FAKULTA APLIKOVANÝCH VĚD
ZÁPADOČESKÉ UNIVERZITY
V PLZNI

Semestrální práce z předmětu
Hlasové a dialogové systémy

Tvorba neurálního modelu řeči

Autoři:
Daniel Cífka
A23N0100P
dcifka20@students.zcu.cz

Stanislav Kafara
A24N0088P
skafara@students.zcu.cz

Obsah

Kapitola 1

Zadání

Na základě dodaných (nebo vlastních) nahrávek a odpovídajících textových přepisů natrénujte neurální model řeči. Podrobný návod na trénování je součástí zadání na Google Classroom. Popis používaných ztrátových funkcí a návod na jejich konstrukci je uveden zde.

Vstup

- řečové nahrávky
  - dodaný "erární" hlas: `/storage/plzen4-ntis/projects/korpusy-public/vyuka/HDS2024/SP1/FulTo.cs-CZ.m/wavs`
  - v případě použití vlastního hlasu (obdržíte odkaz na stažení pomocí služby FileSender);
- textové přepisy nahrávek
  - dodaný "erární" hlas: `/storage/plzen4-ntis/projects/korpusy-public/vyuka/HDS2024/SP1/FulTo.cs-CZ.m/sentences_phnfriendly.csv`
  - vlastní hlas: `sentences_phnfriendly.csv` k dispozici na Classroomu
- textový soubor testovacími větami `vety_HDS.ortho.txt`, tj. soubor, pro který musíte v rámci SP0 vytvořit jeho fonetický přepis (upravený podle návodu, aby byl kompatibilní s vytvořeným neurálním modelem vašeho hlasu).

Výstup

- vysyntetizované promluvy odpovídající textům v souboru `vety_HDS.ortho.txt`

Práci odevzdávejte pomocí Google Classroom, natrénovaný model (má obvykle velikost cca 1 GB) posílejte pomocí služby FileSender na adresu dtihelka@ntis.zcu.cz.

Součástí odevzdání je i stručný referát popisující postup trénování modelu a syntézy textů. Uveďte zejména jakékoliv změny v trénovacím procesu oproti návodu (např. počet trénovacích kroků).

Vygenerované promluvy si poslechněte a kriticky zhodnoťte schopnosti vašeho modelu, zaměřte se především na srozumitelnost (případné problémy s výslovností některých hlásek v určitých kontextech) a přirozenost (plynulost mluvy, adekvátní frázování a jeho prozodická realizace, vkládání pauz apod.).

Kapitola 2

Trénování modelu

Pro účely této semestrální práce byl využit dodaný TTS model VITS. Tento model jsme fine-tunovali na dodaném „erárním“ hlasu, jelikož jsme pro tuto práci vlastní hlas nenahráli. Model byl učen na dodaných nahrávkách a jejich textový přepis jsme transkribovali do fonetické podoby pomocí vlastního skriptu, odevzdaného jako součást semestrální práce č. 0.

Model byl trénován celkem na třech konfiguracích trénování, na výchozí konfiguraci a dvou dalších, jež jsou z ní odvozené. Tyto konfigurace lze vidět v Tabulce 2.1. Model byl učen po dobu 1000 epoch.

Parametr	**Konfig. #1**	**Konfig. #2**	**Konfig. #3**
`lr`	0.0005	0.0005	0.0005
`lr_gen`	0.0001	0.0001	0.0001
`lr_disc`	0.0001	0.0001	0.0001
`kl_loss_alpha`	5.0	5.0	5.0
`disc_loss_alpha`	1.0	1.0	1.0
`gen_loss_alpha`	1.0	1.0	1.0
`feat_loss_alpha`	1.0	2.0	5.0
`mel_loss_alpha`	45.0	50.0	55.0
`dur_loss_alpha`	1.0	2.0	3.0

Tabulka 2.1: Konfigurace trénování

Průběh ztrátových funkcí pro konfiguraci trénování #1 lze vidět na Obrázku A. Trénink modelu probíhá relativně stabilně. Většina ztrát má oscilující, ale ustálený průběh, což je u GAN architektur běžné. Např. `avg_loss_kl` a `avg_loss_feat` ale vykazují postupný růst, což může naznačovat určité problémy trénování modelu.

Kombinované ztrátové funkce `avg_loss_0` diskriminátoru a `avg_loss_1` generátoru se pohybují ve stabilním rozsahu. `avg_loss_0` osciluje mezi hod-

notami 1,25 až 1,5 a vykazuje mírně klesající trend, zatímco `avg_loss_1` zůstává stabilní kolem hodnoty 39 až 43 bez výrazného zhoršení.

Ztrátová funkce diskriminátoru má klesající trend, poté osciluje kolem nízké hodnoty. Diskriminátor je tedy schopen stabilně dobře rozpoznávat reálná data.

Ztráta spojená s predikcí délky fonémů `avg_loss_dur` prudce klesá už na začátku trénování a rychle se ustaluje kolem hodnoty 0,4. Tento průběh naznačuje, že model se naučil délku fonémů dobře predikovat a že tento aspekt generace je stabilní.

Ztráta typu feature matching `avg_loss_feat` osciluje v rozmezí od 10,2 do 11,8. Je zde patrný mírně rostoucí trend, který může naznačovat, že generátor ztrácí krok s diskriminátorem a dochází k mírné nevyváženosti.

Ztráta generátoru `avg_loss_gen` osciluje kolem hodnoty 3,7 až 4,1, bez zjevných známek zhoršení. Tento průběh je typický pro GAN architektury, kde se generátor a diskriminátor navzájem dynamicky přizpůsobují.

Signálem k pozornosti může být průběh KL divergence `avg_loss_kl`, která od hodnot kolem 2 trvale roste až nad 4. To je nežádoucí a může značit problém s učením latentního prostoru, kdy se distribuční aproximace postupně vzdaluje od skutečné distribuce.

Mel-spektrogramová ztráta `avg_loss_mel` plynule klesá, což značí, že model se postupně zlepšuje ve schopnosti generovat řeč, jejíž spektrální vlastnosti se stále více blíží skutečným vzorkům.

Kapitola 3

Hodnocení modelu

Syntetizované řeči lze rozumět a je relativně kvalitní. Kvalita je avšak omezena námi nedokonale implementovanou fonetickou transkripcí, odevzdanou jako součást semestrální práce č. 0. Syntetizovaná řeč je velmi monotónní a mírně robotická. Nejlepších výsledků je dosaženo při použití výchozí konfigurace trénování. Při ostatních konfiguracích nebylo pozorováno výrazné zlepšení v kvalitě výstupu.

Při poslechu syntetizované řeči jsme pozorovali některé nevyžádané vlastnosti. Řečník má relativně dlouhou pauzu mezi věty, oddělené čárkami, kde by měla být jen „krátká pauza“. Konečná část promluvy je obvykle zatížena chybou, což většinou způsobí nesprávné vyslovení posledního písmena promlouvaného textu nebo podivný zvuk po ukončení promluvy. Někdy se stane, že při vyslovování „r“ řečník „ráčkuje“, obvykle je to ale bez problému. Slovo „ale“ je často vyslovováno podobně jako slovo „a“.

Na základě těchto pozorování byly navrhnuty další dvě konfigurace trénování, kde jsme především zvyšovali důležitost odhadu trvání fonémů, přesnost syntézy ve smyslu mel-spektrogramů a věrohodnost syntézy ve smyslu feature matching generátoru. Tyto konfigurace jsou uvedeny v Tabulce 2.1.

Kapitola 4

Závěr

Během vypracovávání této práce jsme se setkali s několika problémy. Domníváme se, že dodané vstupní soubory s přepisem nahrávek nemají obsahovat znak „^“, což může způsobit problémy při trénování. Další problémy, s kterými jsme se potýkali, se týkají MetaCentra. Na přidělených strojích nějakého typu nefungovalo trénování v interaktivním režimu z nejasného důvodu. Dále jednomu z nás nefungovalo trénování v dávkovém režimu, přičemž druhému z nás toto fungovalo bez problémů (bez provedení jiných změn než změn příslušných cest).

V této semestrální práci byly implementovány všechny požadované ztrátové funkce. Byly natrénovány celkem tři modely, z toho jeden s výchozí konfigurací trénování a další dva s konfiguracemi vycházející z pozorování výsledné syntetizované řeči modelu s výchozí konfigurací trénování. Nejlepší natrénovaný model byl kriticky zhodnocen. Veškeré požadavky práce byly splněny.

Příloha A

Průběh ztrátových funkcí