

# Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

**Autor práce** Bc. Štěpán Procházka  
**Název práce** Adaptive Handwritten Text Recognition  
**Rok odevzdání** 2021  
**Studijní program** Informatika      **Studijní obor** Umělá inteligence

**Autor posudku** Milan Straka      **Role** Vedoucí  
**Pracoviště** Ústav formální a aplikované lingvistiky

## Text posudku:

Cílem diplomové práce bylo navrhnout systém, který by dokázal převést ručně psaný text do digitální podoby, se zaměřením na současné (20. a 21. století) české vázané písmo. Největší překážkou byla neexistence datasetů v této oblasti, a tak bylo hlavní zaměření práce na technikách, které by umožnily trénování modelů s žádným nebo velmi malým množstvím ručně anotovaných dat.

Celý proces rozpoznání ručně psaného textu má několik fází (skenování a jeho digitální úprava, detekce rozložení a řádek, extrakce textu z řádek). Diplomová práce se zabývá především poslední z těchto úloh, tj. extrakcí textu z řádek, ale poskytuje základní řešení i pro zbývající fáze a je použitelná na skeny dokumentů skládajících se z jednoho bloku řádek. Zároveň je pokročilejší analýza rozložení v podstatě nezávislá na extrakci textu, takže výsledky diplomové práce je možné aplikovat i v případě dokumentů s komplikovanějším rozložením.

Zvolené řešení pochopitelně staví na technikách hlubokého učení, které je v oblasti rozpoznání textu dominantním přístupem. První kapitola diplomové práce popisuje algoritmy používané specificky v oblasti OCR, především CTC (Connectionist Temporal Classification), což je trénovací a inferenční algoritmus, který při trénování nevyžaduje zarovnání (přiřazení) mezi obrázky řádek a extrahovanými znaky, a při inferenci umožňuje použití externího jazykového modelu. Popis obou variant tohoto algoritmu je velmi precizní a tvoří tak jednoznačně první příspěvek diplomové práce.

Druhá kapitola popisuje navržený generátor syntetických dat, který produkuje obrázky ručně psaného textu pomocí syntetických fontů, v různých náhodně zvolených stylech (tloušťka, sklon, rotace, velikost). Výsledná implementace je modulární a efektivní.

Třetí kapitola práce popisuje zvolenou architekturu neuronové sítě (což je standardní CNN-LSTM architektura s CTC loss), C++ implementaci CTC dekodéru používající originální kombinaci slovního a znakového jazykového modelu, a předtrénování modelu na syntetických datech.

Trénink a vyhodnocení na skutečných datech rozebírá práce v kapitole čtvrté. Řešitel v práci vytvořil dva české datasety, jejichž skeny převzal z Archivu hlavního města Prahy – školní kroniku z let 1913-1933 obsahující 6.5 tisíce řádek textu, a městskou kroniku z let 1931-1944 obsahující 18 tisíc řádek. Tyto datasety rozdělil řešitel rozdělené na jednotlivé řádky a ručně přepsal prvních 1024 řádek z každého datasetu. Vzniklé datasety jsou veřejně přístupné. Kromě českých datasetů práce používá také anglický dataset Washington pro porovnání s existujícími metodami – konkrétně model předtrénovaný na syntetických českých a anglických datech dotrénovaný na trénovací části datasetu Washington podstatně překonává nejlepší známé výsledky, což demonstruje efektivitu zvolené architektury a tréninku na syntetických i reálných datech.

Nejzajímavější část práce pak popisuje algoritmus self-training, který počínaje předtrénovaným modelem iterativně extrahuje text aktuálním modelem, zvolí část „nejlepších“ predikcí a pomocí nich natrénuje model nový. Řešitel vyhodnocuje jednak korelaci několika metrik s úspěšností predikce (aby bylo možné vybrat „nejlepší“ predikce) a také množství rozvrhů, jaké při trénování používat. Výsledky dosažené na všech třech datasetech bez anotovaných trénovacích dat jsou velmi dobré v porovnání s výsledky s použitím anotovaných dat. Zároveň se trénování s anotovanými daty a self-training nijak nevyklučují a mohou se doplňovat. Tato část práce je dle mého mínění na vysoké úrovni a vyžadovala tvůrčí přístup a schopnost provedení velkého množství experimentů v relativně neprobádané oblasti.

Celkově hodnotím diplomovou práci jako velmi povedenou a prokazující schopnost samostatné výzkumné činnosti, vzhledem k tomu, že spojuje teoreticky náročné algoritmy (CTC beam search CTC dekodování s language modelem) včetně jejich efektivní C++ implementace, vytvoření flexibilního systému syntetického generování dat, návrh a provedení komplexních experimentů a v neposlední řadě dosažení velmi dobrých výsledků v méně zkoumané oblasti self-training.

**Práci doporučuji k obhajobě.**

**Práci nenavrhuji na zvláštní ocenění.**

**Datum** 23. srpen 2021

**Podpis**