Posudek vedoucího diplomové práce

Václav Honzík: Multi-modální zpracování dokumentů

Cílem práce je návrh, implementace a otestování prototypu systému pro multi-modální zpracování dokumentů. V případě dobrých výsledků bude systém integrován do historického portálu Porta fontium (PF) pro zlepšení jeho vyhledávacích možností.

Práce autora začíná stručným popisem neuronových sítí, které budou pro řešení využity. Následující kapitola, která se věnuje samotnému popisu multi-modálních přístupů, logicky navazuje na kapitolu první. Další kapitola popisuje relevantní datové sady pro multi-modální zpracování. S ohledem na cíle práce je důraz kladen na datasety obsahující text a obrázky. Následující text se věnuje popisu knihoven a systémů optického rozpoznávání znaků (OCR) s cílem identifikace metody pro integraci do výsledného prototypu. Na základě porovnání vlastností jednotlivých systémů byl vybrán nástroj Tesseract, jehož volbu považuji za vhodnou. Vzhledem k tomu, že žádný z dostupných datasetů nebyl dostatečně podobný datům z portálu PF, bylo rozhodnuto vytvořit vlastní datovou sadu. Nově vytvořený dataset obsahuje celkem 4640 vzorků, které patří do sedmi klasifikačních tříd objektů na stránce (např. nadpis, odstavec, tabulka, apod.). Jeho popisu se věnuje následující kapitola. Zde bych chtěl uvést, že samotné vytvoření datasetu je velkým přínosem nejen pro výzkum v dané oblasti, ale zejména pro vylepšení uvedeného portálu. Při tvorbě datasetu byl použit systém OCR, proto tato kapitola popisuje i tuto úlohu. Autor dosáhl s použitím doučení (fine tuning) CER ~ 1,2%, což jsou výsledky srovnatelné ze state of the art na podobných datech.

Dále se diplomant zabývá vlastním prototypem, kde je nejdříve popsán návrh a implementace celého systému, který je složen ze tří hlavních částí: segmentační, OCR a klasifikační. V segmentační části byly využity tři architektury neuronových sítí. Pro multi-modální klasifikaci byly srovnány dva přístupy: model LayoutLMv3 a fúzní model kombinující model BERT pro textovou a model Vision Transformer (ViT a Swin v2) pro obrazovou modalitu. Jako doplněk je možné použít ještě informace o rámečku (bounding boxu). Pro implementaci prototypu je použit jazyk python s frameworkem PyTorch. Pro implementaci modelů BERT, ViT, Swin V2 a LayoutLMv3 byla použita knihovna Huggingface Transformers. Volby sítí, způsob kombinace modalit i implementační nástroje považuji pro danou úlohu za velmi vhodné.

Následující kapitola obsahuje popis realizovaných experimentů a dosažené výsledky. Experimenty jsou rozděleny na segmentační a klasifikační. Student provedl obrovské množství experimentů se zajímavými výsledky. Ukázalo se, že výsledky multi-modálního přístupu klasifikace oproti využití pouze obrazové modality jsou velmi podobné. Na základě dalších experimentů bylo zjištěno, že je možné klasifikovat objekty na stránce s velmi dobrými výsledky s využitím pouze jedné modality. Nicméně obrazový vstup dává výrazně lepší výsledky (96,52 % vs. 83.89 %). Je tedy velmi pravděpodobné, že pro naši úlohu textová modalita neobsahuje doplňkovou informaci pro klasifikaci. Vzhledem k vysoké přesnosti dosažených výsledků je v plánu využít dosažené výsledky pro vylepšení portálu PF.

Průvodní dokument je vytvořen v systému LaTeX a je psaný anglickým jazykem. Má přehlednou strukturu. Je psán kvalitní angličtinou, neobsahuje pravopisné chyby ani překlepy. Při testování byl systém plně funkční, nebyly nalezeny žádné chyby.

Předložená diplomová práce je na vynikající úrovni a splňuje zadání. Je třeba dále zdůraznit, že téma práce je velmi rozsáhlé a složité a vyžadovalo nastudování řady informací z oblasti umělé inteligence. Autor zde prokázal výborné znalosti nejen z informatiky, ale i strojového učení. Přesvědčivě zde ukázal, že dokáže samostatně analyzovat a řešit složité problémy. Práci doporučuji k obhajobě a hodnotím klasifikačním stupněm

„výborně“

V Plzni 6. června 2023

doc. Ing. Pavel Král, Ph.D.
vedoucí DP