

Přepis ozdobných nápisů

Ondřej Dacík (xdacik01)¹, Kryštof Andrýsek (xandry12)¹

¹ Faculty of Information Technology Brno University of Technology

Abstrakt

Následující práce se zabývá přepisem textu z obrázků z reálného světa. Jedná se o úlohu STR (Scene Text Recognition) se zaměřením na ozdobné, umělecké nápisy, které jsou často obtížně čitelné, ať z důvodu rozmazání, zakřivení nebo celkové komplexnosti textu. V rámci této práce analyzujeme existující řešení se zaměřením na soutěž ICDAR24. Vyhodnocujeme řešení, která jsme vyzkoušeli, počínaje využitím konvolučních neuronových sítí s LSTM a CTC loss. Více se poté zaměřujeme na model PARSeq založený na transformerové architektuře. Představujeme experimenty, které jsme s ním provedli, mimo jiné natrénování od nuly na český jazyk a mezery. Výsledné řešení vyhodnocujeme metrikami přesnosti a editační vzdálenosti a nabízíme srovnání, jak s originálním modelem, tak jinými přístupy řešení STR úloh.

Úvod

Schopnost automaticky rozpoznávat a přepisovat text v reálných obrázcích hraje stále důležitější roli v mnoha praktických aplikacích. Ať se jedná o čtení dopravních značek, až po digitalizaci ručně psaných dokumentů. Zatímco klasické OCR (Optical Character Recognition) přístupy dosahují velmi dobrých výsledků na naskenovaných dokumentech s dobře čitelným a předvídatelným písmem, úlohy STR (Scene Text Recognition) přináší mnohem náročnější výzvy.

Text v reálných scénách může být deformovaný, zakřivený, částečně zakrytý, rozmazaný, v různých typech písma a barevných kombinacích, a často se vyskytuje na nestandardních pozadích.

V této práci se zabýváme právě touto náročnější variantou rozpoznávání textu, konkrétně potom rozpoznáváním uměleckých a ozdobných nápisů, jako jsou například cedule u obchodů. Cílem je vytvořit model, který je schopen takové texty číst a převádět je do digitální textové podoby.

V rámci řešení jsme analyzovali existující přístupy, experimentovali s různými architekturami a nakonec se zaměřili na moderní transformerové modely, které se pro tuto úlohu ukázaly jako velmi účinné.

Definice úlohy

Tato práce se zabývá tvorbou modelu schopného řešit úlohu Scene Text Recognition (STR), jedná se o rozpoznávání textu z oblastí detekovaných v obrázcích reálného světa (např. značky, etikety, nápisy, umělecké nadpisy atd.). Je to úloha náročnější než běžné OCR, neboť text může být zkreslený, nakloněný, různých stylů, barev a často nečitelný kvůli šumu nebo překážkám. Vstupem modelu je obrázek obsahující výřez s nápisem, který má být přečten. Jeho výstupem je přepis textu.

Existující řešení

V rámci této kapitoly jsou prezentována existující řešení, která dokáží číst alespoň do určité míry umělecké texty a nápisy. Jednotlivá řešení jsou rozdělena do tří podsekcí. První z nich se zaměřuje na modely, které jsou obecně schopné číst velké množství textů, ale jsou zaměřené spíše všeobecně, nikoliv na konkrétní úlohu čtení extrémně složitých nápisů, jak je tomu u druhé sekce. Třetí sekce pak pouze velmi stručně představuje další jedince, které budeme využívat pro srovnání v tabulce 2 na konci této práce.

Standardní přístupy

Prvním zástupcem této kategorie je model TrOCR[1] od společnosti Microsoft. Je to model založený na transformerové architektuře využívající enkodér a dekodér. Skládá se z vizuálního enkodéru, který zpracovává obrazové vstupy ve formě patchů (podobně jako Vision Transformer – ViT), a jazykového dekodéru, generujícího sekvenci textových tokenů na základě vizuálních reprezentací a předchozích predikcí. Enkodér je inicializován s využitím modelů typu DeiT a Beit. Dekodér pak přímo využívá předtrénované modely RoBERTa nebo MiniLM. Model byl natrénován na velkém množství syntetických dat a poté dotrénován na konkrétní úlohy[1].

Druhým zástupcem je Gemini[2], tento model je vyvinutý společností Google DeepMind. Je založený na architektuře Mixture-of-Experts Transformer. Využívá podmíněného výpočtu, čímž dosahuje vysoké efektivity při zachování výkonu. Podporuje vstupy o délce až 10 milionů tokenů, což umožňuje práci s velmi dlouhými texty, kódem nebo obrázky. V oblasti čtení uměleckých textů kombinuje OCR schopnosti s kontextovým jazykovým

porozuměním, čímž umožňuje přímou interpretaci i obtížně čitelných textů[2].

Specializované přístupy

Tato sekce se zabývá modely, které vznikly v rámci soutěže ICDAR 2024[3]. Soutěž se zabývala rozpoznáváním uměleckých jednoslovných textů.

Prvním zástupcem je model, který se v této soutěži umístil na prvním místě. Jedná se o model Ocr For WordArt. Tým, který vytvořil tento model, využil kombinaci supervizovaného, semi-supervizovaného i self-supervizovaného učení. Autoři předpokládali, že takto obtížné texty nelze rozpoznávat pouze s jedním modelem, takže natrénovali celkem 4 modely, kde každý provádí predikci zvlášť a finální predikce se pak vybere na základě hlasování[3].

Dalším modelem je ViettelAi-OCR (druhé místo), využívající architekturu PARSeq s ViT a VIPTR enkodéry. PARSeq využívá permutační jazykové modelování, při němž se model učí z náhodně přeházené sekvence tokenů. Model byl natrénován na milionech vzorků z datových sad textových obrázků (Union14M-L) a bylo zde využito self-supervizované učení pomocí metody Character-to-Character Distillation, která vylepšuje schopnost modelu odlišit text od pozadí a udržet konzistentní reprezentaci znaků při různých transformacích[3].

Další zástupci

Mezi další zavedené modely pro rozpoznávání textu z obrázků patří ABINet¹, využívající paralelní jazykový model. Dalšími zástupci jsou TRBA² kombinující konvoluční, rekurentní přístup a attention vrstvy, nebo také ViTSTR-S³ fungující hlavně na vizuálních transformerech. Posledním příkladem je potom CRNN⁴, to jest klasický přístup založený na kombinaci CNN a RNN pro extrakci znaků v posloupnosti.

Popis řešení

V této kapitole jsou prezentována řešení, která jsme použili. V úvodu je krátká sekce zabývající se triviálními přístupy, které se ukázaly jako nepříliš dobré. V druhé části je pak už prezentována výsledná architektura modelu, který byl trénován, a model, který byl dotrénován.

¹ <https://github.com/FangShancheng/ABINet>

² <https://github.com/clovaai/deep-text%2Drecognition-benchmark>

³ <https://github.com/roatienza/deep-text-recognition-benchmark>

⁴ <https://github.com/meijieru/crnn.pytorch>

Jednodušší architektury

Nejprve jsme zkoušeli pro přepis textů využít konvolučních neuronových sítí v kombinaci s LSTM a CTC loss funkcí. Struktura modelu byla následující:

- Konvoluční neuronová síť – Z obrázků extrahovala příznaky, jako jsou okraje a tvary charakteristické pro jednotlivá písmena, výstupem byla tedy feature mapa.
- Long Short-Term Memory (LSTM) – Tato vrstva modelovala sekvence a umožňovala rozpoznat pořadí znaků s využitím kontextu mezi jednotlivými znaky.
- Connectionist Temporal Classification loss (CTC) – Pomocí této loss funkce byl model trénován. CTC je pro OCR úlohy užitečná zvláště díky tomu, že umožňuje pracovat s texty různých délek. Má ovšem i jisté problémy, jako například přepis slov s opakujícími se znaky.

Výsledky natrénovaného modelu byly ovšem relativně špatné, což bylo zřejmě způsobeno i tím, že jsme ze začátku ještě neměli dostatečné množství trénovacích dat. Rozhodli jsme se tedy zkusit nějaké existující robustnější řešení, konkrétně transformery. Ostatně právě tento přístup dobře fungoval pro většinu modelů s velmi dobrými výsledky.

PARSeq

Následující sekce prezentuje výsledné řešení problému, které jsme zvolili. Využili jsme architekturu a model PARSeq[4] navržený pány Darwinem Bautistou a Rowelem Atienza. V první části je prezentována hlavní myšlenka tohoto přístupu, a později je i stručněji shrnuta architektura.

Hlavní myšlenka Běžně se při řešení těchto úloh využívá auto regresivního přístupu, kde každý znak závisí na tom předchozím. To ovšem způsobuje nevýhody, jsou jimi směrová závislost a pomalejší dekodování. PARSeq tento problém řeší tím, že při učení využívá permutační jazykové modelování (PLM). Dále podporuje více způsobů dekodování (NAR, AR, cloze). Také využívá iterativního vylepšování (iterative refinement), které spočívá v tom, že model opakovaně předvídá to stejné slovo, přičemž v každé další předpovědi už využívá informace získané v té předchozí. Jedná se tedy o formu oboustranného kontextu.[4]

Architektura PARSeq hojně využívá Multi-head Attention (MHA) přístup. Kde q , k a v označují query, key, value a m značí volitelnou attention masku. Architektura vychází z enkodér-dekodér přístupu, který lze vidět na obrázku 1. Enkodér se skládá z dvanácti vrstev, zatímco dekodér je tvořen pouze

jednou vrstvou. Tento přístup byl zvolen pro minimalizaci výpočetního času, aniž by příliš ovlivňoval kvalitu výsledného modelu[4].

Enkodér je založen na Vision transformeru ViT. ViT vrstva obsahuje jeden MHA modul používaný pro self-attention. Samotný enkodér se poté skládá z dvanácti ViT vrstev bez klasifikační hlavy a [CLS] tokenů. Vstupní obrázek je převeden na stejně velké obdélníkové „patche“, které jsou následně projektovány na vektory pomocí naučené matice embeddingů. K tokenům jsou přičteny naučené poziční embeddingy před prvním průchodem ViT[4].

Jako dekodér se využívá Visio-Lingual Decoder, jedná se o speciálně navržený transformer dekodér, který využívá pre-LayerNorm architekturu a dvojnásobný počet attention hlav (head) oproti běžnému nastavení. Přijímá tři hlavní vstupy: poziční tokeny, které určují, na jaké pozici má být výstup předpovězen, kontextové tokeny s pozicemi a obrázkové tokeny z enkodéru. První attention blok propojuje pozice a kontext (context-position attention), čímž se model učí předpovídat na základě pozice nezávisle na pořadí kontextu, což výrazně napomáhá kvalitě modelu. Druhý blok přidává obrazové informace pomocí attention vrstvy mezi výstupem z předchozí vrstvy a obrazovými tokeny (image-position attention). Maska je v dekodéru flexibilní, to znamená, že během trénování se používají náhodné permutace, při inferenci pak buď autoregresivní maska, cloze maska pro iterativní zpřesnění výstupu, nebo žádná maska pro ne autoregresivní dekodování. Tento design umožňuje dekodéru efektivně kombinovat jazykové a vizuální informace a zároveň podporuje různé přístupy generování textu[4].

Metriky pro hodnocení modelu

Pro vyhodnocení modelu byly využity dvě základní metriky: klasická přesnost a normalizovaná editační vzdálenost (NED).

- **Přesnost**

Přesnost je základní metrikou pro hodnocení modelů, která měří podíl správných výstupů vůči celkovému počtu výstupů. Je tedy definována jako:

$$\text{Přesnost} = \frac{\text{Počet správně rozpoznaných textů}}{\text{Celkový počet textů}} \times 100$$

- **Normalizovaná editační vzdálenost**

Druhou metrikou je normalizovaná editační (Levenštejnova) vzdálenost (NED), která udává podobnost dvou řetězců na základě minimálního počtu operací (vlození, smazání nebo nahrazení znaku), které jsou potřeba k přechodu z jednoho

textu na druhý. Je definována jako Levenštejnova vzdálenost mezi predikovaným a referenčním řetězcem dělená délkou delšího z nich:

$$\text{NED} = \frac{\text{Levenštejnova vzdálenost}}{\text{Délka delšího ze dvou řetězců}}$$

Normalizovaná editační vzdálenost takto nabývá hodnoty 0, pokud jsou řetězce identické a hodnoty 1, pokud jsou maximálně rozdílné (neshodují se ani v jednom znaku). Na základě NED je pak možné definovat procentuální přesnost predikce následovně:

$$\text{Přesnost predikce} = 100 \times (1 - \text{NED})$$

Experimenty

Následující kapitola shrnuje experimenty, které jsme v rámci naší práce provedli. Experimenty jsou prezentovány v pořadí, jakém jsme je prováděli. Kapitola začíná dotrénováním modelu a později je prezentováno i trénování našeho nového modelu úplně od začátku. Ve svém závěru se zabývá dotrénováním našeho modelu na historických datech.

Dotrénování původního PARSeq modelu na datové sadě Wordart

V rámci prvního experimentu jsem se pokusil o fine-tuning předtrénovaného modelu PARSeq⁵ datovou sadou Wordart-v1.5⁶. Tato sada obsahuje 6000 anotovaných uměleckých nápisů a byla použita v soutěži rozpoznávání uměleckých textů ICDAR 2024[3]. Cílem bylo zvýšit úspěšnost rozpoznávání graficky stylizovaných textů.

Fine-tuning modelu s dávkami o velikosti 128 trval 5 epoch a výsledky nebyly příliš uspokojivé. Na obrázku 2 lze vidět, že validační přesnost značně kolísala a celkově nevykazovala jasný trend zlepšení. Na testovací sadě datasetu Wordart se model také příliš nezlepšil, přesnost zůstala stejná (86.96 %) a metrika normalizované editační vzdálenosti (1-NED) se pouze nepatrně zlepšila z 95,90 na 96,06. Na ostatních testovacích sadách přesnost o několik procent poklesla.

⁵ <https://github.com/baudm/parseq>

⁶ <https://sites.google.com/view/icdar-2024-competition-wordart/>

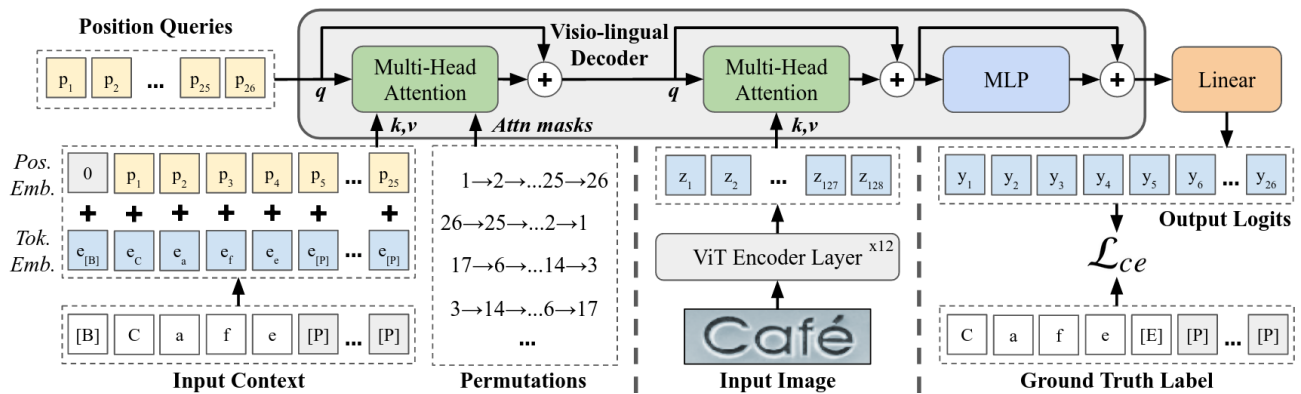


Figure 1: Schéma znázorňuje architekturu a trénování modelu PARSeq, kde poziční tokeny slouží jako dotazy i embeddingy, attention masky se používají jen pro context–position attention a [B], [E], [P] označují speciální tokeny.[4]

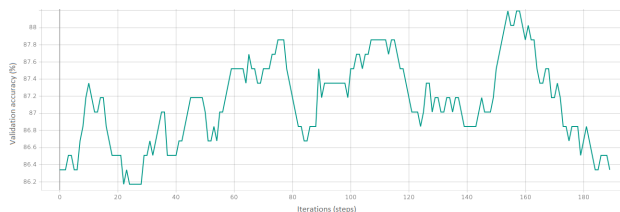


Figure 2: Graf validační přesnosti při dotrénování modelu PARSeq na datové sadě Wordart

Experiment lze považovat za neúspěšný, protože nedošlo k výraznějšímu zlepšení u žádné datové sady. Jedním z důvodů může být fakt, že už výchozí přesnost na datové sadě Wordart byla relativně vysoká a model tak neměl dostatečný prostor pro zlepšení.

Adaptace modelu pro přepis českých textů s mezerami

OCR modely jsou typicky trénované na specifických znakových sadách, což způsobuje potíže s generalizací na textech jiných jazyků. V tomto experimentu jsme se tedy zaměřili na rozpoznávání českých textů, včetně mezer mezi slovy. Předtrénovaný model PARSeq byl sice trénován na velkém množství reálných i syntetických dat, ale s omezenou znakovou sadou, která obsahovala pouze číslice, anglickou abecedu a speciální znaky (mimo mezeru). Model se tak naučil jak mezery, tak diakritiku úplně ignorovat.

Pro tento experiment jsme vytvořili syntetickou datovou sadu⁷ obsahující 100 000 nápisů, skládající se z náhodných českých slov. Nápisů mají různé fonty, barvy a pozadí.

Dotrénování původního PARSeq modelu českou syntetickou datovou sadou

Nejprve jsme se pokusili o fine-tuning předtrénovaného modelu s rozšířenou znakovou sadou pomocí nové syntetické datové sady. Změna znakové

sady je ovšem při fine-tuningu problematická, jelikož výstupní klasifikační vrstva modelu má pevně daný počet tříd podle původní sady znaků a po přidání nových znaků pro diakritiku a mezery se rozměry vah neshodují. Je tedy nutné nahradit výstupní vrstvu (head) a její váhy přetrénovat od začátku.

Při fine-tuningu jsme tedy při načítání předtrénovaného modelu vynechali výstupní hlavu a trénovali 20 epoch na syntetické datové sadě. Výsledky ovšem byly velmi špatné, model se musel opět učit mapování jednotlivých znaků, což vedlo k částečné ztrátě jeho původních schopností. Také se na syntetické sadě zřejmě přetrénoval, takže ačkoliv testovací přesnost na syntetické datové sadě dosáhla 72.55 %, u ostatních datových sad se úspěšnost propadla na průměrných 50 %, v případě sady Wordart dokonce až na 30 %. Vzhledem k tomuto neúspěchu jsme se rozhodli změnit přístup a model natrénovat od začátku.

Trénování modelu na český jazyk a mezery od nuly

Druhým pokusem bylo natrénování modelu od začátku už s rozšířenou sadou znaků. K tomu jsme zkombinovali naši syntetickou datovou sadu s těmi, které byly použity k trénování modelu PARSeq: COCO, RCTW17, Uber, ArT, LSVT, MLT19, ReCTS, TextOCR a OpenVINO⁸.

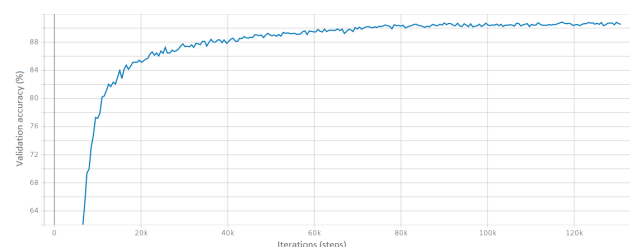


Figure 3: Graf ukazuje validační přesnost při trénování.

Model jsme trénovali celkem 10 epoch. V tabulce 1

⁷ https://drive.google.com/drive/folders/1Aqv98d605m63FeysN_s6VxpA8n1GTKJ

⁸ <https://github.com/audm/parseq/blob/main/Datasets.md>

Datová sada	Přesnost (%)		1 - NED	
	Parseq	My	Parseq	My
IIIT5K	98.93	95.77	99.76	98.57
SVT	98.30	96.29	99.54	98.82
IC13_857	98.25	95.22	99.61	98.68
IC15_1811	90.72	85.04	97.22	94.98
SVTP	95.04	93.02	98.56	97.78
CUTE80	98.96	94.79	99.69	98.05
Synth100k	8.23	74.08	70.80	93.87
Noviny	26.32	22.73	75.64	72.87
Wordart	86.96	78.33	95.90	92.43

Table 1: Srovnání výsledného modelu z druhého experimentu a původního modelu PARSeq na různých datových sadách.

Lze vidět, že testovací přesnost výsledného modelu na datových sadách přibližně o 4 % poklesla, u sady Wordart pokles činil 8 %. Přesnost na syntetické sadě však dosáhla 74 %, což je lepší výsledek, než v předchozím pokusu o přímý fine-tuning se syntetickou sadou. Je nutné vzít v potaz i to, že pokles u datových sad bez češtiny se dal očekávat, jelikož původní předtrénovaný PARSeq nemohl být zvláště u uměleckých nápisů zmaten různými čarami, které připomínají diakritiku.

Experiment lze jednoznačně považovat za úspěch. Výsledný model úspěšně rozpoznává texty s mezery a diakritikou a zároveň si zachoval velmi dobrou úspěšnost na testovacích sadách. Porovnání s původním modelem PARSeq a dalšími na konkrétních příkladech lze vidět v tabulce 2.

Dotrénování nového modelu na texty z historických novin

Jako poslední experiment jsme se pokusili zlepšit úspěšnost českého modelu z předchozího experimentu na textech z historických časopisů a novin. Historická datová sada, kterou jsme měli k dispozici, obsahovala přibližně 18 000 neanotovaných výstřížků historických nápisů. Ručně jsme jich přibližně 2000 anotovali a rozdělili na trénovací, validační a testovací sadu v poměru 8:1:1. Na této menší sadě jsme pak 4 epochy trénovali český model z druhého experimentu.

Model	Přesnost (%)	1-NED
PARSeq	26.32	75.64
Český model	22.73	72.87
Dotrénovaný český model	58.08	87.51

Table 3: Tabulka ukazující přesnost a 1-NED metriku modelů na datové sadě historických novin.

Výsledky lze vidět v tabulce 3. Úspěšnost

dotrénovaného modelu na historické sadě značně vzrostla a model byl schopen rozpoznat i komplikovanější texty. Podle očekávání ovšem lehce poklesla úspěšnost u ostatních datových sad, v průměru o necelá 2 %. Vzhledem k zaměření na historické české texty i tak lze experiment považovat za relativní úspěch.

Závěr

V rámci této práce se nám podařilo seznámit se s možnými architekturami řešícími úlohu STR. Vytvořili jsme syntetickou datovou sadu vhodnou pro natrénování modelu řešícího obtížnou úlohu přepisu těžko čitelných ozdobných a uměleckých textů. Provedli jsme několik experimentů. Prvním z nich bylo dotrénování modelu na datové sadě Wordart. Ten nám ukázal, že původní model PARSeq je už natolik kvalitní, že pouhé dotrénování nepřinese výrazné zlepšení. Dále se nám podařilo natrénovat model, který je na rozdíl od modelu PARSeq schopen rozpoznávat mezery a taktéž dokáže přepisovat texty v českém jazyce. Nakonec jsme provedli dotrénování našeho nového modelu na datové sadě historických textů, čímž se nám podařilo značně zlepšit úspěšnost na odpovídající testovací sadě.

Co se dalšího výzkumu týče, je zde prostor pro natrénování robustnějšího modelu s větší datovou sadou pro český jazyk. Dále by tréninku napomohlo opravení a upravení datových sad používaných při tréninku celého modelu, neboť některé neobsahují mezery a vyskytují se zde také chyby. Pro zlepšení výsledků při čtení novinových článků by bylo vhodné datovou sadu lépe pročistit a taktéž pro větší přesnost natrénovat model na dalších jazycích, jako je ruština nebo staré německé písmo švabach.

Poděkování

Děkujeme MetaCentru⁹ za poskytnutí prostředků, které nám umožnily vytvořit tuto práci.

Zdroje

- [1] Minghao Li et al. "TrOCR: Transformer-based Optical Character Recognition with Pre-trained Models". In: *CoRR* abs/2109.10282 (2021). arXiv: 2109.10282. URL: <https://arxiv.org/abs/2109.10282>.

⁹ <https://metavo.metacentrum.cz/cs/>



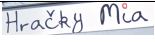






Obrázek	Skutečný text	Náš model	Parseq	ABINet	TRBA	ViTSTR-S	CRNN
	Fajntričko	Fajntričko	Fajntricko	Fayntricho	Fojntricko	Fajntricho	Fapabahe
	NašeHory	našeHory	naseHory	naseHory	nascHory	naseHory	nasellery
	Hračky Mia	Hračky Mia	HrackyMia	HrackyMia	HrackyMia	HrackyMia	HrackyMia
	LYŽE KOLA MO-TOOBLEČENÍ	LYŽE KOLA MO-TOOBLEČENÍ	LYZEKOLAINO-TOOBLECENT	LIYZEROLAMO-OOOOLEEEENN	LV2ENOLA-MOTOGBEEN	LLVIEACCA-IINTOUBLEEEII	ATECANO-TOMACEN
	AUTO SERVIS	AUTO SERVIS	AUTOSEIVIS	AUTOSEEnS	HUTOSENIS	AUTOSENVs	MUTOSEnS
	Verbandstoffe	Verbandsteffe	Verbandsteffe	Verbandsteffe	Verbandstelle	Verbandsteffe	Verbandsleffe
	HOUSTON	HOUSTON	HOUSTON	HOUSTON	HOUSTON	MOUSTON	HOLSICH
	Kappa	Kaopa	Kappa	Kappa	Kaspa	Kappa	Kaada
	3rd Ave	3rd Ave	3rdAve	3=-Ave	3rdAve	3rdAve	Coke

Table 2: Porovnání OCR výstupů různých modelů na příkladech i s českými slovy. Srovnávané modely i obrázky ‘Verbandstoffe’, ‘Kappa’ a ‘3rd Ave’ byly převzány z PARSeq[4] github repozitáře.

- [2] Gemini Team et al. *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. 2024. arXiv: 2403.05530 [cs.CL]. URL: <https://arxiv.org/abs/2403.05530>.
- [3] Xudong Xie et al. “ICDAR 2024 Competition on Artistic Text Recognition”. In: *Document Analysis and Recognition - ICDAR 2024*. Ed. by Elisa H. Barney Smith, Marcus Liwicki, and Liangrui Peng. Cham: Springer Nature Switzerland, 2024, pp. 301–314. ISBN: 978-3-031-70552-6.
- [4] Darwin Bautista and Rowel Atienza. “Scene Text Recognition with Permuted Autoregressive Sequence Models”. In: *European Conference on Computer Vision*. Cham: Springer Nature Switzerland, Oct. 2022, pp. 178–196. doi: 10.1007/978-3-031-19815-1_11. URL: https://doi.org/10.1007/978-3-031-19815-1_11.