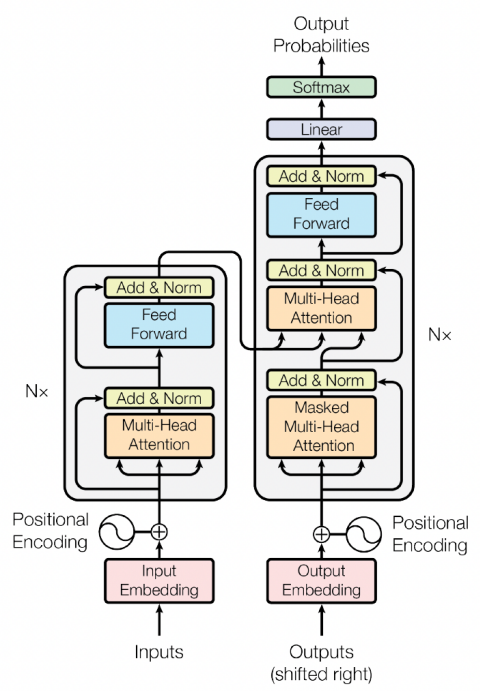
lt89715@student.uni-lj.si

TODO: povzetek.

1 Uvod

Pojav GPT (Generative Pre-trained Transformers) modelov je pomenil pomemben mejnik na področju obdelave naravnega jezika (NLP). Modeli GPT so revolucionarno spremenili naš pristop k nalogam, kot so generiranje besedil, jezikovno prevajanje in semantična analiza. Njihova arhitektura transformerjev (slika 1) se odmika od običajne uporabe rekurentnih nevronskih mrež (RNN) in konvolucijskih nevronskih mrež (CNN) ter namesto tega uporablja kodirnike in dekodirnike, ki so učinkovitejši pri razumevanju in generiranju jezika v kontekstu (*1*).

Z GPT-1 so bili postavljeni temelji arhitekture transformerjev, model GPT-2 je to nadgradil z večjo izpopolnjenostjo in obsegom, model GPT-3 pa je s svojimi 175 milijardami parametrov pomenil velik preskok in pokazal sposobnosti razumevanja in ustvarjanja jezika, ki so bile izjemno podobne človeškim. Vsaka različica je prinesla izboljšave pri razumevanju jezika, razumevanju konteksta in ustvarjanju človeku podobnega besedila v različnih aplikacijah. Kljub prelomnemu potencialu modeli GPT organizacije OpenAI niso odprtokodni, predvsem zaradi



Slika 1: Arhitektura transformerjev (*1*).

pomislekov glede zlorabe in etičnih vidikov. Vendar je vzpon GPT spodbudil odprtokodne alternative, katerih cilj je ponoviti zmogljivosti GPT-3 z odprtokodno etiko.

Namen tega članka je primerjalna analiza različnih odprtokodnih knjižnic GPT, od katerih vsaka ponuja edinstvene zmogljivosti in prednosti za naloge NLP. Cilj je ne le primerjati te knjižnice z vidika zmogljivosti in uporabnosti, temveč tudi preučiti izvedljivost razvoja enostavnega primera modela GPT. Literaturo smo iskali predvsem prek storitve Google Scholar, pri čemer smo uporabili ključne besede, kot sta "GPT models" in "opensource GPT", ter izbrali najbolj relevantne članke. V naslednjih razdelkih bo podan poglobljen pregled in primerjava teh knjižnic, sledil pa bo primer praktične uporabe modela GPT pri posodabljanju starejšega znanstvenega učbenika.

2 Odprtokodne GPT knjižnice

2.1 Pregled

2.1.1 Hugging Face Transformers

Knjižnica Transformers podjetja Hugging Face je postala ključno orodje na področju NLP. Ponuja široko paleto vnaprej-izurjenih (pre-trained) modelov in je znana po svoji vsestranskosti in enostavni uporabi. Pomembna značilnost knjižnice je njena celovita pokritost, ki zajema modele, kot so BERT, GPT-2 in GPT-3, ki so pomembni pri različnih nalogah, od ustvarjanja besedila do razumevanja jezika (*2*). O uporabnosti in priljubljenosti knjižnice priča tudi njena obsežna podpora skupnosti. Knjižnica Hugging Face je ustvarila aktivno skupnost, ki prispeva k nenehnemu izboljševanju knjižnice. To vključuje obsežno dokumentacijo, navodila in skladišče modelov, ki so jih prispevali člani skupnosti, s čimer ponuja bogat vir za začetnike in izkušene strokovnjake na tem področju (*3*).

2.1.2 GPT-Neo/GPT-NeoX by EleutherAI

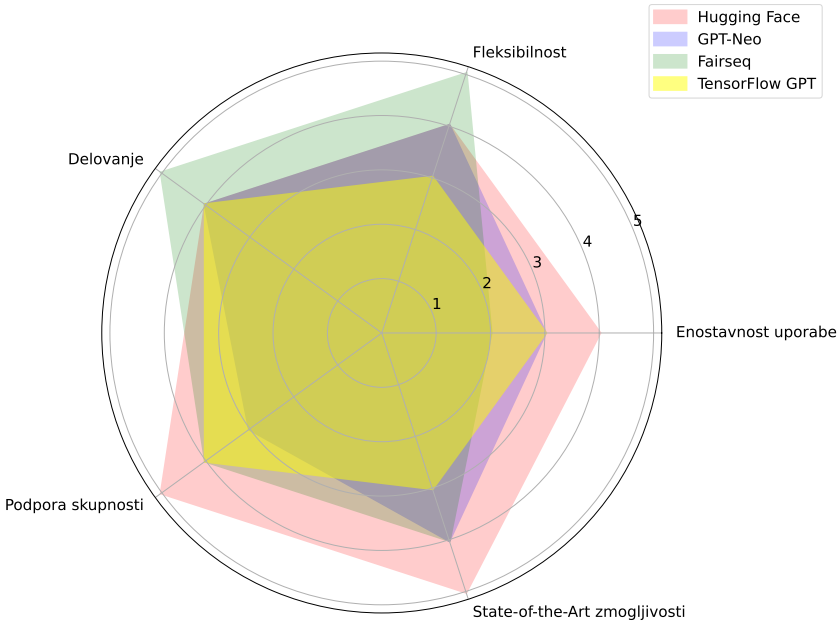
GPT-Neo in GPT-NeoX, ki ju je razvilo podjetje EleutherAI, sta odprtokodni alternativi GPT-3. Zlasti GPT-NeoX je zasnovan za skalabilnost in prilagodljivost ter je namenjen demokratizaciji obsežnih jezikovnih modelov. Ti modeli so pokazali obetavne rezultate pri različnih nalogah NLP, čeprav z nekaterimi razlikami v zmogljivosti v primerjavi z GPT-3 (*4*). V primerjavi s Hugging Face Transformers sta GPT-Neo in GPT-NeoX bolj osredotočena na skalabilnost in sta še posebej primerna za uporabnike z dostopom do znatnih računalniških sredstev. Kljub temu sta zaradi svoje odprtokodne narave dostopna širšemu občinstvu, kar pa spodbuja raziskave in razvoj na področjih, kjer so bili tako veliki modeli prej nedostopni (*4*).

2.1.3 Fairseq

Fairseq, ki ga je razvil Facebook AI Research, je robusten in prilagodljiv nabor orodij za učenje sequence-to-sequence, ki podpira različne modele, vključno s tistimi, ki temeljijo na arhitekturi transformerjev. Njegova sposobnost učinkovitega usposabljanja in napovedovanja ter zmožnost sodelovanja z več grafičnimi procesorji in različnimi računalniškimi vozlišči sta ga uvrstili med priljubljene možnosti za raziskovalce in razvijalce, ki se ukvarjajo z večjimi NLP projekti (*5*). Fairseq ima modularno zasnovo, ki uporabnikom omogoča eksperimentiranje z različnimi arhitekturami modelov in režimi urjenja. Ta prilagodljivost je spodbudila inovacije na področju strojnega prevajanja, jezikovnega modeliranja in drugih področjih NLP. Odprtokodna narava knjižnice in aktivna skupnost prispevata k njenemu nenehnemu izboljševanju in prilagajanju najnovejšim dosežkom na področju globokega učenja in NLP (*6*).

2.1.4 TensorFlow-based GPT modeli

Modeli GPT, ki temeljijo na TensorFlow, se nanašajo na implementacije arhitektur, podobnih GPT, ki uporabljajo TensorFlow, odprtokodno platformo za strojno učenje. Široka uporaba TensorFlow v skupnosti strojnega učenja je privedla do razvoja različnih modelov GPT, prilagojenih za izkoriščanje prednosti TensorFlow, kot sta učinkovita obdelava računskih grafov in močan ekosistem za nameščanje ML modelov v produkcijskih okoljih. Ti modeli pogosto povzemajo arhitekturo GPT podjetja OpenAI, vendar so zasnovani tako, da so bolj dostopni in spremenljivi ter ustrezajo posebnim potrebam raziskovalcev in razvijalcev, ki delujejo v ekosistemu TensorFlow. Modeli GPT, ki temeljijo na TensorFlow, se uporabljajo tudi v izobraževalnih in eksperimentalnih okoljih, saj ponujajo alternativo tistim, ki jim je programska paradigma TensorFlow bolj udobna (*7*).



Slika 2: Primerjava GPT knjižnic iz različnih vidikov.

2.2 Primerjava

Na področju NLP ima vsaka knjižnica GPT svoje prednosti. Hugging Face transformerji imajo uporabniku prijazen vmesnik in širok nabor vnaprej-izurjenih modelov, ki jih podpira močna skupnost. Zaradi obsežnih virov in enostavne uporabe je idealna tako za začetnike kot za strokovnjake. GPT-Neo/GPT-NeoX pa poudarja skalabilnost in odprtokodno dostopnost ter je namenjen uporabnikom, ki potrebujejo obsežne jezikovne modele brez lastniških omejitev. Fairseq je prilagojen za napredne naloge NLP, saj ponuja modularnost in učinkovitost, zlasti v raziskovalnih okoljih. Njegova zasnova omogoča prilagodljivo eksperimentiranje s konfiguracijami modelov. Tensorflow-based GPT modeli pa se odlikujejo po skalabilnem izračunavanju ter integraciji razvoja in uvajanja modelov. Zaradi tega so primerni za uporabnike TensorFlow, ki iščejo učinkovito okolje, pripravljeno za proizvodnjo. Vsaka knjižnica služi različnim potrebam NLP, od enostavne uporabe in podpore skupnosti do skalabilnosti in raziskovalne prilagodljivosti (slika 2).

3 Izdelava preprostega primera

Za svoj primer smo si izbrali implementacijo modela, ki v starejšem učbeniku posodobi vsebino s članki iz zadnjih let. Za razvoj našega modela GPT bodo uporabljeni najnovejši članki iz PubMeda, povezani s tematiko izbrane znanstvene knjige. Ti članki, omejeni na objave iz zadnjih nekaj let, bodo zagotavljali aktualna spoznanja, potrebna za usposabljanje modela.

TODO: Izbira ustrezne knjižnice GPT na podlagi primerjave, vsi koraki za razvoj osnovnega modela, pridobitev podatkov in preprocessing podatkov... Treniranje modela GPT z naborom podatkov, testiranje in validacija...

4 Zaključek

TODO

Literatura

1. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023. arXiv:1706.03762 [cs].

2. Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. arXiv:1910.03771 [cs].

3. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, October 2019. arXiv:1910.13461 [cs, stat].

4. Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, USVSN Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. GPT-NeoX-20B: An Open-Source Autoregressive Language Model, April 2022. arXiv:2204.06745 [cs].

5. Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. fairseq: A Fast, Extensible Toolkit for Sequence Modeling, April 2019. arXiv:1904.01038 [cs].

6. Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook FAIR's WMT19 News Translation Task Submission, July 2019. arXiv:1907.06616 [cs].

7. Martın Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning.