Fakultet primijenjene matematike i informatike Osijek

Sveučilišni diplomski studij Matematika i Računarstvo

Generiranje recepata pomoću GPT-2

Članak

Autor:

Mirna Ladnjak

Osijek, 2023

Fakultet primijenjene matematike i informatike Osijek

Sveučilišni diplomski studij Matematika i Računarstvo

Generiranje recepata pomoću GPT-2

Članak

Mentor:

doc.dr.sc. Domagoj Ševerdija

Autor:

Mirna Ladnjak

Osijek, 2023

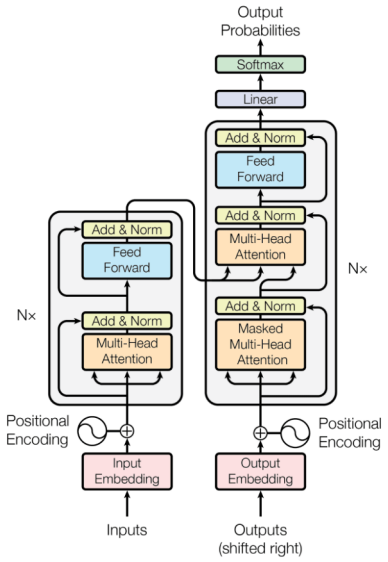
Sadržaj

Uvod

U ovom članku bavit ćemo se problemom generiranja teksta korištenjem GPT-2 modela. GPT-2 je veliki jezični model temeljen na Transformer arhitekturi i treniran na skupu podataka od 8 milijuna web stranica s ciljem prevođenja teksta, odgovaranja na pitanja, sumarizacije teksta, generiranja teksta, itd. Ovaj model daje vrlo dobre rezultate za razne NLP (*eng.* Natural Language Processing) zadatke i može se ponovno natrenirati korištenjem drugog skupa podataka. [4]

Transformer arhitektura

Transformer je arhitektura dubokog učenja koja se oslanja na paralelni mehanizam pažnje s više glava. [2] Prvi puta je predložena sredinom 2017.godine u dokumentu [3] pod nazivom "*Attention is all you need*". Od tada je ova arhitektura korištena u stvaranju najsuvremenijih rezultata od kojih je jedan bio GPT/GPT-2. Poznat je i po tome što zahtijeva manje vremena za treniranje od prijašnjih rekurentnih neuronskih mreža, kao na primjer LSTM (*eng.* Long Short-Term Memory). [7] Transformer arhitektura slijedi enkoder-dekoder arhitekturu, ali se ne oslanja na ponavljanje i konvolucije kako bi se generirao izlaz.



Slika 1: Enkoder-dekoder struktura transformer arhitekture

Ukratko, zadatak enkodera, na lijevoj polovici Transformerove arhitekture na Slici 1, je mapirati ulazni niz u niz kontinuiranih reprezentacija, koje se zatim unose u dekoder.

Dekoder, na desnoj polovici arhitekture, prima izlaz enkodera zajedno s izlazom dekodera u prethodnom vremenskom koraku za generiranje izlazne sekvence. [8]

Korišteni podaci

Ideja je koristiti skup podataka [6] sa *keggle.com* koji u sebi sadrži preko 2 000 000 recepata i na njemu odraditi fino ugađanje (*eng.* fine-tuning) GPT-2 modela te nakon toga generirati nove recepte. Iz skupa podataka prikazanog u obliku tablice na Slici 2 uzimamo informacije o nazivu recepta, potrebnim sastojcima te uputstvima.

	Unnamed: 0	title	ingredients	directions	link	source	NER
0	0	No-Bake Nut Cookies	["1 c. firmly packed brown sugar", "1/2 c. eva...	["In a heavy 2-quart saucepan, mix brown sugar...	www.cookbooks.com/Recipe-Details.aspx?id=44874	Gathered	["brown sugar", "milk", "vanilla", "nuts", "bu...
1	1	Jewell Ball'S Chicken	["1 small jar chipped beef, cut up", "4 boned ...	["Place chipped beef on bottom of baking dish....	www.cookbooks.com/Recipe-Details.aspx?id=699419	Gathered	["beef", "chicken breasts", "cream of mushroom...
2	2	Creamy Corn	["2 (16 oz.) pkg. frozen corn", "1 (8 oz.) pkg...	["In a slow cooker, combine all ingredients. C...	www.cookbooks.com/Recipe-Details.aspx?id=10570	Gathered	["frozen corn", "cream cheese", "butter", "gar...
3	3	Chicken Funny	["1 large whole chicken", "2 (10 1/2 oz.) cans...	["Boil and debone chicken.", "Put bite size pi...	www.cookbooks.com/Recipe-Details.aspx?id=897570	Gathered	["chicken", "chicken gravy", "cream of mushroo...
4	4	Reeses Cups(Candy)	["1 c. peanut butter", "3/4 c. graham cracker ...	["Combine first four ingredients and press in ...	www.cookbooks.com/Recipe-Details.aspx?id=659239	Gathered	["peanut butter", "graham cracker crumbs", "bu...
...	...	...	...	...	...	...	...
995	995	Heath Bar Pie	["3 Heath bars, chopped fine", "1 medium conta...	["Mix chopped Heath bars with whipped topping ...	www.cookbooks.com/Recipe-Details.aspx?id=976718	Gathered	["graham cracker pie crust", "chocolate curls"]
996	996	Victorian Baked French Toast	["1 c. brown sugar", "1/3 c. butter", "2 Tbsp....	["Cook brown sugar, butter and corn syrup in s...	www.cookbooks.com/Recipe-Details.aspx?id=908190	Gathered	["brown sugar", "butter", "light corn syrup", ...
997	997	Quick Swedish Meatballs	["1 lb. ground beef", "1 c. soft bread crumbs"...	["Combine meat, bread crumbs, cheese, soup mix...	www.cookbooks.com/Recipe-Details.aspx?id=850050	Gathered	["ground beef", "bread crumbs", "cream cheese"...
998	998	Irish Stew(Microwave)	["2 lb. lamb, cut in 1-inch cubes", "2 c. wate...	["In 4-quart casserole, combine lamb, 1 1/4 cu...	www.cookbooks.com/Recipe-Details.aspx?id=1017368	Gathered	["lamb", "water", "onion soup", "bay leaf", "c...
999	999	Peach Salad	["2-3 oz boxes Peach Jello", "1 large jar Peac...	["Mix jello with hot water and sugar. Allow to...	www.cookbooks.com/Recipe-Details.aspx?id=65771	Gathered	["Jello", "food", "Condensed Milk", "cream che...

Slika 2: Prvih 1000 redaka iz odabranog skupa podataka

Obrada podataka

Za korištenje GPT-2 modela na odabranim podatcima prije svega potrebno je "očistiti" tekst odnosno ukloniti bilo kakve posebne znakove te odraditi tokenizaciju teksta.

Podaci će modelu biti proslijeđeni kao lista s članovima tipa *ime/upute/sastojci*. Svaki član liste u tom obliku predstavlja informacije za jedan zasebni recept iz skupa podataka kako je prikazano u primjeru na Slici 3. Prije samog treniranja tako uređeni skup podataka dijeli se na skup za treniranje i skup za validaciju.

```
['name: Michigan Sauce Southern Style\ndirections: Put ingredients pot mix
together well cooks medium heat. The meat ketchup mixed prior cooking prevents
meat clumping gives best chili texture. This also cooked crock pot\ningredients:
extra lean ground chuck, ketchup, onion, garlic, cumin, chili powder, cayenne
pepper, salt',
 'name: Alsatian Stuffed Chicken Breasts\ndirections: Saute ham shallots
mushrooms together oil. Slit pocket chicken breast. Divide ham mixture evenly
among breasts. Bake 375 covered dish 25 minutes cooked. Remove cover top chicken
shredded cheese. Broil cheese bubbles browns 5 minutes\ningredients: chicken
breast, deli ham, shallots, baby portabella mushrooms, olive oil, shredded
gruyere',
 'name: Potato Bake\ndirections: Blend ingredients together keeping cheddar
aside. Pour lightly oiled baking dish sprinkle top remaining cheese. Bake hot 30
minutes 350 degrees F. To brown cheese broil minutes end cooking\ningredients:
potatoes, mushroom, onion, paprika, mustard powder, basil, garlic, yogurt, egg,
cheddar cheese']
```

Slika 3: Oblik liste kakva se prosljeđuje modelu

Treniranje modela i rezultati

Treniranje odnosno *fine-tuning* za odabrani skup podataka odrađen je na GPT-2 modelu sa 117 milijuna parametara. Ti parametri su težine i *bias*-i naučeni tijekom procesa treniranja, a koriste se za predviđanje sljedeće riječi u tekstu. To je jedna od manjih verzija GPT-2 modela, budući da je OpenAI trenirao modele sa do 1,5 milijardi parametara.

Nakon 10 izvršenih epoha treniranja, najbolji dobiveni rezultat odnosno generirani recept prikazan je na Slici 4. Promjenom parametara poput količine podataka, duljine rečenica te veličine *batch*-a mogu se dobiti i drugi, ne nužno bolji, rezultati kao što je vidljivo na primjeru sa Slike 5.

```
New Recipe: Grilled Squid with Corn, Tomatoes, and Freshly Tomatoes
ingredients: corn, tomato paste, extra virgin olive oil, unsalted butter, light brown sugar, Kosher salt, egg, baby portabella
mushrooms, shallots, shallots, shallots with feta cheese
directions: Bring a large pot of salted water to a boil. Add the corn and tomato paste and cook stirring occasionally until the
liquid separates from the liquid. Add the tomato paste
```

Slika 4: Generirani tekst Primjer 1

```
New Recipe: Chicken Stock
directions: In large bowl combine chicken stock, salt, pepper, garlic, onion,
cumin, cumin powder, salt, pepper, cumin powder, cayenne pepper. Stir well.
Cover chicken stock tightly. Cover chicken wire rack. Place chicken wire rack.
Place chicken wire rack. Place chicken wire rack. Place chicken wire rack. Place
chicken wire rack. Place chicken wire rack. Place chicken wire rack. Place
chicken wire rack. Place chicken wire
```

Slika 5: Generirani tekst Primjer 2

Evaluacija

Postoji nekoliko evaluacijskih metrika koje se koriste za procjenu izvedbe modela u NLP zadacima, uključujući generiranje teksta. Perpleksnost (*eng.* perplexity) je mjera za koliko pouzdano jezični model predviđa uzorak teksta. Drugim riječima, mjeri koliko je model "iznenađen" kada vidi nove podatke. Što je niža perpleksnost, to bolje model predviđa odnosno generira tekst. [9]

```
New Recipe: Cabbage Salad
directions: Combine cabbage tomatoes dressing mix well. Add salt pepper. Mix well. Serve salad
ingredients: cabbage, tomatoes, salt, pepper, cumin, garlic, cumin, cumin, cumin, cumin, cumin, cumi
n, cumin, cumin, cumin, cumin, cumin, cumin, cumin, cumin, cumin, cumin, cumin, cumin

Perplexity:  5.8645
```

Slika 6: Evaluacija generiranog teksta Primjer 1

```
New Recipe: Brown Rice Casserole
directions: Mix together all ingredients. Cook at 350u00b0 for 30 minutes or until browned on top. Re
move from heat and stir in remaining ingredients. Serve with rice. Makes 4 servings
ingredients: brown rice, brown sugar, eggs, sour cream, sour cream, cream of chicken soup, sour crea
m, cream of chicken soup, sour cream, cream of chicken soup, sour cream, cream of chicken soup, sour
cream,

Perplexity:  9.8221
```

Slika 7: Evaluacija generiranog teksta Primjer 2

Zaključak

Iako ih je ponekad teško razlikovati od ljudskog govora, uzorci teksta proizvedeni GPT-2 modelom mogu postati vrlo repetitivni uz besmislene rečenice kada je riječ o dužim tekstovima. [5] Ipak, ako pogledamo mjeru perpleksnosti za generirani tekst nakon *fine-tuneanog* GPT-2 modela i uzevši u obzir da je korišten samo mali udio iz originalnog skupa podataka, moglo bi se reći da ovaj model daje dovoljno dobre rezultate. U svrhu poboljšanja rezultata može se pokušati uzeti veći udio podataka i zatim ponovo trenirati model mijenjajući parametre kako bi se postigli najbolji mogući rezultati.

Literatura

[1] Web izvor dostupan na https://blog.knoldus.com/what-are-transformers-in-nlp-and-its-advantages/. (*Transformers In NLP*).

[2] Lewis Tunstall, Leandro von Werra, Thomas Wolf., *Natural Language Processing with Transformers.*, Published by O'Reilly Media, 2022.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar., *Attention Is All You Need*, 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[4] Cari na Geerlings, Albert Mero No-Pe Nuela., *Interacting with GPT-2 to Generate Controlled and Believable Musical Sequences in ABC Notation*

[5] owen Tan, Zichao Yang, Maruan Al-Shedivat., *Progressive Generation of Long Text with Pretrained Language Models.*, Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 4313-4324, June 6-11, 2021. ©2021 Association for Computational Linguistics

[6] Web izvor dostupan na https://www.kaggle.com/datasets/paultimothymooney/recipenlg?datasetId=1025978&sortBy=dateRun&tab=profile. (*Cooking recipes dataset*).

[7] Web izvor dostupan na https://en.wikipedia.org/wiki/Transformer_(machine_learning_model).

[8] Web izvor dostupan na https://machinelearningmastery.com/the-transformer-model/.

[9] Web izvor dostupan na https://www.techslang.com/perplexity-in-nlp-definition-pros-and-cons/. (*Perplexity metric*).