KATEDRA INFORMATIKY A VÝPOČETNÍ TECHNIKY

Josef Baloun

Diplomová práce

Inženýrská informatika

Medicínská informatika

2019/2020

Vedoucí práce:

doc. Ing. Pavel Král, Ph.D

FAKULTA APLIKOVANÝCH VĚD ZÁPADOČESKÉ UNIVERZITY V PLZNI

Segmentace stran rukopisných dokumentů

Abstrakt

Analýza stran dokumentů hraje významnou roli v procesu jejich elektronického zpřístupnění. Dokonce i v současné době může představovat nelehkou výzvu pro historické ručně psané dokumenty vzhledem k jejich různorodé struktuře a možné degradaci kvality. V rámci této práce je vypracován přehled možných metod pro řešení tohoto problému a vytvořena datová sada složená ze stran ručně psaných kronik. Dále je navržen prototyp systému pro analýzu stran dokumentů. Segmentace a klasifikace do tříd text, obrázek a pozadí jsou řešeny označením každého obrazového bodu strany dokumentu vhodnou třídou. Základem prototypu je plně konvoluční neuronová síť založená na síti U-Net. Pro trénování sítě je uvažováno např. váhování chybové funkce, automatické rozšíření trénovacích dat a možnosti načítání vstupu.

Úvod

V současné době je značná snaha o digitalizaci a elektronické zpřístupnění dokumentů ve většině oblastí. Toho se týká i projekt Bavorsko-česká síť digitálních historických pramenů, jehož cílem je za pomoci rozsáhlé digitalizace a webové prezentace spojit do jednoho virtuálního celku archiválie státních archivů České republiky a Bavorska.

Základním krokem k elektronickému zpřístupnění je naskenování stran dokumentů. Dalším krokem může být analýza stran dokumentů (angl. page layout analysis), která se skládá ze segmentace strany dokumentu na homogenní komponenty a jejich klasifikace např. na bloky textu a obrázky. Na výstup analýzy může být dále aplikováno např. optické rozpoznávání znaků (OCR). Původní stránka dokumentu tak může být převedena do značkované, strojově čitelné podoby umožňující efektivní vyhledávání v jejím obsahu. Je snadné si představit, kolik práce by to ušetřilo historikům.

Tato práce se týká zmiňované analýzy stran ručně psaných dokumentů s důrazem na kroniky z 19. a 20. století.

Systém pro segmentaci stran

K problému segmentace je přistupováno jako k pixel -labeling problému, kde je každý pixel zařazen do odpovídající třídy.

Základem systému pro segmentaci stran rukopisných dokumentů je plně konvoluční neuronová síť, která je oproti síti U-Net paměťově úspornější, využívá padding v konvolučních vrstvách a je navržena pro zpracování celých stran. Sdílení parametrů v konvolučních vrstvách je využito pro umožnění zpracování různých vstupních rozměrů.

Prototyp systému pro segmentaci stran

Postup zpracování začíná načtením vstupního obrázku ve stupních šedi a úpravou jeho rozměrů pro splnění požadavků na vstup, které jsou dány architekturou sítě. Poměry stran vstupního obrázku jsou přitom zachovány v maximální možné míře. Následuje segmentace pomocí natrénované sítě. Výstup sítě tvoří hodnoty poslední vrstvy s aktivační funkcí sigmoid. Tím se pro každou požadovanou třídu získá maska se stejnými rozměry jako jsou rozměry vstupu sítě. Tyto masky se nejdříve zvětší na původní rozměry obrázku a následně jsou prahovány. Výsledkem segmentace jsou binární masky pro jednotlivé třídy, jejichž rozměry odpovídají původnímu obrázku.

Datová sada

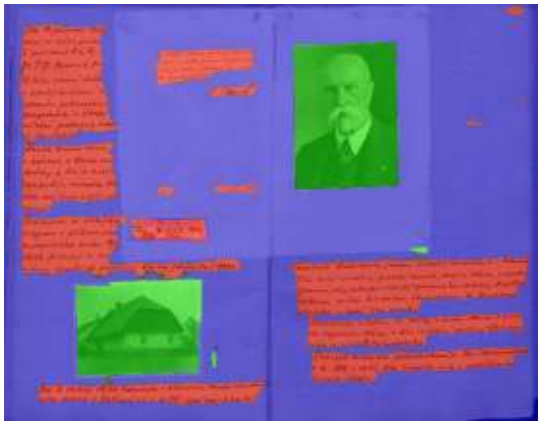
Pro úspěšné natrénování sítě je nutná odpovídající datová sada. Z tohoto důvodu byla vytvořena datová sada založená na stranách kronik poskytnutých z portálu Porta fontium.

Vedle stran kronik obsahuje datová sada i strany s tištěným textem pro možné experimenty. Anotováno bylo celkem 72 obrázků stran či dvoustran.

Dosažené výsledky

Nejlepších výsledků bylo dosaženo s prototypem, který ve fázi trénování využíval načítání celých stran, váhování chybové funkce a byla automaticky rozšířena trénovací množina.

Prototyp systému dosáhl v segmentaci stran kronik výborných výsledků: 0,908 IoU a 0,991 FgPA.



Ukázka segmentace strany kroniky: červeně text, zeleně obrázek a modře pozadí

Závěr

Velkou výhodou implementovaného řešení je možnost segmentace prakticky libovolných rozměrů strany.

Síť se díky možnosti zpracování celých stran a využití paddingu v konvolučních vrstvách dokáže velmi dobře vypořádat se šumem na okraji obrázků stran.

Zajímavou vlastností sítě je i schopnost generalizace. Přestože je síť trénovaná na starších ručně psaných kronikách, dokáže dobře segmentovat např. i moderní tištěné strany. Dobrých výsledků je dosaženo dokonce i se sítí natrénovanou na pouhých šesti stranách kronik.