NLP – Peti domaći zadatak

Višejezična klasifikacija toksičnih komentara

Sadržaj

Uvod

Višejezičnu klasifikaciju toksičnih komentara, po nivou toksičnosti u Python programskom jeziku, nad ogromnim skupom podataka, nešto više od 300 000 redova, gde najduži komentar u redu ima 1403 reči, radimo koristeći BERT transformers, kao i Tenzorsku procesorksu jedinicu, unutar Google Colab okruženja. Skup podataka potiče sa Kaggle vebsajt-a. TPU koristimo zato što bi treniranje pomenutog modela, nad ovako ogromnim skupom podataka, trajalo veoma dugo.

Glavna oblast fokusa su modeli mašinskog učenja koji mogu da identifikuju toksičnost u razgovorima, gde se toksičnost definiše kao bilo šta nepristojno, nepoštovanje ili na drugi način, verovatnoću da će nekoga naterati da napusti diskusiju. Ako se ovi toksični doprinosi mogu identifikovati, mogli bismo imati bezbedniji internet sa više međuljutske saradnje.

Tenzorska procesorska jedinica (TPU)

Tenzorska procesorska jedinica, TPU je integrisano kolo (ASIC), specifiče namene za aplikaciju akceleratora veštačke inteligencije, koje je razvio Google, a pomaže u smanjenju vremena obuke na modelima dubokog učenja posebno u Google Tensorflow paketu. Obuka modela dubokog učenja na TPU-u uređaju je mnogo jednostavnija u Tensorflow Python biblioteci u poređenju sa Py-Thorch bibliotekom.

Pošto je veštačka inteligencija zasnovana na principu množenja matrica, nije teško sastaviti ASIC čip, koji ima specijalnu namenu, množenje matrica, a to je ujedno i sve što taj čip može da radi. Time je postignuta ogromna efikasnost i brzina, kojom se rešava specifični problem.

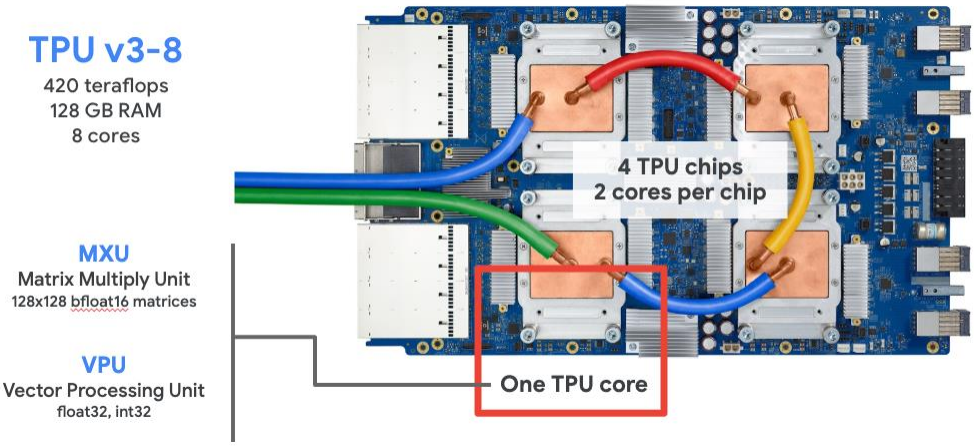

Figure 1 - TPU v3

TPU v3 ploča sadrži četiri TPU čipa i 32GB HBM memorije, jedne od najbržih i najskupljih RAM memorija na svetu (postoji HBM2). Svaki TPU čip sadrži dva jezgra. Svako jezgro ima matričnu jedinicu za množenje (MXU), vektorsku jedinicu i skalarnu jedinicu. Svaka matrična jedinica je sposobna da izvrši 16 000 operacija množenja i akumulacije u svakom ciklusu koristeći format brojeva bfloat16.

Koliko je dobra integracija TPU uređaja u samoj Tensorflow biblioteci pokazuje i količina koda, neophodnog za uspešno povezivanje TPU-a. U slučaju da TPU ne bude detektovan, definisan je fallback na GPU ili CPU, kao primarne uređaje za treniranje modela.

```
try:
    tpu = tf.distribute.cluster_resolver.TPUClusterResolver()
    print('TPU konekcija ', tpu.master())
except ValueError:
    tpu = None
if tpu:
    tf.config.experimental_connect_to_cluster(tpu)
    tf.tpu.experimental.initialize_tpu_system(tpu)
    strategy = tf.distribute.experimental.TPUStrategy(tpu)
else:
    strategy = tf.distribute.get_strategy()
```

Figure 2 - TPU konekcija

Prilikom uspešno uspostavljene konekcije sa TPU uređajem, dobijamo povratnu informaciju u vidu porta i IP adrese. Komunikaciju sa TPU uređajem vrši specijalna virtualna mašin, putem gRPC protokola.

```
TPU konekcija  grpc://10.0.0.2:8470
```

Figure 3 - TPU IP

Transformersi i BERT

Tehnologija Transformersa, koja je u potpunosti promenila NLP svet, razlog je današnje neverovatne NLP tehnologije. Transformersi nadmašuju model prevođenja Google Neural Machine u određenim zadacima. Najveća korist, je vidljiva u načinu paralelizacije.

BERT-ova ključna tehnička inovacija je primena dvosmerne obuke Transformersa. Ovo je u suprotnosti sa prethodnim naporima koji su posmatrali sekvencu teksta bilo s leva na desno ili kombinovanu obuku s leva na desno i zdesna nalevo. Rezultati pokazuju da jezički model koji je dvosmerno obučen, može imati dublji osećaj jezičkog konteksta i toka, nego jednosmerni jezički modeli.

Nakon izbora uređaja definišemo enkoder za kodiranje teksta u niz celih brojeva, koji će služiti kao ulaz za BERT.

```
chunk=256, maxlen=512
tokenizer.enable_truncation(max_length=maxlen)
tokenizer.enable_padding(length=maxlen)
ids = []

for i in (range(0, len(texts), chunk)):
    text_chunk = texts[i:i+chunk].tolist()
    encoders = tokenizer.encode_batch(text_chunk)
    ids.extend([e.ids for e in encoders])

return np.array(ids)
```

Figure 4 - Enkoder

Zatim radimo tokenizaciju, korišćenjem poznate HuggingFace biblioteke, zbog pripreme ulaza za model. Delimo podatke na podatke za terning, validaciju i testiranje, a potom kreiramo i setove podataka.

```
tokenizer = transformers.DistilBertTokenizer.from_pretrained('distilbert-base-multilingual-cased')
tokenizer.save_pretrained('.')
# Reload with the huggingface tokenizers library
fast_tokenizer = BertWordPieceTokenizer('vocab.txt', lowercase=False)
fast_tokenizer

x_train = fast_encode(train1.comment_text.astype(str), fast_tokenizer, 256, MAX_LEN)
x_valid = fast_encode(valid.comment_text.astype(str), fast_tokenizer, 256, MAX_LEN)
x_test = fast_encode(test.content.astype(str), fast_tokenizer, 256, MAX_LEN)

y_train = train1.toxic.values
y_valid = valid.toxic.values
```

Figure 5 - Tokenizacija

Podaci su podeljeni, tako da imamo 63 811 ulaznih redova za test podatke i 223 548 redova komentara za trenažne podatke.

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
223544	fff8f64043129fa2	:Jerome, I see you never got around to this...! ...	0	0	0	0	0	0
223545	fff9d70fe0722906	==Lucky bastard== \n http://wikimediafoundatio...	0	0	0	0	0	0
223546	fffa8a11c4378854	==shame on you all!!!== \n\n You want to speak...	0	0	0	0	0	0
223547	fffac2a094c8e0e2	MEL GIBSON IS A NAZI BITCH WHO MAKES SHITTY MO...	1	0	1	0	1	0
223548	fffb5451268fb5ba	" \n\n == Unicorn lair discovery == \n\n Suppo...	0	0	0	0	0	0

Figure 6 - Trenažni podaci

Sada je moguće konstruisati BERT model, koji je treniran tokom 3 epohe.

```
def bertModel(transformer, max_len=512):
    words = Input(shape=(max_len,), dtype=tf.int32, name="words")
    out = transformer(words)[0]
    token = out[:, 0, :]
    out = Dense(1, activation='sigmoid')(token)
    model = Model(inputs=words, outputs=out)
    model.compile(Adam(lr=1e-5), loss='binary_crossentropy', metrics=['accuracy'])
    return model
```

Figure 7 - BERT model

Treniranje BERT modela na TPU uređaju je trajalo 7,5 minuta, pri čemu je nivo iskorišćenosti matrične jedinice za množenje bio svega 20%.

```
Epoch 1/3
1746/1746 [==============================] - 182s 81ms/step - loss: 0.1773 - accuracy: 0.9316 - val_loss: 0.4778 - val_accuracy: 0.8471
Epoch 2/3
1746/1746 [==============================] - 132s 75ms/step - loss: 0.0917 - accuracy: 0.9638 - val_loss: 0.4267 - val_accuracy: 0.8506
Epoch 3/3
1746/1746 [==============================] - 132s 76ms/step - loss: 0.0791 - accuracy: 0.9678 - val_loss: 0.5202 - val_accuracy: 0.8504
```

Figure 8 - TPU treniranje

Pokušaj treniranja istog modela preko Tesla P100 grafičke karte, sa 16GB HBM2 memorije i 250w maksimalne potrošnje električne energije, demonstrirao je pravu moć TPU uređaja. Za samo jednu epohu, procenjeno vreme trajanja treninga, bilo je oko 50 minuta pri samom pokretanju. Ova grafička karta je,

poređenja radi, duplo brža od Nvidia RTX 2080Ti karte, sudeći po Tensor TFLOPS performansama. Maksimalna potrošnja električne energije TPU v3 uređaja je takođe 250w, ali je TPU isti posao odradio neuporedivo brže uz samo 20% iskorišćenosti.

```
345/13971 [..............................] - ETA: 48:11 - loss: 0.2323 - accuracy: 0.9109
```

Figure 9 - GPU trening

Provera

Komentara na srpskom jeziku nije bilo u ovom setu podataka, pa da vidimo kako će se BERT model izboriti sa našim neprimerenim govorom. Pregledom naredne tabele, možemo zapaziti da se model ne snalazi loše sa našim jezikom, iako nije imao ulazne podatke na srpskom, tokom treniranja.

lang	toxic	content
sr	0.0232905	Govedo onobrdsko
sr	0.6357938	Gospodjica mi je licno veoma nesimpaticna, njenom partijom koja opljackala i ponizila narod Srbije, trebalo bi da se pozabave policijski i istrazni orga
sr	97.911766	O junače, obrni se k meni -mrtvoga te majka obrtala
sr	0.6188363	E, šipak ćeš da dobiješ ono što hoćeš
sr	87.365204	NE UPOZORAVAJU ME JER GUBE IZ UPOZORENJA SA DIPLOMIRANIM INFORMATICIMA DRŽAVE ARGENTINE I VERUJTE MI, IZGUBIĆE
sr	99.89342	Tokom rada na Vikipediji pokazao si se kao štreber, kučka, kreten, seronja, kondom, je**k, f***b, 3,14 zda, 2,71 f*** ti. Tvoja majka! Gubi se odavde

Figure 10 - Srpski komentari

Zaključak

Demonstracija nadmoći TPU uređaja je zaista impozantna. Uverili smo se, koliko je NLP napredovao poslednjih goidna, najviše zahvaljujući Transformersima. Iako smo imali ogroman set podataka sa višejezičnim komentarima, tačnost pri treniranju je bila poprilično visoka.

Opisani model može biti od velike koristi u sprečavanju sajber nasilja, govora mržnje, kao i zaštite najmlađih pojedinaca našeg društva, filtriranjem neprimerenog, toksičnog sadržaja.

Literatura

1. https://cloud.google.com/tpu/docs/system-architecture-tpu-vm#:~:text=A%20TPU%20core%20contains%20one,power%20in%20a%20TPU%20chip
2. https://www.gpumag.com/gddr5-gddr5x-hbm-hbm2-gddr6/
3. http://jalammar.github.io/illustrated-transformer/
4. https://www.kaggle.com/c/jigsaw-multilingual-toxic-comment-classification/data
5. https://www.kaggle.com/docs/tpu
6. https://www.analyticsvidhya.com/blog/2019/09/demystifying-bert-groundbreaking-nlp-framework/
7. https://towardsdatascience.com/how-to-train-a-bert-model-from-scratch-72cfce554fc6
8. https://nlp.stanford.edu/seminar/details/jdevlin.pdf