



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

POVA

Počítačové vidění (v angličtině)

Projekt - vlastní téma Kvalita historických dokumentů v projektu PERO

Autor:
Petr Buchal

Login:
xbucha02

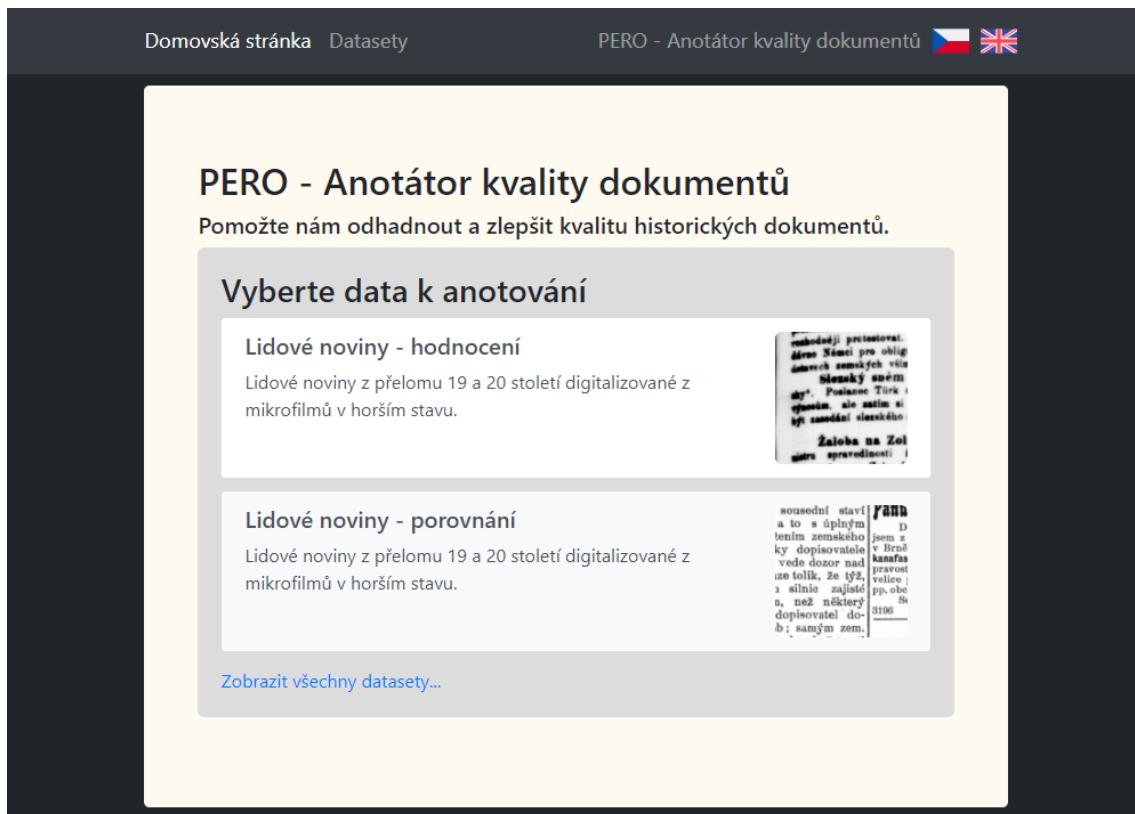
28. Prosinec, 2019

1 Úvod

Po domluvě s Ing. Michalem Hradišem Ph.D. jsme se rozhodl, že jako projekt do předmětu POVa zdokumentují svou práci na projektu PERO (Pokročilá extrakce a rozpoznávání obsahu tištěných a rukou psaných digitalizátů pro zvýšení jejich přístupnosti a využitelnosti) [3], která se týká zkoumání kvality historických dokumentů.

2 Platforma na vytváření anotací

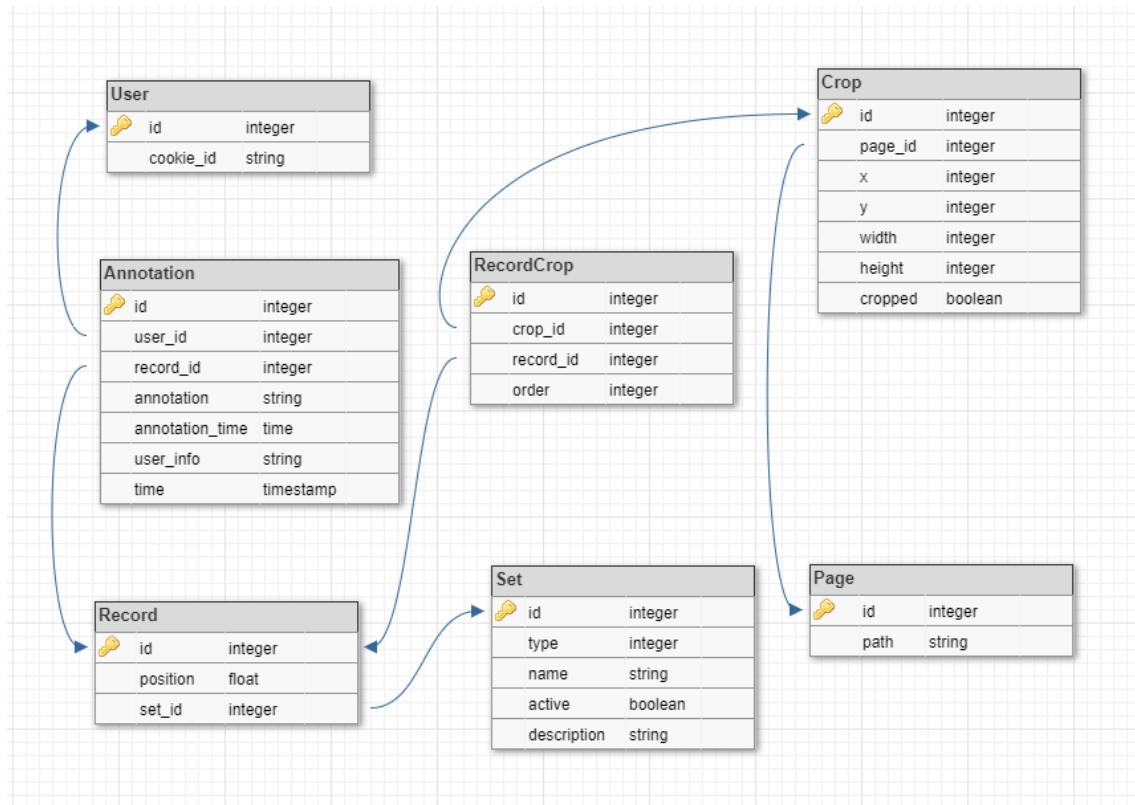
Pro provádění experimentů bylo třeba nejdříve vytvořit patřičný dataset a proto jsem vytvořil webovou platformu na získávání anotací od uživatelů. Platforma je implementována v jazyce Python 3 ve frameworku Flask. Stránka je responzivní, anotovat data tedy lze i na mobilním telefonu (mimo mód seřazování) a podporuje češtinu a angličtinu. Vzhled webové stránky je zobrazen na obrázku 1.



Obrázek 1: Úvodní stránka anotovací platformy.

2.1 Databáze

Na anotovací platformu je navázáná databáze SQLite ve které jsou uložená data k anotaci i samotné anotace. Schéma relační databáze se nachází na obrázku 2.



Obrázek 2: Schéma relační databáze.

Tabulka User slouží k ukládání uživatelů, kteří navštíví anotační platformu. Záznam vzniká při navštívení stránky, nikoliv při vytvoření anotace. S tím souvisí atribut `cookie_id`, jedná se o cookie identifikující uživatele v prohlížeči, které dále využívá Flask. Jako primární klíč pak slouží atribut `id`. Tabulka Set obsahuje datasety k anotování. Každý dataset má jako primární klíč `id`. Dále tabulka Set obsahuje atribut `type`, ten označuje mód, kterým jsou data v datasetu anotována (0 - porovnání, 1 - seřazování, 2 - hodnocení). Pak obsahuje atribut `name`, jedná se o název datasetu, který se bude objevovat na webové stránce. Dále obsahuje atribut `active`, jedná se o vlastnost, která určuje, jestli se dataset objevuje na webu nebo nikoliv (0 - neaktivní - nezobrazuje se, 1 - aktivní - zobrazuje se). V poslední řadě obsahuje dataset atribut `description`, jedná se o popis datasetu, který se rovněž bude objevovat na webové stránce. Každému datasetu jsou přiřazeny záznamy (tabulka Record), ten obsahuje atribut `id` (primární klíč), atribut `position` (podle něj se řídí pořadí zobrazování

záznamů na webu) a atribut set_id (cizí klíč odkazující na primáří klíč tabulky Set). Ke každému záznamu může být 0 až n anotací (tabulka Annotation), kde každou anotaci vytvořil právě jeden uživatel (tabulka User). Tabulka Annotation obsahuje atribut id (primární klíč), atribut user_id (cizí klíč odkazující na id uživatele), atribut record_id (odkazující na id záznamu), atribut annotation (samotná anotace), atribut annotation_time (čas, který zabralo vytvoření anotace uživateli), atribut user_info (jedná se o json s užitečnými informacemi o prostředí ve kterém uživatel vytvořil anotaci) a atribut time (čas, kdy uživatel vytvořil anotaci). Každý záznam obsahuje 1 až 5 výřezů (tabulka Crop) podle módu anotování (porovnání - 2 výřezy, seřazování - 5 výřezů, hodnocení - 1 výřez). Výřez (tabulka Crop) může být obsažen ve více záznamech (tabulka Record). Vztah mezi těmito dvěma tabulkami je m:n. Databáze tedy obsahuje tabulku RecordCrop, která mapování mezi tabulkami popisuje. Tabulka RecordCrop obsahuje atribut id (primární klíč), atribut crop_id (cizí klíč odkazující na id výřezu), atribut record_id (odkazující na id záznamu) a atribut order (určuje pořadí výřezů zobrazených na webové stránce). Výřez pak obsahuje atributy id (primární klíč), page_id (cizí klíč odkazující se na id stránky ze které byl výřez získán), x (nejlevější souřadnice na x-ové ose), y (nejhornější souřadnice na y-ové ose), width (šířka výřezu), height (výška výřezu), cropped (příznak značící zda byl již výřez vyřezán a uložen na disk). Při vkládání výřezu do databáze je příznak cropped nastaven na false, na true se mění ve chvíli, kdy má být prvně zobrazený na webové stránce. V ten okamžik se otevře soubor stránky z něj se získá výřez a uloží se na disk. Při dalším zobrazení stejného výřezu se načítá již přímo ze souboru, kde je uložený. Tabulka Page obsahuje atributy id (primární klíč) a path (cesta k souboru, obsahující naskenovanou stránku).

2.2 Módy

Platforma obsahuje tři anotovací módy. Prvním módem je porovnávací mód (obrázek 3). Když jej uživatel vyvolá, tak mu stránka zobrazí 2 obrázky a on má za úkol označit ten lépe čitelný. Může tak učinit pomocí kliknutí přímo na obrázky nebo pomocí kliknutí na tlačítko "Lépe čitelný" nebo pomocí šipek doleva a doprava. Po vybrání lepšího obrázku se do databáze ukládá buď anotace 01 (levý obrázek je lépe čitelný) nebo 10 (pravý obrázek je lépe čitelný).

Druhým módem je seřazovací mód (obrázek 4). Když jej uživatel vyvolá, tak mu stránka zobrazí 5 obrázků a on je má za úkol pomocí tahů myší seřadit od nejlépe čitelného po nejhůře čitelný. Tento mód funguje pouze pokud je v prohlížeči umožněn drag and drop. Příklad anotace v databázi pak může vypadat následovně: 23014, kde druhý obrázek je nejčitelnější, poté je nejčitelnější třetí a tak dále, kde nejhůře čitelný je čtvrtý obrázek.

Lépe čitelný

Lépe čitelný

nemají žádné. Němci jily, jakoby Hlasy z Hané“ sdělují, byl dopustil traté při volební agitaci edků, které již sami dotazu kuž na jiné strany někdy je z obmezování svobody železničních candidátní listinu dosavila se všecka živnosti a, rolnictvo i obchodníků = **Mateřská řeč v těstí**, že jim

Arturu. stranách. Tu Takuš prý bud týdne máme sil dá na sebe. Třetí čtyři divis 10. a 11.,

[Nápověda](#)

Obrázek 3: Mód porovnávání anotační platformy.

nejlépe čitelný

ský sněm. září. (Zvl. tel.) Sněm se usnesl, aby na oslavu a darského jubilea a slezského zemského i vyslné zřízení druhým pro 100 osob mužského bylo usneseno po delší zemskou p 100 káranců přes 14

nejhůře čitelný

ský sněm. září. (Zvl. tel.) Sněm se usnesl, aby na oslavu a darského jubilea a slezského zemského i vyslné zřízení druhým pro 100 osob mužského bylo usneseno po delší zemskou p 100 káranců přes 14

přijel v 8 v pátek o dročen dají ct. spolu s obecen u ještě dvě schůze , minuli vy mnouho Ms schůze.

Dr. W. Z Bud výboru pro vyšetření na Moravě a být konečně kanc. označení v Brně Alexandr Z vidně, 13. března. ledne příští době sešel se dnes i, posavadní Hr. A. tvé a v Kraslicích. Předseda různých krvavých feditelství v

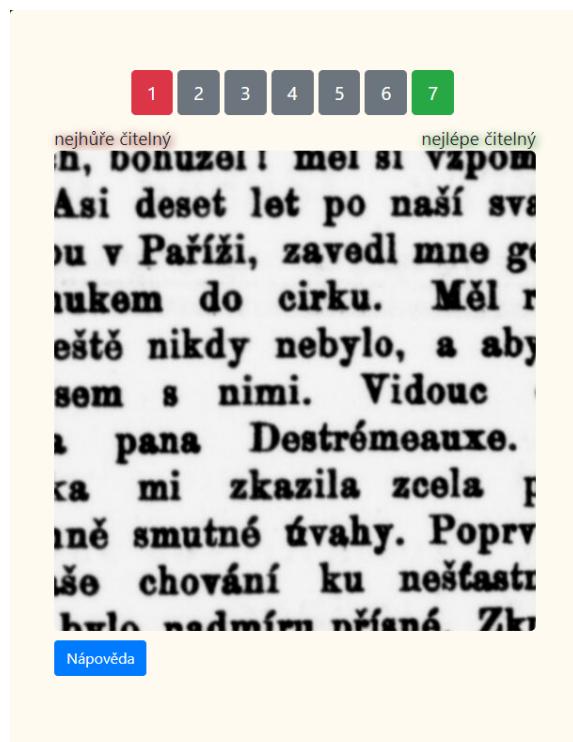
válel se v ven do nehosty v L Mnozí záprudkými tedy divu, zotavení h domů se v a nejmíň i Tataž nem

Sefazeno

[Nápověda](#)

Obrázek 4: Mód seřazování anotační platformy.

Třetím módem je hodnotící mód (obrázek 5). Když jej uživatel vyvolá, zobrazí se mu jeden obrázek. Ten pak má uživatel za úkol ohodnotit za pomocí kliknutí na tlačítka od 1 do 7 (1 - nejhůře čitelný, 7 - nejlépe čitelný), zároveň k tomu může použít číslice na klávesnici.



Obrázek 5: Mód hodnocení anotační platformy.

2.3 Získaná data

V průběhu výzkumu metod určování kvality vzniklo několik datasetů, na jejichž různých kombinacích byly prováděny experimenty. Veškerá data na kterých jsem experimenty prováděl pocházela z Lidových novin z počátku 20. století. Celkem bylo vytvořeno 9 843 anotací výřezů pocházejících z Lidových novin, z toho bylo 8 429 anotací unikátních. V režimu hodnocení bylo vytvořeno celkem 475 anotací, v režimu porovnání to bylo 9 368 a v režimu řazení nebyly vytvořeny žádné anotace.

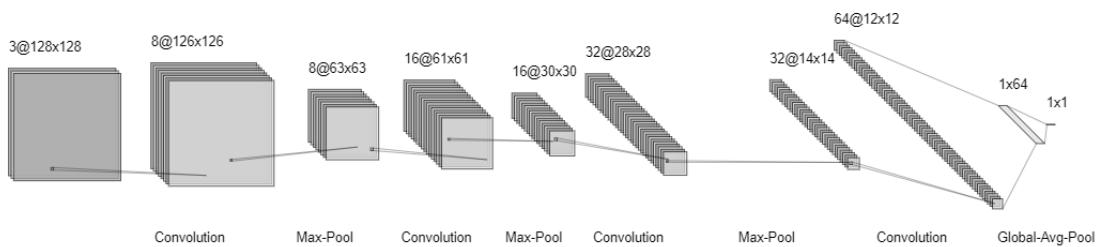
Souřadnice výřezů, které se vkládaly do databáze k anotování, byly generovány náhodně, při čemž byl následně výřez posouzen jednoduchým klasifikátorem, který určil zda výřez obsahuje text. Pokud obsahoval text byl přidán do databáze. Tento klasifikátor nebyl ve všech případech úspěšný, ale odfiltroval téměř všechny výřezy obsahující obrázky nebo prázdný papír, jejichž zobrazování se v platformě jevilo jako problém, protože uživatelé nevěděli, jak s takovými výřezy zacházet.

První vytvořený dataset obsahuje 2406 unikátních porovnání výřezů z náhodných stránek. Druhý dataset obsahuje 473 unikátních hodnocení výřezů z náhodných stránek. Třetí dataset obsahuje 1653 unikátních porovnání výřezů ze špatně čitelných stránek. Čtvrtý dataset obsahuje 3 297 unikátních porovnání 200 různých výřezů z

náhodných stránek. Pátý dataset obsahuje 600 unikátních porovnání 48 různých výřezů, které byly vybrány v té době nejlepším modelem jako výřezy napříč spektem kvality (od těch nejhorších po ty nejlepší).

3 Experimenty s dvojicemi obrázků

Cílem mé práce bylo získat neuronovou síť, která na vstupu zpracuje obrázek a na výstupu uživateli poskytne míru kvality historického dokumentu. První přístup, který jsem pro vytvoření takové neuronové sítě použil, bylo trénování popsané v následujících rádcích. Vytvořil jsem neuronovou síť, která na vstup dostane 2 obrázky a na výstupu poskytne číslo v rozmezí od 0 do 1 podle toho, který obrázek byl čitelnější. Dataset porovnání dvou obrázků jsem získal z platformy popsané v kapitole 2. Tato architektura se dá rozdělit na dva bloky, první částí je konvoluční zpracování obrazu (obrázek 6). Během učení touto částí postupně projdou oba dva obrázky tak, že na výstupu tohoto bloku budou dvě čísla. Ve druhém bloku neuronové sítě pak proběhne odečtení těchto dvou čísel a na výsledek je aplikována aktivační funkce sigmoida. Na výstupu takové neuronové sítě je tedy jedno číslo, které se učením snaží přiblížit štítkům dat (0 - je-li čitelnější první obrázek, 1 - je-li čitelnější druhý obrázek).



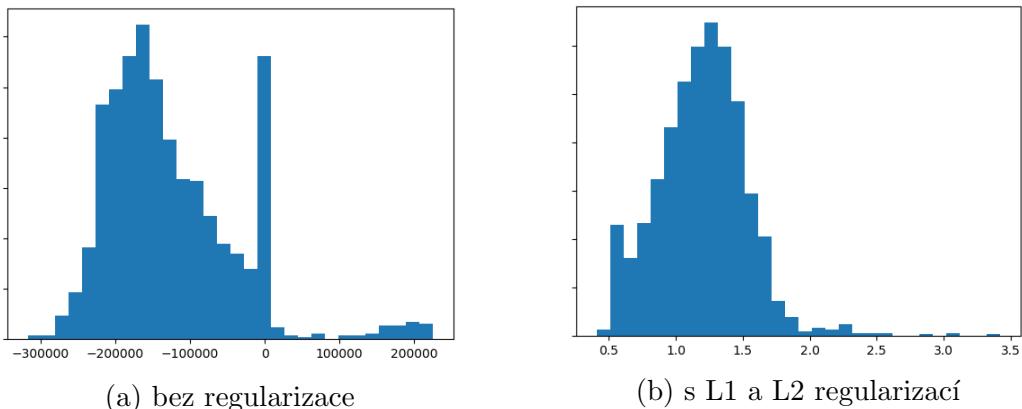
Obrázek 6: Architektura konvolučního bloku neuronové sítě při trénovaní.

Po takovém trénování by konvoluční část měla být schopna rozpozнат kvalitu historických dokumentů. Vyzkoušel jsem několik variant konvolučních bloků s drobnými úpravami (např. Max Pooling byl nahrazen Average Poolingem, byl změněn počet konvolučních vrstev, či velikost jejich filtrů, byla přidána vrstva Dropout). Úspěšnost různých variant při trénování se na testovací sadě pohybovala mezi 70 a 80 % (binary accuracy). Na vstup neuronové sítě jsem zároveň zkoušel dávat různé velikosti výřezů (šlo o výřezy z výřezů uložených v databázi). Ukázalo se, že nejlepších výsledků neuronová síť dosahovala, když na vstup dostávala výřezy originální velikosti uložené v databázi (512 x 512 pixelů). Tento model ovšem trpěl rychlým přetrénováním, protože těchto dat bylo o dost méně než, když neuronové sítě na vstup šly výřezy z

výřezů (128 x 128 pixelů a 256 x 256 pixelů). Zároveň větší vstup znamenal zpomalení sítě. Jako konvoluční blok jsem rovněž vyzkoušel použít část předtrénované sítě VGG, její výsledky ale byly horší než u navrhnuté architektury. Jako nejúspěšnější se ukázal model, který je na obrázku 2, který používá L1 a L2 regularizaci.

3.1 Problémy s vizualizací

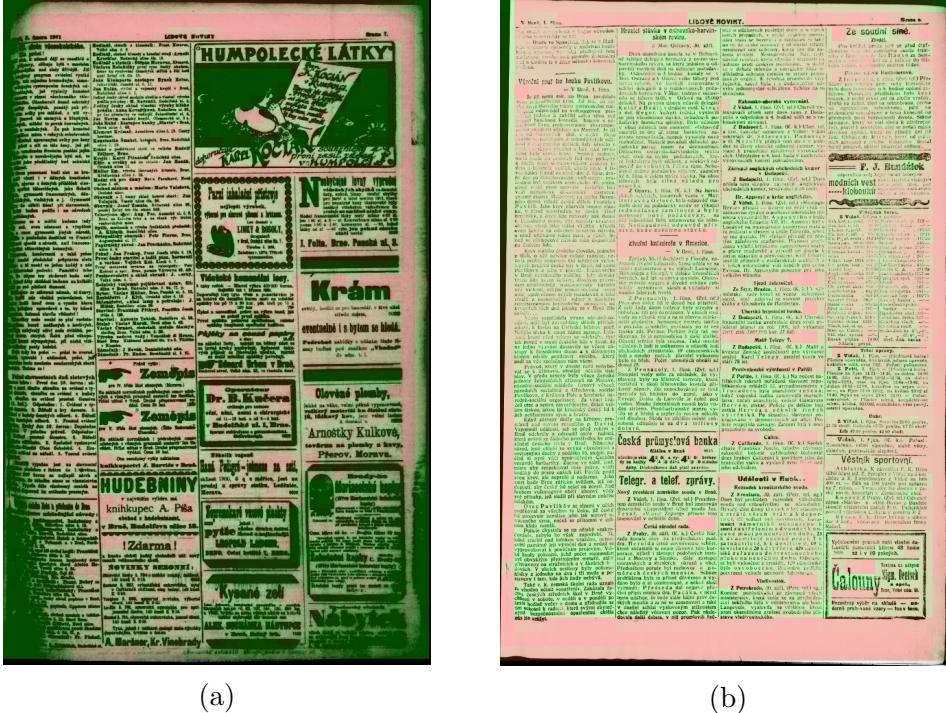
I přes úspěšnost mezi 70 a 80 % při vyhodnocování jednotlivých modelů vyvstalo několik problémů. Když jsem chtěl pouze konvoluční část neuronové sítě pro vykreslení teplotní mapy čitelnosti historického textu, ukázalo se, že hodnoty, které neuronová síť dává na výstup, mají velké rozpětí. Nešla z nich tedy určit obecná kvalita obrázku, protože chyběla minimální a maximální hodnota, která se může na výstupu takové sítě objevit a tedy nešla stanovit stupnice. Teplotní mapa byla možná vykreslit za použití normalizace všech hodnot získaných ze zkoumané stránky. V některých případech ale nefungovala nikterak skvěle, protože se stávalo, že rozptyl většiny hodnot nebyl velký a poté se našlo pár hodně vzdálených hodnot (obrázek 7a), které informační hodnotu tepelné mapy zničily. Tenhle problém jsem se pokusil odstranit regularizací v poslední plně propojené vrstvě konvoluční části neuronové sítě. Vyzkoušel jsem jak kombinaci regularizací L1 a L2, tak jejich použití zvlášť, nejvíce se osvědčila jejich kombinace. Hodnoty v neuronové síti s regularizací byly pro vykreslení heatmapy daleko lépe použitelné než bez ní, viz obrázek 7. Stále nicméně přetrvával problém chybějí minimální a maximální hodnoty pro vytvoření obecné stupnice kvality, tím se zabývám v kapitole 4.



Obrázek 7: Rozložení výstupních hodnot konvoluční části neuronové sítě z jedné historické stránky.

Architektura neuronové sítě pak tíhla k tomu, že tmavé části stránek považovala za dobře čitelné, i když dobře čitelné nebyly (obrázek 8). Zkusil jsem neuronovou síť

trénovat s kombinací originálních a invertovaných výřezů, ale tohoto neduhu se zcela nezbavila.



Obrázek 8: Teplotní mapy prvních natrénovaných modelů zobrazující kvalitu historického textu - chybě zvýrazňují tmavé části stránek.

4 Experimenty s hodnocením obrázků

Dat z hodnotícího módu anotační platformy bylo od začátku málo (druhý dataset - 473 unikátních hodnocení). I přes to jsem zkoušel natrénovat neuronovou síť regresí, vizuální výsledky ovšem neodpovídaly skutečnosti, a tak jsem se kvůli nedostatku dat těmto experimentům přestal věnovat. Dat z porovnávacího módu jsem měl ovšem dostatek, a tak jsem se rozhodl vyzkoušet tyto data na základě porovnání seřadit a podle seřazení opět vyzkoušet regresní trénování. K seřazení výřezů podle kvality na základě porovnání jsem si vybral Bradley-Terryho model [1]. Tento model byl v Pythonu již implementovaný, a tak jsem použil modul `choix` [4]. Nejdříve jsem ho použil na první dataset, ukázalo se ale, že každý výřez má v tomto datasetu porovnání s průměrně 3 dalšími a síť porovnání není dostatečně hustá pro vytvoření modelu, který dokáže výřezy seřadit. Vytvořil jsem tedy dataset, který obsahoval pouze 200 výřezů a každý výřez měl porovnání v průměru se 40 dalšími výřezy (s 20

% výřezů v celém datasetu). Po aplikaci Bradley-Terryho modelu na tento dataset, jsem dosáhl dobrých vizuálních výsledků při seřazení výřezů. Kvůli jejich nízkému počtu se ovšem dataset nedal využít pro trénování a škálování počtu výřezů by značně zvýšilo počet potřebných anotací.

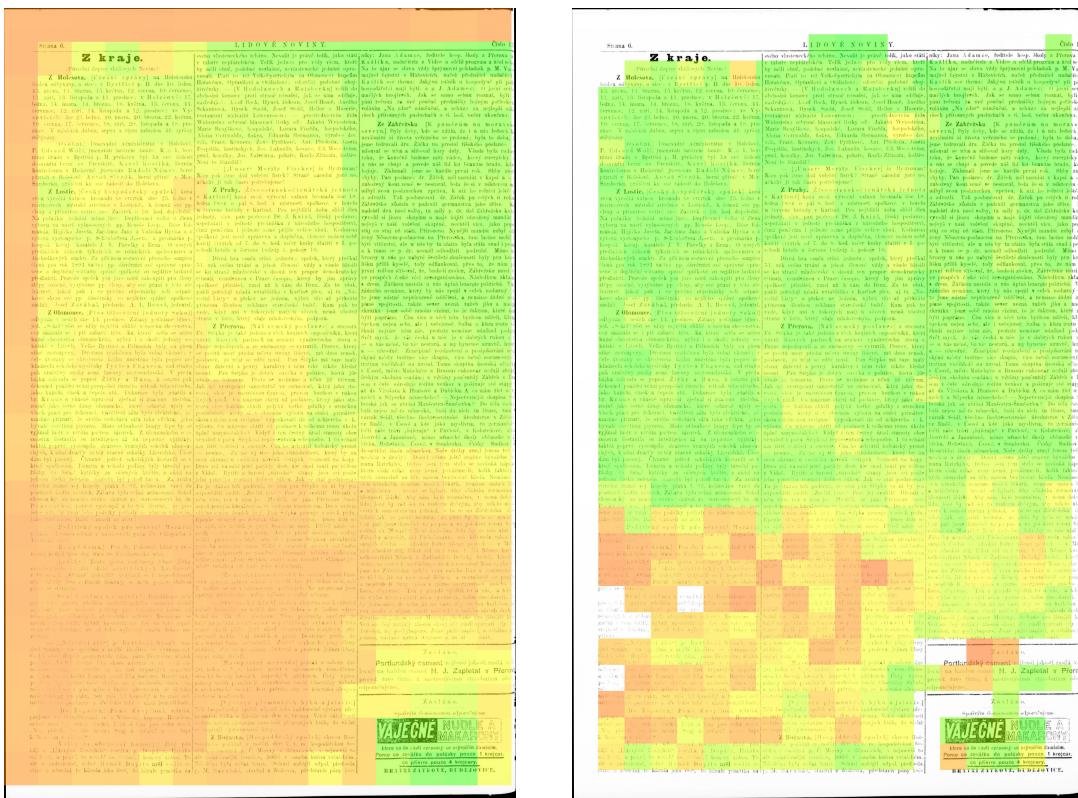
Získané znalosti o funkčnosti modelu jsem využil po vytvoření malého datasetu (pátý dataset), který jsem s Bc. Matúšem Bako využíval k hodnocení modelů pro určení kvality na základě korelace hodnot z obou modelů (Bradley-Terry a náš model) na pátém datasetu. Zároveň mi tento malý dataset umožnil mapovat výstupní hodnoty konvolučního bloku neuronové sítě na hodnoty Bradley-Terryho modelu získaného z pátého datasetu a to díky rozložení kvality výřezů po celém spektru. Tím jsem získal univerzální stupnici pro porovnání hodnot různých stránek.

5 Experimenty s regresním trénováním podle OCR

Jelikož se modely z předchozích kapitol ukázaly jako použitelné, ale ne nedokonalé, vyzkoušel jsem ještě jeden způsob trénovaní, jehož výsledkem jsou na základě vizuální kontroly nejpřesnější modely. Ukázka výstupu nejlepšího modelu je vidět na obrázku 9b. Těmito experimenty jsem navázal na práci kolegy Bc. Matúše Bako, který určoval kvalitu historických dokumentů na základě jistoty OCR. Jeho model měl na pátém testovacím datasetu korelací cca 86 %. Problémem architektury jeho OCR neuronové sítě byla její rychlosť. Vzal jsem tedy jím vytvořený dataset, který obsahoval výřezy z Lidových novin a jejich kvalitu a regresně jsem trénoval přibližně stejný model jako je na obrázku 6. Jeho korelace s pátým datasetem dosahuje 87 %.

5.1 Segmentace textu

Problém s tímto modelem je ten, že se učil jen z výřezů, které obsahovali nějaký text. Model tedy nevěděl jak přistupovat k částem stránek, kde žádný text není a tyto části zhoršovaly přesnost určení kvality. Bylo tedy třeba provádět vyhodnocení kvality stránky pouze tam, kde je nějaký text. K segmentaci textu jsem nejdříve použil naivní metodu. Ta určovala místa, kde se nachází text tam, kde je průměrný jas výřezu v určitém rozmezí. Pokud je na výřezu text, neblíží se průměrný jas výřezu horní hranici, a zároveň pokud je průměrný jas nízký, je na výřezu pravděpodobně tmavý obrázek. Tato metody měla pevně nastavené prahy rozmezí a tedy fungovala s různou úspěšností pro různé stránky. Následně jsem místo této naivní metody zkousil použít rychlou neuronovou síť na segmentaci obrazu od Ing. Oldřicha Kodyma. Tato metoda se ukázala jako účinnější a s jejím využitím jsem dosáhl vizuálně nejlepších výsledků, viz obrázek 9b.



(a) nejlepší model trénovaný pomocí metody popsané v kapitole 3 (b) nejlepší model trénovaný pomocí metody popsané v kapitole 5

Obrázek 9: Teplotní mapy nejlepších modelů.

6 Závěr

Vytvořil jsem platformu na tvorbu anotací, díky které bylo vytvořeno přes 10 000 anotací. Platforma je snadno modifikovatelná pro podobné účely ke kterým jsem ji vytvořil. V průběhu semestru jsem pomáhal Ing. Oldřichu Kodymovi platformu modifikovat pro sbírání dat o kvalitě rádků, kterým byla několika způsoby zlepšována kvalita. Tímto bych zároveň chtěl poděkovat všem, kteří mi ve svém volném čase pomohli s vytvářením anotací. Po vytvoření anotační platformy jsem vyzkoušel několik přístupů, jak trénovat modely zjišťující kvalitu historických dokumentů. Některé se ukázaly jako použitelné s určitými neduhy, jiné se ukázaly jako nefunkční. Nejlepších výsledků dosahoval regresivně trénovaný model podle OCR s použitím neuronové sítě pro segmentaci, viz kapitola 5. Nejúspěšnější model, který dosahoval požadované vizuální kvality, je nyní dostupný v oficiálním Githubovém repozitáři projektu pero [2].

Reference

- [1] Bradley–terry model, Oct 2019. https://en.wikipedia.org/wiki/Bradley-Terry_model.
- [2] DCGM. Dcgm/pero_quality_web, Dec 2019. https://github.com/DCGM/pero_quality_web.
- [3] Fakulta informačních technologií Vysokého učení technického v Brně. Project pero. <https://pero.fit.vutbr.cz/>.
- [4] Lucas Maystre. choix. <https://pypi.org/project/choix/>.