Jebavý, Filip

Pomocný software na tvorbu obrazových datových sad v digitální knihovně s využitím strojového učení

ProInflow. 2023, vol. 15, iss. 2, pp. [155]-175

ISSN 1804-2406 (online)

Stable URL (DOI): https://doi.org/10.5817/ProIn2023-36869
Stable URL (handle): https://hdl.handle.net/11222.digilib/digilib.79620

Odborný článek

ProInflow: časopis pro informační vědy
Vol. 15, No. 2, 2023, str. 155–175
https://doi.org/10.5817/ProIn2023-36869

POMOCNÝ SOFTWARE NA TVORBU OBRAZOVÝCH DATOVÝCH SAD V DIGITÁLNÍ KNIHOVNĚ S VYUŽITÍM STROJOVÉHO UČENÍ

ASSISTIVE SOFTWARE FOR THE CREATION OF IMAGE DATASET IN A DIGITAL LIBRARY USING MACHINE LEARNING

Filip Jebavý

Moravská zemská knihovna v Brně

Abstrakt

Účel – Tento odborný článek popisuje možnosti využití pomocného softwaru za účelem efektivní tvorby obrazových datových sad z dokumentů digitální knihovny. Popisovaný software, kromě běžných způsobů práce s daty, využívá prvky strojového učení, které mají potenciál jak práci anotátorů usnadnit, tak také změnit anotační praktiky. Zároveň je kladen důraz na jednoduchost a otevřenost celého procesu. Cílem je na tyto prvky upozornit pomocí praktických ukázek.

Design / metodologie / přístup – Po úvodní části jsou představeny možnosti výběru a separace dat z dokumentů digitální knihovny. Zároveň je poukázáno na limity těchto přístupů. Na základě těchto poznatků jsou poté zkoumány možné přístupy a využití pomocného softwaru za účelem tyto limity překonat. Metody jsou popisovány na základě praktického využití softwaru při anotačním procesu. Validace prvků strojového učení je provedena mimo jiné vizualizační technikou *Class Activation Mapping* a pomocí metriky *F-score*.

Výsledky – Popisované přístupy a využití pomocného softwaru s prvky strojového učení se ukázalo jako velmi přínosné. Software nejen práci anotátorů ulehčuje, ale zároveň značným způsobem urychluje a zpřesňuje. Za velké pozitivum lze považovat univerzálnost testovaného modelu strojového učení, která umožňuje rozšířit anotační procesy za zprvu předpokládané využití, a dává tedy prostor pro další výzkum v této oblasti.

Originalita / hodnota – Odborný článek poukazuje na možné přístupy využití pomocného softwaru, usnadňující tvorbu obrazových datových sad u dokumentů s omezeným množstvím identifikátorů, jako je například digitální knihovna, a to bez potřeby komerčních nástrojů. Dále ukazuje praktické příklady, jak lze pomocí strojového učení tyto procesy zefektivnit. Podstatné jsou také příklady možností univerzálního využití těchto procesů.

Klíčová slova: datové sady, software, strojové učení, digitální knihovna, anotace

Abstract

Purpose – This paper describes the possibilities of using assistive software to efficiently create image datasets from digital library documents. The software described, in addition to the usual ways of working with data, uses machine learning features that have the potential to both make the work of annotators easier and to change annotation practices. At the same time, the emphasis is on simplicity and openness of the whole process. The aim is to highlight these elements through practical examples.

Design / methodology / approach – After an introductory section, the possibilities for selecting and separating data from digital library documents are presented. At the same time, the limitations of these approaches are pointed out. Based on these insights, possible approaches and the use of assistive software are then explored in order to overcome these limits. The methods are described based on the practical use of the software in the annotation process. The validation of the machine learning features is performed using, among others, the visualization technique Class Activation Mapping and the F-score metric.

Results – The described approaches and the use of assistive software with machine learning features proved to be very beneficial. The software not only makes the work of the annotators easier but also considerably faster and more accurate. The versatility of the tested machine learning model also proved to be a great positive, allowing to extend the annotation processes beyond the initially assumed use and thus giving room for further research in this area.

Originality / value – The technical paper highlights possible approaches to use assistive software to facilitate the creation of datasets for documents with a limited number of identifiers, such as a digital library, without the need for commercial tools. It also shows practical examples of how machine learning can be used to make these processes more efficient. Examples of how these processes can be used universally are also provided.

Keywords: datasets, software, machine learning, digital library, annotation

ÚVOD

V posledních několika letech jsme svědky výrazného posunu ve vývoji strojového učení a umělých neuronových sítí. Mohlo by se zdát, že za tento relativně rychlý pokrok může především rozvoj hardwarových technologií – například v oblasti grafických karet (GPU) – a průlom v oblastech algoritmických technik a s tím spojená problematika hlubokého učení. Ostatně tyto dvě oblasti jsou historicky vzájemně provázány. Funkčnost algoritmů hlubokého učení vznikajících koncem 20. století se začala projevovat až s příchodem výkonnějších infrastruktur, a naopak tato infrastruktura umožnila vznik novým a komplexnějším modelům, které na odhalení svého plného potenciálu možná teprve čekají (Goodfellow et al., 2016). Tyto faktory však hrály zásadní roli především na přelomu 20. a 21. století. V současné době můžeme úspěšnost a výrazný pokrok v oblasti strojového učení zdůvodnit zejména vznikem nových, rozsáhlejších a lépe strukturovaných datových sad (Ratner, De Sa, et al., 2017). Obrovské množství volně dostupných dat v digitální podobě, pečlivě spojených do jednotlivých sad, umožňují modelům umělých neuronových sítí získat širší kontext, lepší generalizaci a zlepšenou schopnost odhalovat složité vzorce. Pečlivé spojení těchto dat však často představuje mnohem náročnější úkol, než jakým je samotná kompletizace modelu strojového učení (Ratner, Bach, et al., 2017). Výhodnou oblastí pro sběr dat jsou tedy místa, kde jsou data již určitým způsobem kategorizována. Jedním z těchto míst je právě digitální knihovna.

Digitální knihovny jsou velmi dobrým zdrojem pro tvorbu datových sad v oblasti strojového učení, a to hned z několika důvodů. Zaprvé obsahují obrovské množství textových informací, včetně knih, periodik, vědeckých studií a dalších literárních či akademických děl. Tato data poskytují kvalitní základnu pro trénování modelů na různé úlohy – rozpoznávání obrazů, strojový překlad, sémantickou analýzu a podobně. Zadruhé jsou digitální dokumenty opatřeny příslušnými identifikátory a zároveň jsou navázány na metadata. Samotná metadata jsou již velmi cenná a v mnoha případech usnadňují kategorizaci dokumentů bez potřeby manuální anotace. Posledním důležitým bodem je fakt, že každý dokument prochází častou kontrolou katalogizačních záznamů, čímž se zvyšuje kvalita a spolehlivost dat. Zde je však důležité poznamenat, že tyto výhody digitální knihovny lze aplikovat především na tvorbu datových sad z pohledu celých exemplářů, na které se zmíněné identifikátory vztahují. Jelikož digitální dokumenty vznikají zejména pro potřeby knihovny a jejích uživatelů – nikoliv tedy z důvodu tvorby a ukládání datových sad – některé prvky nelze jednoduše kategorizovat pouhým využitím informací z metadat. S počtem desítek milionů stran už jen v digitálních knihovnách

České republiky (*Registr Krameriů*, b.r.) je žádoucí zaměřit se právě na tyto elementy, které jsou cenné z pohledu tvorby datových sad především svým obsahem, nikoliv jednoduchostí jejich kategorizace. To je také hlavní důvod vzniku pomocného softwaru, který je, společně s principy jeho funkčnosti, hlavním předmětem této studie. Pomocí softwaru se snažíme řešit problematiku třemi hlavními způsoby: 1. umožnit snadnou a intuitivní anotaci obrazových dokumentů; 2. usnadnit a urychlit anotaci pomocí technik strojového učení; 3. zautomatizovat procesy spojené s přípravou tvorby datových sad. Pro správné pochopení problematiky je však nejprve potřeba poukázat na základní principy možností extrakce dat z digitální knihovny.

EXTRAKCE OBRAZOVÝCH DAT Z DIGITÁLNÍ KNIHOVNY

Digitální knihovna nám díky svému obsahu umožňuje tvorbu velkého množství různých druhů datových sad. Pokud bychom se například zaměřili pouze na monografie, můžeme separovat hned několik prvků – obálku, titulní list, obsah, text, textové prvky, netextové prvky, druh papíru, písmo, prázdné strany a mnoho dalšího. Tyto prvky lze dále dělit do konkrétnějších kategorií. Abychom mohli z těchto dat vytvořit obrazovou datovou sadu, musíme nejprve data získat a kategorizovat. V našem případě pracujeme s digitální knihovnou Moravské zemské knihovny v Brně (dále jen MZK), kde je každý samostatný dokument i každá samostatná strana označena Univerzálním unikátním identifikátorem (dále jen UUID). Po zjištění tohoto identifikátoru tedy budeme považovat dokument či stranu za získanou. Samotné získání dokumentu či strany z UUID je již jednoduché – v našem případě například volání API frameworku IIIF (*API Specifications – International Image Interoperability Framework™*, b.r.). Prvním krokem je tedy stanovit, z jakých prvků chceme datovou sadu vytvořit. Pokud bychom chtěli například vytvořit datovou sadu obsahující pouze rukopisné strany, hledaným prvkem jsou pro nás UUID stran rukopisných dokumentů. V tomto případě můžeme využít informaci z metadat a pomocí příkazu query a REST API systému Kramerius (systém digitální knihovny používaný v MZK) získat UUID požadovaných stran velmi jednoduše (*API v7 · ceskaexpedice/kramerius Wiki*, b.r.). Query předpis by tedy vypadal takto:

own_parent.model:manuscript AND model:page

Celé volání poté takto:

https://api.kramerius.mzk.cz/search/api/client/v7.0/search?q=own_parent.model:manuscript%20AND%20model:page&fl=pid,own_parent..pid&rows=100000

Tím bychom získali potřebné UUID pro naši datovou sadu. V tomto případě jsme mohli využít již zmiňovanou výhodu digitální knihovny – metadata. Tento způsob tvorby datových sad můžeme využívat v takovém rozsahu, v jakém jsou informace obsaženy v metadatech příslušných dokumentů. Ostatně daný příklad můžeme dále rozšířit a požadovaný prvek naopak zúžit. Jedním z používaných identifikátorů je například typ stránky.

Pokud bychom se tedy chtěli zaměřit pouze na rejstříky rukopisných dokumentů, příkaz rozšíříme takto:

own_parent.model:manuscript AND model:page AND page.type:Index

Celé volání takto:

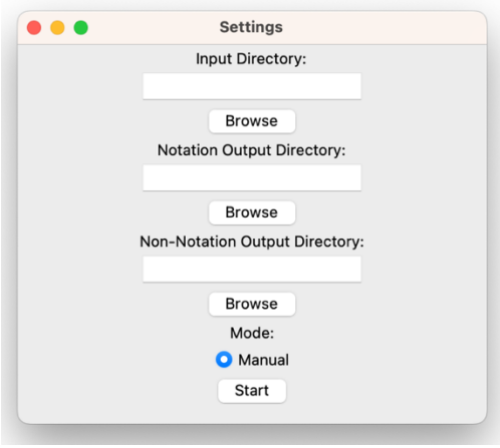
https://api.kramerius.mzk.cz/search/api/client/v7.0/search?q=own_parent.model:manuscript%20AND%20model:page%20AND%20page.type:Index&fl=pid&rows=100

Ve chvíli, kdy metadata obsahují příslušný typ dokumentu či stránky, můžeme již mluvit o anotovaných datech. Získání těchto dat a následná tvorba datových sad není komplikovaným procesem. Problém nastává ve chvíli, kdy metadata žádaný údaj neobsahují. Jedná se především o netextové objekty, které z pohledu katalogizačních pravidel není potřeba označovat, protože při tvorbě metadat by bylo jejich označování velmi pracné a zdlouhavé. Netextovým prvkem je například notový záznam, který v digitální knihovně MZK můžeme nalézt hned v několika podobách. Jedná se o dobrý příklad už jen z toho důvodu, že metadata hudebnin sice obsahují informaci „parent.model:sheetmusic“, ale tuto informaci bychom již v typu stránky ve většině případů nenašli. Pomocí příkazu query tedy můžeme nalézt a získat veškeré hudebniny, ale pokud bude jako základní prvek datové sady strana obsahující notový záznam, vzniká v tomto přístupu problém. Navíc netextový objekt notového záznamu se nenachází pouze v dokumentech označených jako hudebnina, tedy modelem „sheetmusic“. Velmi dobrým příkladem jsou liturgické knihy, které jsou často označovány modelem „monograph“. Specifické druhy liturgických knih – kancionál, graduál, notovaný misál atd. – přitom notový záznam většinou obsahují, a to i ve velkém množství. Na tvorbu tohoto druhu datové sady, kde je hlavním prvkem netextový objekt notový záznam, jsme tedy museli přijít s jiným řešením.

POMOCNÝ SOFTWARE NA TVORBU OBRAZOVÝCH DATOVÝCH SAD

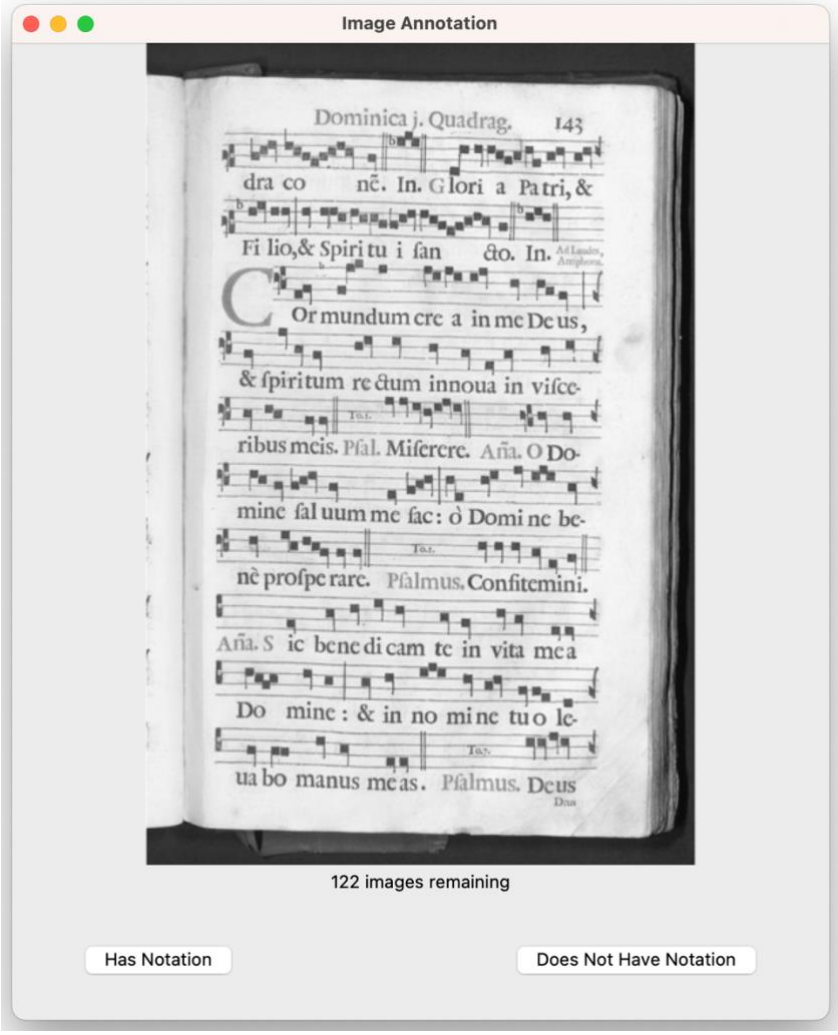
Z důvodu velkého spektra dokumentů a počtu stránek některých liturgických knih je zřejmé, že manuální prohledávání digitální knihovny, procházení každého dokumentu a zapisování UUID odpovídajících stran není vhodným postupem. Na druhou stranu poměr mezi počtem exemplářů celých liturgických knih a množstvím stran v těchto knihách nám pomůže zúžit hledané řešení na pouhý binární problém, a to tedy rozlišit strany s notovým záznamem a bez něj. Zjednodušeně řečeno, najít liturgické knihy v digitální knihovně je o tolik jednodušší úloha než separovat tyto dokumenty na jednotlivé stránky podle zadaného prvku, že si u hledání celých dokumentů vystačíme pouze s již existujícími fasetami a indexem digitální knihovny. Na získání celých dokumentů můžeme navíc velmi často využít příkazu query, u něhož si ve většině případů vystačíme pouze s dotazem na model dokumentu. Typ stránky či objekt samotný v takovém případě nemusíme řešit. Poté, co jsme si stanovili, že hledané řešení bude mimo systém Kramerius, je zřejmé, že odpovědí na tuto otázku je nezávislý software. Tomuto způsobu řešení nahrává také fakt, že tok samotných dat není komplikovanou záležitostí, a tedy hledat řešení jiné než lokální – ač by to bylo z mnoha pohledů výhodné – je pouhou komplikací.

Software má tedy pomoci se separací celých dokumentů na jednotlivé prvky. Musí být také dostatečně jednoduchý a intuitivní pro tým anotátorů, díky čemuž se zbavíme potřeby zaškolovat anotátory v používání příkazů query a API volání systému Kramerius. Na základě těchto zásad vznikl jednoduchý anotační software, který uživateli umožňuje zvolit vstupní data – formát PDF, JPG, PNG – a výstupní místo pro strany s notovým záznamem a pro ty, které jej neobsahují.



Obr. 1 Nastavení vstupních a výstupních složek

Po spuštění anotačního procesu se zobrazí první strana vstupních dat a uživatel má možnost určit, jestli se na ní daný prvek nachází, či nikoliv.



Obr. 2 Proces manuální anotace

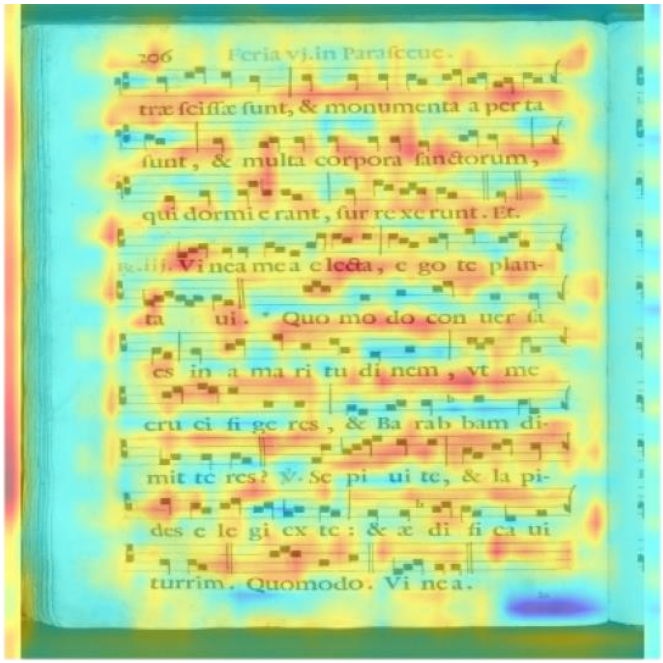
V uživatelském rozhraní je také zobrazován počet zbývajících dokumentů. Pro urychlení tohoto procesu jsou tlačítka navázána na kurzorové šipky klávesnice. Výstupem jsou tedy dvě složky, každá obsahující svou cílenou kategorii či prvek. Abychom však splnili zásadu, že získaný dokument pro nás může představovat i pouhé UUID strany, je v každé výstupní složce automaticky vytvořen soubor ve formátu txt, do kterého se okamžitě po rozhodnutí o anotaci zapíše název zkoumaného souboru. V našem případě identifikátor dané strany. Výsledkem je tedy seznam UUID stran, které splňují určený požadavek a spadají do jedné ze zvolených kategorií. Tento seznam lze poté jednoduše použít ke stažení těchto dokumentů nebo například k obohacení metadat v digitální knihovně.

Ačkoliv se jedná o velmi jednoduchý princip, tento software představuje velké zjednodušení a urychlení tvorby datových sad prvků, které nelze z digitální knihovny jiným způsobem automaticky vytěžit. I přesto je celý proces stále manuální prací anotátora, a tedy po dokončení anotace je nezbytná další kontrola výsledných dokumentů v kategorii. Nesprávné označení stran by totiž u tvorby datových sad za účelem strojového učení mohlo mít v některých případech za následek nesprávné fungování trénovaného modelu (Northcutt et al., 2021). S porovnáním se získáváním potřebných dokumentů pomocí příkazů query se tedy z tohoto pohledu stále jedná o velmi zdlouhavý proces.

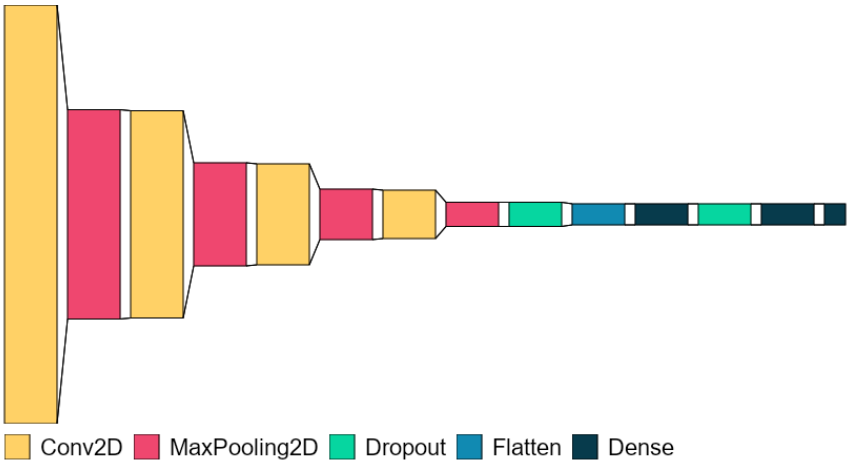
TVORBA OBRAZOVÝCH DATOVÝCH SAD S ASISTENCÍ

Chybovost datových sad je do určité míry úměrná rychlosti vytvoření anotace. Jednoduché prostředí softwaru může dále svádět k rychlému proklikávání daty a s tím spojené zdržení u následné kontroly výsledných kategorií. Abychom se tomuto problému vyhnuli a zároveň zlepšili přesnost anotovaných dat, rozhodli jsme se vytvořit model umělé neuronové sítě, který by o správnosti dat dokázal rozhodnout. I toto řešení musí splňovat několik zásad. Architektura modelu musí být dostatečně jednoduchá, abychom si vystačili s menším množstvím trénovacích dat. Tento bod je zásadní, protože využití velmi hluboké či široké umělé neuronové sítě předpokládá, že máme potřebné, a hlavně efektivní nástroje na tvorbu dostatečně velké datové sady. To je však v našem případě právě ta problematika, kterou se snažíme vyřešit. Pokud bychom totiž neměli dostatečně velké množství dat, vystavili bychom se nebezpečí v podobě přetrénování neboli *overfitting* (Zhang et al., 2017). Na druhou stranu architektura modelu nemůže být příliš jednoduchá, aby rozdíly mezi kategoriemi dokázala zachytit. Navíc jsme se chtěli vyhnout příliš velkému užití technik augmentace dat. Zaměřili jsme se tedy na jiné techniky, a to především na regulaci L2 (*weight decay*), vrstvy *dropout* a *early stopping* (Ying, 2019). To nám umožnilo začít s relativně malou datovou sadou o velikosti přibližně 2000 vzorků. Po několika testech a tvorbě odlišných modelů podle druhu hledaných parametrů (v našem případě podle druhu notace), jsme se dopracovali až na datovou sadu o velikosti přibližně 12000 vzorků obsahující různé druhy notace a stran z liturgických knih (již za pomocí softwaru).

To nám umožnilo vytvořit výsledný model s úspěšností F skóre 0.92 (k tomuto bodu se ještě vrátíme), schopný najít notový záznam bez ohledu na to, jestli se jedná o chorální, či menzurální notaci.



Obr. 3 Validace funkčnosti modelu pomocí techniky Class Activation Mapping. Červená barva znázorňuje nejvyšší hodnoty aktivace, které označují nejdůležitější a vysoce relevantní oblasti. Modrá naopak nejnižší hodnoty aktivace, které představují oblasti menšího

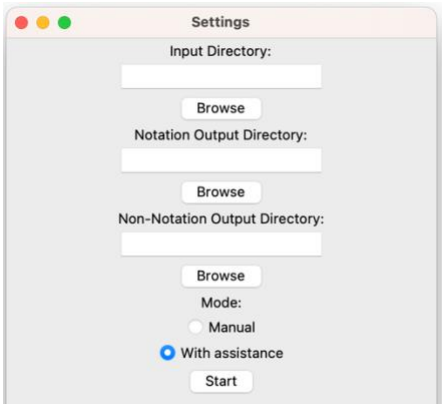


Obr. 4 Grafické znázornění podoby výsledné architektury modelu – základem jsou čtyři konvoluční vrstvy a dvě vrstvy plně propojené

Neuronová síť si dokáže poradit i v případě, že je hledaný prvek jen malou částí celé strany. Přesnost nesnižují dokonce ani jiné netextové objekty.

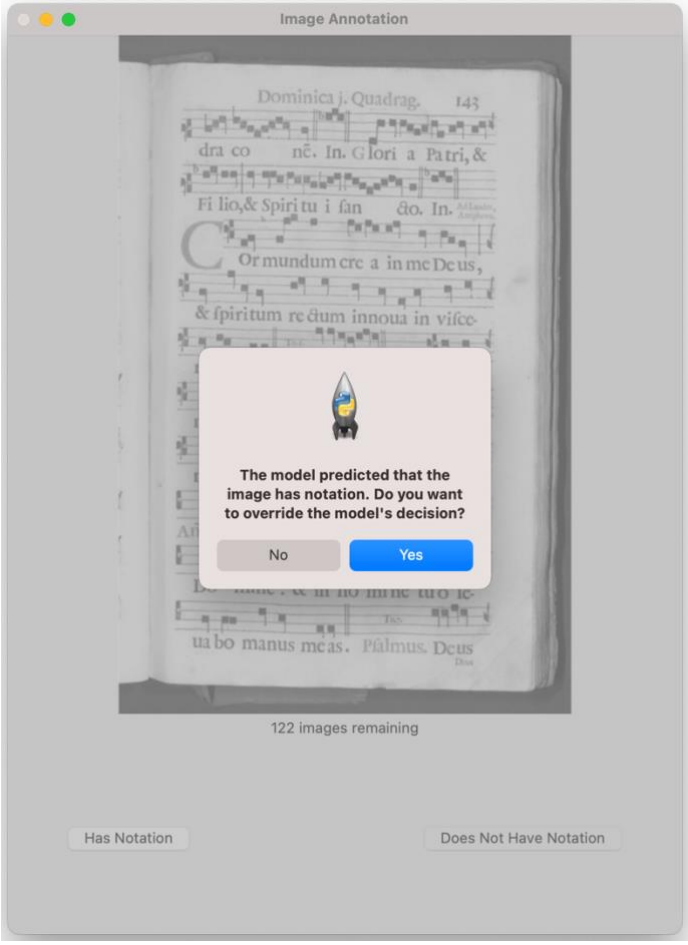
Obr. 5 Validace funkčnosti modelu pomocí techniky Class Activation Mapping

Poté, co jsme vytvořili a natrénovali funkční model musíme, najít způsob, jak jej co nejlépe zaimplementovat do procesu tvorby datové sady. Samozřejmě se hned nabízí nechat umělou neuronovou síť automaticky kategorizovat všechna data, která následně anotátor překontroluje. Tím bychom však faktor lidské chybovosti omezili jen částečně. Opačný proces také nelze v praxi aplikovat, protože samotný princip chybovosti modelů by mohl naopak vytvořit chyby nové. Vymysleli jsme tedy způsob, který využívá výhody obou variant a zároveň celý proces urychlí. Nazvali jsme ho mód s asistencí.



Obr. 6 Nastavení vstupních a výstupních složek – mód s asistencí

Mód s asistencí funguje na první pohled úplně stejně, jako mód manuální. Uživatelské rozhraní zůstává zachované. Změna se projeví ve chvíli, kdy anotátor vzorek špatně vyhodnotí. Ve stejnou chvíli totiž o kategorizaci vzorku rozhoduje také model umělé neuronové sítě a pokud se rozhodnutí modelu a anotátora liší, anotátor je na tuto skutečnost upozorněn.



Obr. 7 Model upozorňuje anotátora, že se jejich rozhodnutí neshodují

Ve chvíli, kdy je anotátor upozorněn, může rozhodnout, jestli chce upřednostnit predikci modelu, či nikoliv. Podle tohoto rozhodnutí se poté vzorek přesune do příslušné složky jako obvykle a proces pokračuje dalším vzorkem. Když se rozhodnutí shoduje, je tento asistenční mód od manuálního nerozeznatelný. Model umělé neuronové sítě bude pouze na pozadí kontrolovat správnost anotovaných vzorků.

Tento princip považujeme za velmi výhodný zejména proto, že ač nelze vyloučit chybovost jak lidského anotátora, tak vytvořeného modelu, předpokládáme, že pravděpodobnost chyby dvou subjektů ve stejnou chvíli je menší, než kdyby procesy probíhaly nezávisle na sobě. Anotátor

nemůže vzorky pouze rychle proklikat, protože ho model při chybách nepustí. Naopak pečlivější přístup je u tohoto asistenčního módu časově výhodnější, protože se tak lze vyhnout potenciálnímu upozorňování modelu.

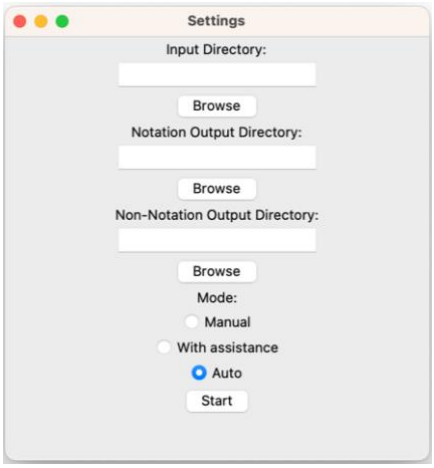
Tento způsob se tedy ukázal jako velmi efektivní pro tvorbu větších datových sad. Zde je také důležité zmínit, že představovaný software je velmi flexibilní. Ačkoliv bylo jeho primárním účelem řešit binární problém, implementace více kategorií je velice jednoduchá. Stejně tak lze využívaný model velmi jednoduše změnit. Software tedy můžeme s menšími úpravami využít na celou řadu anotačních problematik.

Obtížným úkolem by však mohla být snaha vytvořit pomocí tohoto principu datovou sadu zaměřující se na specifický druh prvku, který se však vyskytuje v širokém spektru dokumentů – nehledáme stránku, ale přímo prvek. Pomineme-li principy segmentace stran pomocí umělých neuronových sítí, které jsou problematikou samy o sobě, vytvořit model s dostatečně jednoduchou architekturou, ale dostačující hloubkou u takovéhoto typu prvku není lehkým úkolem. Jak se však ukázalo, i tento problém dokážeme alespoň z části vyřešit.

AUTOMATICKÁ PREANOTACE

S problematikou hledání stejného prvku ve velkém spektru různě formovaných dokumentů se můžeme obecně setkat u periodik, jako jsou například noviny. Jejich struktura a velký počet různých prvků na jednotlivých stránkách vyžaduje ve většině případů použití komplexnějších druhů architektur, včetně transponovaných konvolučních vrstev a *upscalingu* (Meier et al., 2017). Datová sada v podobě „správných a nesprávných" stránek tedy v tomto případě nemusí stačit. Naším úkolem nicméně není tyto prvky automaticky dohledat, ale pouze tvorbu datových sad usnadnit. Jak se při našich testech modelů umělých neuronových sítí ukázalo, i zde můžeme z části využít stejného principu jako u liturgických knih. Správně naučený model je totiž schopen hledat podobné prvky i u stran novin. Zde se můžeme vrátit k tvrzení z předešlé kapitoly o úspěšnosti modelu – F skóre 0.92. U tvorby různých architektur se nám totiž dařilo dosahovat F skóre v rozmezí 0.96 až 0.98. Důvodem, proč jsme však zvolili architekturu modelu s nižším F skóre, je právě mnohem lepší schopnost rozeznávat jednotlivé samostatné prvky a větší citlivost v pozitivním měření (jak správně pozitivní, tak nesprávně pozitivní). Naopak nesprávně negativní rozhodnutí je potlačeno na minimum. Zjednodušeně řečeno je pro nás mnohem přínosnější zaměřit se na co největší správnost vzorku „bez notace". Tím se také vysvětluje celkově nižší skóre při validaci. Největším rizikem tohoto principu je samozřejmě přílišná citlivost – pokud jsou výsledky pouze správně pozitivní nebo nesprávně pozitivní, k žádným výsledkům se ve skutečnosti nedopracujeme (model by všechny vzorky řadil pouze do jedné složky). Po několika testech se nám však podařila najít přijatelná rovnováha a z testů vychází, že je tento model schopen určit správně negativní případy (strany bez notace) s přesností přibližně 99,91 %.

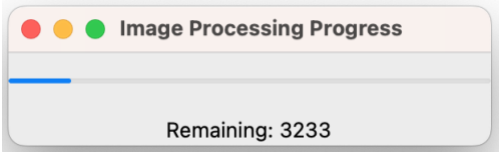
Jak již však bylo řečeno, stejného principu jako u liturgických knih lze využít pouze z části. Pokud by anotátor v těchto případech využil mód s asistencí na více druhů dokumentů, moc by mu nepomohl. Naopak by samotná anotace velmi pravděpodobně trvala mnohem déle než s využitím módu manuálního. Nejvýhodnější je tedy nechat model celou sadu zkoumaných vzorků anotovat automaticky. Z tohoto důvodu jsme do představovaného softwaru implementovali poslední mód, a to plně automatický.



Obr. 8 Nastavení vstupních a výstupních složek – automatický mód

V tomto případě však plní odlišnou roli než módy předešlé. To je důležité zmínit především z toho důvodu, že plně automatická anotace, jak jsme již poznamenali, není vhodným řešením. Jelikož se zde spoléháme především na přesnost správně negativních výsledků, konečným výstupem nebude hotová datová sada. Tento postup můžeme nazývat spíše preanotací, protože díky vyloučení správně negativních výsledků zůstane menší množina správně pozitivních a špatně pozitivních vzorků. Tímto způsobem tedy můžeme potřebnou část na anotaci výrazně zmenšit i u komplexnějších druhů dokumentů, což jsme využili například na datovou sadu vybraných periodik o velikosti 10455 stran. Po dokončení automatického módu nám zůstalo 6724 správně negativních stran, 5 špatně negativních a 3726 správně a špatně pozitivních. Místo celé datové sady nám tedy stačilo manuálně anotovat přibližně jednu třetinu.

Ačkoliv jsme nemohli využít výhody módu s asistencí, i tak nám software tímto způsobem ušetřil mnoho času.



Obr. 9 Průběh procesu automatického módu

FLEXIBILITA MODELU

Rozhodneme-li se pro tvorbu obrazové datové sady použít námi představovaný software, můžeme si určit, jakým způsobem chceme k problematice přistupovat. V případě, že bude zkoumanou problematikou prvek v určitém druhu dokumentu, můžeme s relativně malým množstvím dat natrénovat zmiňovaný model a využít módu s asistencí k efektivní tvorbě větší datové sady. Pokud bude zkoumanou problematikou prvek v širokém spektru různých druhů dokumentů, můžeme stejným způsobem natrénovat model na menším vzorku dokumentů podobného druhu a využít automatického módu k separaci správně negativních vzorků od ostatního. Pro tvorbu menších datových sad – například na natrénování základního modelu softwaru – pak můžeme využít manuální mód. Zde je nicméně potřeba zmínit, že úspěšnost modelu je přímo závislá na druhu sledovaného prvku. Ačkoliv byla architektura tvořena s důrazem na flexibilitu a podle dosavadních testů je model velmi přesný vždy alespoň jako preanotační nástroj, může se stát, že na nějaký druh prvku vhodný nebude. V takovém případě je však velmi jednoduché implementovat do softwaru model jiný. V kódu softwaru je z tohoto důvodu zaimplementován pouze jako cesta k modelu a není závaznou podmínkou pro fungování (chyba se objeví až u samotné anotace pomocí jednoho z módů s modelem). Je tedy teoreticky možné vytvořit mnohem komplexnější architekturu, či natrénovat hlubší a širší umělou neuronovou síť, která bude v asistenčním módu fungovat i u složitějších prvků. To již však z našeho pohledu přesahuje zamýšlené využití celého systému a v některých případech může být tento způsob kontraproduktivní.

ZÁVĚR

Digitální knihovna je díky metadatům, která jsou součástí dokumentů, velmi dobrým zdrojem pro tvorbu některých druhů datových sad. Pomocí standardních nástrojů, jako je například REST API, může být i samotná tvorba těchto datových sad jednoduchým procesem. Pokud se však zaměříme na prvky, které nelze jednoduše identifikovat, musíme hledat způsoby jiné. Představovaný software tuto funkci plní velmi dobře. Jak jsme si ukázali, správné využití modelu umělé neuronové sítě může pomoci celý proces urychlit a zmenšit množství potenciálních chyb. Námi navržená architektura navíc umožňuje pomocný model naučit rozeznávat hledané prvky již s relativně malým počtem potřebných vzorků, které lze pomocí manuálního módu jednoduše získat.

Jako inovativní považujeme především funkci anotačního módu s asistencí, která umožňuje využít jak lidského anotátora, tak strojovou kontrolu v jeden okamžik, a díky tomu zkrátit celý anotační proces. Úspěšná byla také implementace automatického módu, který sice sám nevytvoří celou datovou sadu, ale umožňuje zredukovat významným způsobem manuální anotaci. To nám otevírá prostor pro další práci v této oblasti. Za důležitou budoucí funkci považujeme například možnost natrénovat zmiňovaný model přímo prostřednictvím softwaru. Anotátor by mohl buď použít již existující datovou sadu, anebo by pomocí nového „učícího“ módu nejprve anotoval vzorky manuálně, dokud by software nerozhodl, že je pro tento typ architektury již vzorků dostatek a model by se následně začal automaticky učit – tím by se zpřístupnil asistenční a automatický mód.

Dalším způsobem, jak urychlit celý proces, je umožnit zobrazení více vzorků najednou. Ačkoliv je podstata tohoto nového přístupu velmi jednoduchá, existuje mnoho možností, jak tento mód implementovat. Jednou z variant může být například možnost vidět následující vzorek, anebo naopak vidět hned několik vzorků bez předem daného pořadí. Co je z pohledu anotátora jednoduší však již může být v těchto případech spíše otázkou subjektivní preference. Ačkoliv je tu ještě mnoho prostoru na zlepšení a usnadnění celého procesu, cíl vytvořit jednoduchý anotační nástroj, který pomocí modelu umělé neuronové sítě usnadní tvorbu obrazové datové sady, považujeme za splněný.

DEDIKACE

Studie byla publikována v rámci Institucionální podpory na dlouhodobý koncepční rozvoj výzkumné organizace (DKRVO) – Moravská zemská knihovna v Brně.

SEZNAM POUŽITÉ LITERATURY

API Specifications—International Image Interoperability Framework™. (b.r.). Získáno 20. červenec 2023, z https://iiif.io/api/

API v7 · ceskaexpedice/kramerius Wiki. (b.r.). Získáno 20. červenec 2023, z https://github.com/ceskaexpedice/kramerius/wiki/API-v7

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Meier, B., Stadelmann, T., Stampfli, J., Arnold, M., & Cieliebak, M. (2017). Fully Convolutional Neural Networks for Newspaper Article Segmentation. *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, 414–419. https://doi.org/10.1109/ICDAR.2017.75

Northcutt, C. G., Athalye, A., & Mueller, J. (2021). *Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks* (arXiv:2103.14749). arXiv. http://arxiv.org/abs/2103.14749

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2017). Snorkel: Rapid Training Data Creation with Weak Supervision. *Proceedings of the VLDB Endowment. International Conference on Very Large Data Bases*, *11*(3), 269–282. https://doi.org/10.14778/3157794.3157797

Ratner, A., De Sa, C., Wu, S., Selsam, D., & Ré, C. (2017). *Data Programming: Creating Large Training Sets, Quickly* (arXiv:1605.07723). arXiv. http://arxiv.org/abs/1605.07723

Registr Krameriů. (b.r.). Získáno 19. červenec 2023, z https://registr.digitalniknihovna.cz/

Ying, X. (2019). An Overview of Overfitting and its Solutions. *Journal of Physics: Conference Series*, *1168*(2), 022022. https://doi.org/10.1088/1742-6596/1168/2/022022

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2017). *Understanding deep learning requires rethinking generalization* (arXiv:1611.03530). arXiv. http://arxiv.org/abs/1611.03530

POZNÁMKA O AUTOROVI

Filip Jebavý

Filip Jebavý se zabývá problematikou analytických schopností umělých neuronových sítí. V této oblasti se též podílí na několika výzkumných projektech se zaměřením na humanitní vědy a strojové učení. V současnosti pracuje jako vedoucí Odboru správy digitálních dokumentů v Moravské zemské knihovně v Brně.

E-mail: jebavy@mzk.cz

ORCID: https://orcid.org/0009-0004-9448-4380