Recep Oğuz Araz
ELEC390
Prof. Engin Erzin
16.11.2020

Report IV: Data Preperation

Batch Processing

Gradient Descent: Make predictions using current model parameters and calculate error. Update model parameters that will minimize the error by calculating the gradient.

1. Stochastic GD: For each example, calculate the error, update parameters.
2. Batch GD: Calculate error for each example, update after all examples evaluated.
3. Mini-Batch GD: Split training set into mini-batches of size 2^N and apply Batch GD over each.

Parallel Computing: Make use of array multiplications of Python libraries such as Numpy, PyTorch rather than process each input one-by-one.

- Pros: Much faster computation.
- Cons: Parallel data generation requires data augmentation. (CTC may take care of it)

DS2 ➔ Mini-Batch GD, with batch size 16 where all the mini batch is processed at once using parallel computing to exploit GPU structure.

Mozilla Common Voice Dataset, Validated Subset

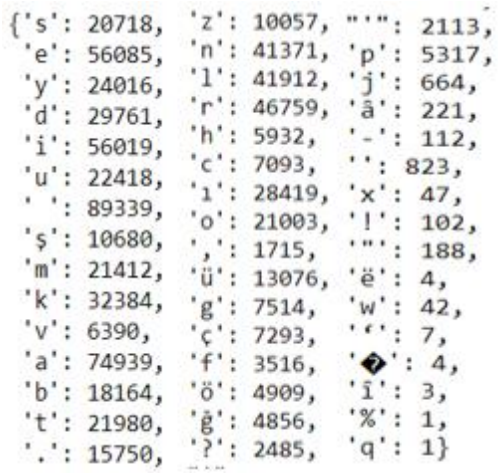
	client_id	path	sentence	up_votes	down_votes	age	gender	accent	lo
0	0c8ba63665303a01117332d5c7de7438ee9e9f5d523530...	common_voice_tr_17343923.mp3	Seydiu şimdi iki mevkiyi de kaybetti.	2	1	teens	male	other	
1	236a01cbb9dc681398aa97c5b5b25df6ab1528757eaa2b...	common_voice_tr_19942053.mp3	Siyasette temiz insanlara ihtiyacımız var.	2	0	NaN	NaN	NaN	
2	38bfb2e08d0f099ba084c2e2bdf4d6beb880e2f2207abd...	common_voice_tr_17354593.mp3	Broz, büyükbabasının başarılarını gururla anıyor.	2	0	NaN	NaN	NaN	
3	49054ba389e2c960c1bb50387c51f32762abfe3e19e0ce...	common_voice_tr_19942151.mp3	Calasan yetenekli çocuk bursu da aldı.	2	0	NaN	NaN	NaN	
4	4a02ad68cf34bc9ac9135432c0070013d02f3903102af1...	common_voice_tr_19821495.mp3	Referandum tarihi henüz belirlenmedi.	2	0	twenties	female	NaN	
...	...	...	...	...	...	...	...	...	
18539	e6ade7149cdb6a5370585393f6db44556d5a9934ad2767...	common_voice_tr_17554012.mp3	Fakat uzmanlar, buna değeceğini söylüyorlar.	2	0	thirties	male	NaN	
18540	e6ade7149cdb6a5370585393f6db44556d5a9934ad2767...	common_voice_tr_17554026.mp3	Sırplar Pekin'de neden daha iyisini yapamadılar?	2	0	thirties	male	NaN	
18541	e6ade7149cdb6a5370585393f6db44556d5a9934ad2767...	common_voice_tr_17554029.mp3	Bundan sonra bir şeylerin değişmesi gerekecek.	2	0	thirties	male	NaN	
18542	e6ade7149cdb6a5370585393f6db44556d5a9934ad2767...	common_voice_tr_17554027.mp3	Slovenya iki bin sekiz yılında Kosova'yı tanım...	2	0	thirties	male	NaN	
18543	e6ade7149cdb6a5370585393f6db44556d5a9934ad2767...	common_voice_tr_17554030.mp3	Ancak Gruevski iki öneriyi geri çevirdi.	2	0	thirties	male	NaN	

18544 rows × 10 columns

Recep Oğuz Araz
ELEC390
Prof. Engin Erzin
16.11.2020

Symbol Analysis

After lower caseing each symbol in the whole dataset, we get the symbol dictionary.



```
{'s': 20718,  'z': 10057,  "'": 2113,
 'e': 56085,  'n': 41371,  'p': 5317,
 'y': 24016,  'l': 41912,  'j': 664,
 'd': 29761,  'r': 46759,  'â': 221,
 'i': 56019,  'h': 5932,   '-': 112,
 'u': 22418,  'c': 7093,   '': 823,
 ' ': 89339,  'ı': 28419,  'x': 47,
 'ş': 10680,  'o': 21003,  '!': 102,
 'm': 21412,  ',': 1715,   '"': 188,
 'k': 32384,  'ü': 13076,  'ë': 4,
 'v': 6390,   'g': 7514,   'w': 42,
 'a': 74939,  'ç': 7293,   '‘': 7,
 'b': 18164,  'f': 3516,   '': 4,
 't': 21980,  'ö': 4909,   'î': 3,
 '.': 15750,  'ğ': 4856,   '%': 1,
              '?': 2485,   'q': 1}
```

Problematic symbols:

Punctuations (removed for now)

- “.” <EOS> is not necessary for ASR ?
- “ ’ ” In Turkish written-read differs only when “ ’ ”.
- “-“ has no sound but it affects intonation
  - **Ex**: Bulgar takımı kıbrıs'ı üç-sıfır yenerek birinci oldu.
- “,” “?” has no sound but it affects intonation.
- “!” no sound but affects intonation
  - **Ex:** yine de girişimlerinden ötürü kutluyoruz!

Turkish non-standard letters

- Şapkalı a,i : pronunciation different than regular a.
  - **Ex:** AB, bu tarihî durumla baş edebilir mi?

Foreign letters

- “x”, ” ë”, “w”, “q”
  - **Ex:** xhema ise mabetex'in yönetim kurulu uyesi.

Word Holders

- “ % ” = “yüzde”

Unkown Symbols

- **:** Only appears 4 time in dataset. Replaces % by listening the examples. Excluded.

How does CTC deal with “ ” ? Blank != “ ” (Next Week)

Recep Oğuz Araz
ELEC390
Prof. Engin Erzin
16.11.2020

If we remove sentences with ['','%','x','ë','w','q',’X’,’Q’,’W’] ➔ 95 sentences discarded.

Replaced şapkalı harfler with their regular versions.

Removed each punctuation for simplicity.

30 symbols remaining, including boşluk and 29 Turkish letters. Blank symbol not included?

More than 10 hours of recording remain (Probably 15-20). Some recordings are different utterances of the same sentences.

Some Problematic Sentences

bosna-hersek'te kampanya dönemi sona ererken medya kuruluşları bunu adil şekilde yayınlıyorlar mı?

firefox

bilim insanı "bunu neden değiştirmek gerekiyor ki?" diye soruyor.

vetëvendosje liderinin tutuklanması hareketin siyasete katılması için yakıt sağladı mı?

operatör huawei ile beş yıllık sözleşme imzaladı.

priştine-belgrad diyaloğu kurtarılabilir mi?

roddick maçı altı-bir ve altı-dörtlük setlerle kazandı.

tesis kırk megawattlık kapasiteye sahip olacak.

her bir firma günde iki virgül otuz iki kwh elektrik sağlayacak.

merwin festivale üçüncü kez katılıyor.

milutin "bu tür bir filmin gösterime girmesi mümkün mü?" diye soruyor blogunda.

ikinci ankete göre bu oran ''den bile yüksek. şimdiye kadar fonların yaklaşık ''i sağlandı.

bojaxhi'ye göre çeşitli yöntemler izlenmeli.

hükümet şirketin %'ini elinde tutmak istiyor.

opa!

yazar böyle bir zihniyetle "aldığımız madalyalar bile büyük ikramiye!" diyor.

kotooshu unvanı on üç-birlik skorla kazandı.

dementieva maçı yedi-altı ve üç-sıfır önde götürüyordu.

moldovan-tilea ciriti elli üç virgül dört metreye attı.

zagrep lego'ları sevdi!

morrisey "sonunda birileri şarkılarımın sözlerini biliyor!" diye bağırdı.