Západočeská univerzita v Plzni
Fakulta aplikovaných věd
Katedra informatiky a výpočetní techniky

Bakalářská práce

Integrovaný systém pro automatizaci záznamu, úprav a publikaci přednášek

Plzeň, 2024 Martin Reich

Obsah

1 Zadání

Hlavní cíle:

1. Seznamte se současnými metodami a možnostmi převodu řeči do textu a zpět a metodami parafrázování.
2. Seznamte se současnými systémy, nástroji a knihovnami pro transkripci řeči na text, převod textu na řeč a parafrázování.
3. Navrhněte integrovaný systém, který dokáže automatizovat převod zvukového a obrazového záznamu přednášky v češtině či angličtině do textu, tento následně parafrázovat a zpětně převést do zvukové a obrazové podoby přednášky.
4. Navržený systém dle bodu 3 implementujte.
5. Ověřte výsledné řešení na dostatečném počtu přednášek v českém i anglickém jazyce.

2 Převod řeči do textu

Převedení řeči do textu (Speech-to-Text, STT) je oblast, která se zabývá automatickým rozpoznáváním mluveného slova a jeho převodem na psaný text. Je to multidisciplinární obor zahrnující akustiku, jazykovědu, statistiku a strojové učení.

2.1 Jak to funguje

Převod řeči na text je software, který funguje tak, že naslouchá zvuku a poskytuje upravitelný, doslovný přepis na daném zařízení. Software to provádí prostřednictvím rozpoznávání hlasu. Počítačový program využívá jazykové algoritmy k třídění zvukových signálů z mluvených slov a převádí tyto signály na text pomocí znaků zvaných Unicode. Převod řeči na text funguje prostřednictvím složitého modelu strojového učení, který zahrnuje několik kroků.

1. Zvuk vycházejí z úst k vytvoření slova, také vytváří sérii vibrací. Technologie převodu řeči na text funguje tak, že zachycuje tyto vibrace a překládá je do digitálního jazyka prostřednictvím analogově-digitálního převodníku.
2. Analogově-digitální převodník přijme zvuk z audio souboru, podrobně zemří jeho vlny a provede na nim filtraci, aby odlišil relevantní zvuky.
3. Zvuk je poté segmentován na setiny nebo tisíciny sekundy a poté je přiřazeny k fonémům. Foném je jednotka zvuku, která odlišuje jedno slovo od druhého v jakémkoli daném jazyce.
4. Fonémy jsou poté zpracovány sítí prostřednictvím matematického modelu, který je porovnává s dobře známými větami, slovy a frázemi.
5. Text je poté prezentován jako text nebo jako počítačový požadavek na základě nejpravděpodobnější verze audia.

Dodat obrázek digital-analog převodník.

2.2 Metody

Současné metody, které se používají k převodu reči na text.

2.2.1 Umělé neuronové sítě

Neuronová síť je model strojového učení, který simuluje rozhodování lidského mozku. Používá procesy, jež napodobují způsob, jakým biologické neurony spolupracují, aby identifikovaly jevy, zhodnotily možnosti a dospěly k závěrům. Každá neuronová síť se skládá z vrstev uzlů, které fungují jako umělé neurony. Tyto vrstvy zahrnují vstupní vrstvu, jednu nebo více skrytých vrstev a výstupní vrstvu. Každý uzel je propojen s ostatními uzly a má svou vlastní váhu a prahovou hodnotu. Pokud výstup některého uzlu překročí stanovený práh, tento uzel se aktivuje a předá data do další vrstvy sítě. Jinak nedochází k žádnému předání dat. Neuronové sítě potřebují tréninková data, aby se učily a zvyšovaly svou přesnost. Jakmile jsou optimalizovány, stávají se silnými nástroji v oblasti informatiky a umělé inteligence, což nám umožňuje rychle klasifikovat a shlukovat data. Úkoly, jako je rozpoznávání řeči nebo obrazů, mohou trvat minuty namísto hodin. Výhody:

- Flexibilita a schopnost učení: Neuronové sítě se dokáží učit z tréninkových dat a zlepšovat svou přesnost, což je činí adaptabilními na různé úkoly a scénáře.
- Schopnost zpracovávat složité vzory: Neuronové sítě jsou schopné identifikovat složité vzory a vztahy v datech, které by tradiční algoritmy mohly přehlédnout.
- Všestrannost: Můžou být aplikovány v různých oblastech, včetně zpracování obrazů, přirozeného jazyka, doporučovacích systémů, medicíny a mnoha dalších.

Nevýhody:

- Požadavky na tréninková data: Neuronové sítě potřebují velké množství kvalitních tréninkových dat, aby byly schopny efektivně se učit a generalizovat na nová data.
- Trénink neuronových sítí může být časově náročný a vyžaduje značné výpočetní zdroje, zejména při práci s velkými datovými sadami nebo složitými architekturami.
- Riziko přetrénování: Pokud jsou neuronové sítě trénovány na příliš malých nebo nevyvážených datových sadách, mohou se "naučit"specifické

vzory v tréninkových datech a mít problémy s generalizací na nová, neznámá data.

Dodat obrázek neuronové sítě

2.2.2 Hluboké učení

Hluboké učení je podmnožina strojového učení, která využívá více vrstvé neuronové sítě, nazývané hluboké neuronové sítě, k simulaci složitého rozhodovacího procesu lidského mozku. Hlavní rozdíl mezi hlubokým učením a strojovým učením spočívá ve struktuře architektury základní neuronové sítě. Tradiční modely strojového učení, které nejsou hluboké, používají jednoduché neuronové sítě s jednou nebo dvěma výpočetními vrstvami. Naopak modely hlubokého učení mají tři a více vrstev, často stovky nebo dokonce tisíce vrstev. Zatímco modely s učením s učitelem vyžadují strukturovaná a označená vstupní data pro dosažení přesných výstupů, modely hlubokého učení mohou pracovat s učením bez učitele. Díky učení bez učitele dokážou modely hlubokého učení extrahovat charakteristiky, vlastnosti a vztahy, které potřebují k dosažení přesných výstupů z neupravených a neorganizovaných dat. Navíc tyto modely mohou vyhodnocovat a zdokonalovat své výstupy pro zvýšení přesnosti.

2.2.3 Rekurentní neuronové sítě

Rekurentní neuronové sítě (RNN) jsou modely hlubokého učení a jsou schopné zpracovávat sekvence dat tím, že mají vnitřní paměť, která uchovává informace o předchozích vstupech. Tato vlastnost jim umožňuje modelovat časové závislosti a vztahy v datech. Na rozdíl od tradičních neuronových sítí, které zpracovávají vstupy nezávisle, RNN zohledňují kontext a historii dat. Výhody:

- Zpracování sekvenčních dat: RNN jsou ideální pro úkoly, kde je pořadí dat důležité.
- Flexibilita: Mohou zpracovávat vstupy a výstupy různé délky (např. různé délky vět).

Nevýhody:

- Vanishing gradient: Při trénování mohou gradienty zmizet, což brání učení dlouhodobých závislostí.
- Vysoké výpočetní nároky: RNN mohou být náročné na výpočetní výkon, zejména při práci s dlouhými sekvencemi.

Dodat obrázek rekuretní neuronové sítě

2.2.4 Transformer

Transformery jsou modely hlubokého učení a používají mechanismus vlastní pozornosti, který umožňuje modelu vážit různá slova v sekvenci na základě jejich relevance. To znamená, že model může posoudit, která slova mají vliv na ostatní slova, což je klíčové pro porozumění kontextu. Jelikož transformery nemají vnitřní strukturu pro zpracování sekvencí (na rozdíl od RNN 2.2.3), používají se k nim pozicové kódování, aby modely mohly rozpoznat pořadí slov v sekvenci. Tato kódování přidávají k vektorům slov informace o jejich pozici v sekvenci. Výhody:

- Paralelizace: Transformery umožňují paralelní zpracování vstupních dat, což urychluje trénink ve srovnání s RNN.
- Zachycení dlouhodobých závislostí: Díky mechanismu pozornosti jsou schopny efektivně sledovat vztahy mezi slovy na delší vzdálenosti.
- Flexibilita: Lze je snadno aplikovat na různé úkoly a adaptovat je pro konkrétní aplikace.

Nevýhody:

- Vysoké nároky na paměť: Vzhledem k tomu, že transformery zpracovávají celou sekvenci najednou, mohou mít vysoké požadavky na paměť, zejména při práci s dlouhými sekvencemi.
- Potřeba velkých dat: K dosažení dobrého výkonu vyžadují transformery velké množství tréninkových dat.

2.2.5 Konvoluční neuronové sítě

Konvoluční neuronové sítě (CNN) jsou specifickým typem neuronových sítí, které se ukázaly jako velmi efektivní při analýze vizuálních dat, jako jsou obrázky a videa. Tyto sítě se široce používají v oblastech, jako je počítačové vidění, rozpoznávání obrazů, analýza videí a dokonce i v zpracování přirozeného jazyka. CNN má tři hlavní typy vrstev.

- Konvoluční vrstva: Tato vrstva aplikuje konvoluční operaci na vstupní data pomocí malých filtrů(filtr pokrývá jen část vstupniách dat). Hodnoty(váhy) ve filtru se vynásobí s hodnotami na vstupních datech, preš který je filtr "umístěn". Filtr se poté přesune, aby vytvořil reprezentaci nové části vstupních dat, a proces se opakuje, dokud nejsou pokrytá celá vstupní data.

- Pooling vrstva: Podobně jako u konvoluční vsrtvy, tato vrstva aplikuje filtry na vstupní data. Jediný rozdíl je, že filtr nemá žádné váhy a ze vstupních dat se vezme hodnotu se největší nebo průměrnou hodnotou.
- Plně propojená vrstva: Na konci CNN se obvykle nachází jedna nebo více plně propojených vrstev, které kombinují výstupy z předchozích vrstev a předpovídají finální výstup.

Výhody:

- Efektivita: CNN dokáží zpracovávat velké množství dat s menšími výpočetními nároky díky sdílení vah a redukci rozměrů pomocí pooling vrstev.
- Vysoká přesnost: Tyto sítě dosahují vynikajících výsledků v úlohách klasifikace obrazů a detekce objektů.

Nevýhody:

- Potřeba velkého množství dat: Pro úspěšné trénování vyžadují CNN velké množství označených tréninkových dat.
- Omezené schopnosti na neobvyklých datech: Modely CNN mohou mít potíže s generalizací na atypická data, která se výrazně liší od tréninkových dat.

3 Zkratky

STT	text-to-speech
RNN	rekuretní neuronová síť
CNN	konvoluční neuronová sít

Tabulka 1: Tabulka zkratek

4 Bibliografie

Reference

[1] Jim Holdsworth and Mark Scapicchio (2024), *Deep learning*, IBM, `https://www.ibm.com/topics/deep-learning`.

[2] *Neural networks*, IBM, `https://www.ibm.com/topics/neural-networks`.

[3] *Transformer model*, IBM, https://www.ibm.com/topics/transformer-model.

[4] *Recurrent neural network*, Wikipedia, https://en.wikipedia.org/wiki/Recurrent_neural_network.