PREPIS HOVORENEJ REČI, PODPORA SLOVENSKÉHO JAZYKA

Samuel Baran

RNDr. Erik Bruoth, PhD.

Abstrakt

Prepis hovorenej reči do je vo svetových jazykoch značne rozšírený a je bežnou súčasťou mnohých moderných technológií. Jazyky, ako je napr. anglický jazyk, sú zvýhodnené oproti minoritným jazykom, kde patrí aj slovenský jazyk, v dostupnosti vstupných dát určených na tréning v podobe rôznych zvukových nahrávok a im prislúchajúcich prepisov. To umožňuje spoľahlivé trénovanie modelov STT a tým pádom aj rýchlejšiu a efektívnejšiu implementáciu STT v praxi. Keďže táto oblasť poskytuje nedostatočnú podporu slovenského jazyka, rozhodli sme sa danú problematiku preskúmať a hľadať spôsob trénovania modelov rozpoznávania reči pre slovenský jazyk.

Rozpoznávanie reči

Rozpoznávanie reči, tiež známe ako Automatic speech recognition (ASR) alebo Speech to text (STT), je jednou z mnohých podoblastí vedného odboru NLP (Natural language processing) zaoberajúceho sa výskumom spracovania textovej alebo zvukovej podoby prirodzeného jazyka za pomoci informačných technológií. Úlohou samotného ASR je vyvíjať techniky a systémy umožňujúce interakciu medzi používateľom a počítačom prostredníctvom zvuku.

DeepSpeech

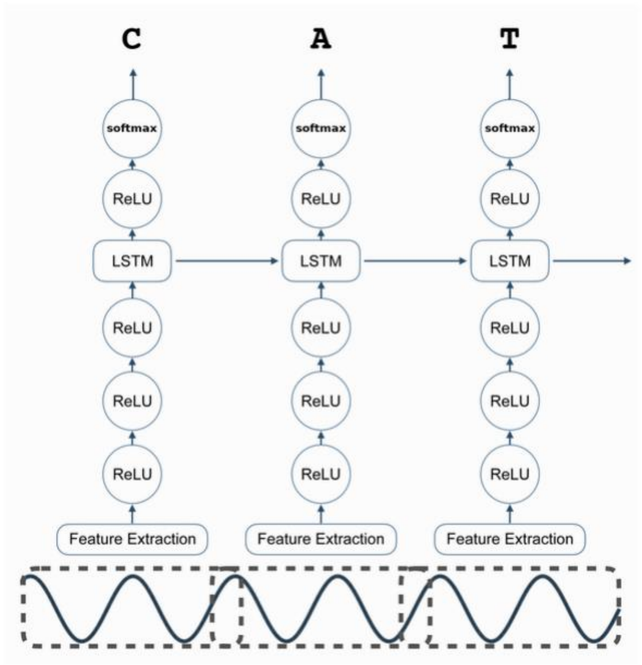
DeepSpeech je open source nástroj rozpoznávania reči vyvinutý firmou Mozilla. Implementácia tohto nástroja využíva model trénovaný pomocou techník strojového učenia, inšpirovaný výskumom Deep Speech: Scaling up end-to-end speech recognition spoločnosti Baidu Research - Silicon Valley AI Lab.

Autori výskumu predstavujú end-to-end systém rozpoznávania reči nazývaný DeepSpeech, ktorý komplikované metódy nahrádza síce robustnou, ale jednoduchou rekurentnou neurónovou sieťou(RNN). Kombináciou s jazykovým modelom tento nástroj dosahuje na náročných úlohách rozpoznávania reči lepšie výsledky ako tradičné metódy, za čo vďačí paralelnému tréningu veľkej neurónovej siete na viacerých jednotkách GPU s použitím tisícok hodín dát.

Architektúra systému

Jadrom tohto nástroja je rekurentná neurónová sieť, trénovaná na generovanie textovej transkripcie vstupných zvukových nahrávok.

Zo zvukového signálu sa extrahujú MFCC vektory, ktoré sú následne doplnené o širší kontext. Takto rozšírené vektory, sú vstupom rekurentnej neurónovej siete pozostávajúcej z piatich skrytých vrstiev. Prvé tri nerekurentné vrstvy pracujú s dátami nezávisle od poradia v akom sa vyskytovali v pôvodnom zázname reči, štvrtá vrstva je rekurentná LSTM vrstva a piata nerekurentná vrstva počíta hodnoty, ktoré korešpondujú s distribúciou pravdepodobností naprieč znakmi abecedy. Na určenie chyby je využitá CTC loss funkcia vyvinutá špeciálne na úlohy klasifikácie na časových radoch.

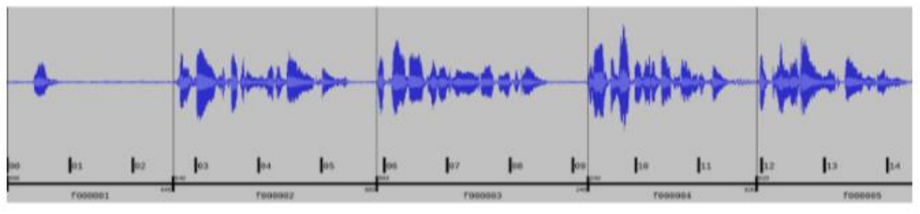


Obrázok č. 1: Ukážka architektúry modelu DeepSpeech.

Príprava datasetu

Jednou z najdôležitejších prvkov systémov využívajúcich neurónové siete je množstvo kvalitných dát, ktorých príprava vyžaduje značné úsilie a zaberá pomerne veľkú časť vývoja a trénovania modelov. Väčšina známych problémov umelej inteligencie má voľne prístupné datasety určené práve na tréning a ohodnocovanie efektivity riešení daných úloh. Ak by sme sa však zamerali na menej rozšírené úlohy, zistíme, že dostupnosť datasetov nie je pravidlom, čo je aj náš prípad. Dáta sme si preto museli pripraviť. Ako zdroj sme využili audioknihy a k nim kompatibilné e-knihy, ktoré sme segmentovali na kratšie nahrávky pomocou implementácie forced alignmentu s názvom aeneas.

```
1                                                  => [00:00:00.000, 00:00:02.640]
From fairest creatures we desire increase,         => [00:00:02.640, 00:00:05.880]
That thereby beauty's rose might never die,        => [00:00:05.880, 00:00:09.240]
But as the riper should by time decease,           => [00:00:09.240, 00:00:11.920]
His tender heir might bear his memory:             => [00:00:11.920, 00:00:15.280]
But thou contracted to thine own bright eyes,      => [00:00:15.280, 00:00:18.800]
Feed'st thy light's flame with self-substantial fuel, => [00:00:18.800, 00:00:22.760]
Making a famine where abundance lies,              => [00:00:22.760, 00:00:25.680]
Thy self thy foe, to thy sweet self too cruel:     => [00:00:25.680, 00:00:31.240]
Thou that art now the world's fresh ornament,      => [00:00:31.240, 00:00:34.400]
And only herald to the gaudy spring,               => [00:00:34.400, 00:00:36.920]
Within thine own bud buriest thy content,          => [00:00:36.920, 00:00:40.640]
And tender churl mak'st waste in niggarding:       => [00:00:40.640, 00:00:43.640]
Pity the world, or else this glutton be,           => [00:00:43.640, 00:00:48.080]
To eat the world's due, by the grave and thee.     => [00:00:48.080, 00:00:53.240]
```



Obrázok č. 2: Grafické znázornenie výstupov metódy Forced alignment. Každá časť prepisu nahrávky má priradený časový interval, v ktorom sa v rámci nahrávky nachádza reč prislúchajúca danému prepisu. Hranice fragmentov sú znázornené zvislými čiarami.

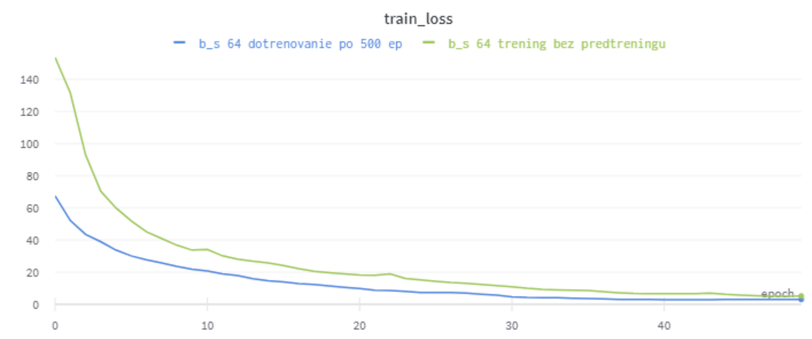
Nekontrolovaný predtréning

V minulosti zohrával predtréning bez učiteľa, tiež známy ako unsupervised predtréning, významnú úlohu v zlepšovaní metód strojového učenia. Príchodom vylepšení sa začalo upúšťať od metód unsupervised predtréningu, no ten si časom našiel uplatnenie v mnohých technológiách spracovania prirodzeného jazyka.

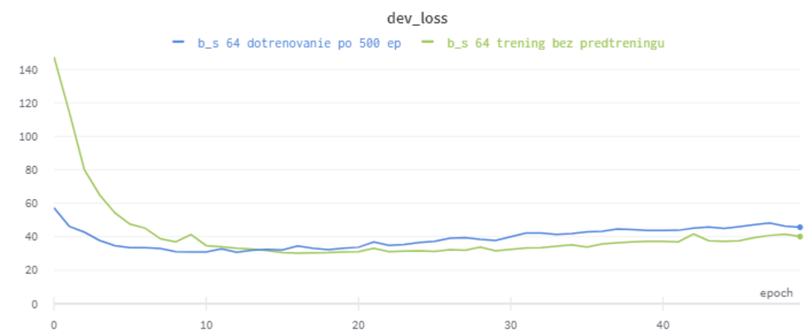
Pri navrhovaní modelu na predtrénovanie sme sa inšpirovali výskumom Generative pretraining from pixels, ktorý pomocou reprezentácií obrázkov zapamätaných pomocou predtréningu dosiahol aplikovaním klasifikačnej vrstvy v určitej časti modelu presnosť presahujúcu mnohé klasifikačné neurónové siete trénované s učiteľom.

Výsledky tréningov

Pri porovnaní tréningu modelu DeepSpeech bez úprav a kontrolovaného dotrénovania modifikovaného modelu predtrénovaného učením bez učiteľa, sme pozorovali zlepšené východiskové hodnoty trénovacej a validačnej stratovej funkcie predtrénovaného modelu a porovnateľné výsledky testovacej stratovej funkcie.
Predtrénovaný model dosiahol minimálnu hodnotu stratovej funkcie skôr ako neupravený model, no následne došlo k preučeniu daného modelu, čo mohol spôsobiť nedostatok dát určených na predtréning.



Obrázok č. 3: Porovnanie trénovacej stratovej funkcie.



Obrázok č. 4: Porovnanie validačnej stratovej funkcie.

Literatúra

- Hannun, A.a kol., 2014.Deep speech: Scaling up end-to-end speech recognition:výskumná práca [online][cit. 2021-04-12]. Dostupné na:https://arxiv.org/abs/1412.5567
- Mozilla Corporation, 2020.DeepSpeech's documentation: DeepSpeech Mo-del [online][cit. 2021-04-15]. Dostupné na:https://deepspeech.readthedocs.io/en/r0.9/DeepSpeech.html
- Chen, M.et al., 2020. Generative pretraining from pixels. Proceedings of the37th International Conference on Machine Learning. In:International Conferenceon Machine Learning[online]. Vol. 119, p. 1691–1703 [cit. 2021-04-25]. ISSN 2640-3498. Dostupné na: http://proceedings.mlr.press/v119/chen20s/chen20s.pdf