Vilniaus universitetas

Ataskaitinė informatikos krypties doktorantų konferencija 2024-03-28

Rolandas Gricius (VU DMSTI doktorantas, Išmaniųjų technologijų tyrimų grupė)

Preliminari darbo tema.

Turinio atpažinimas suskaitmenintuose struktūrizuotuose dokumentuose.

Recognising the contents in digitised structured documents.

Darbo vadovas.

Prof. dr. Igoris Belovas.

Doktorantūros studijų laikotarpis.

2021 m. spalio mėn. 1 d. – 2025 m. rugsėjo mėn. 30 d.

Ataskaitinis laikotarpis.

2023 m. spalio mėn. 1 d. – 2024 m. kovo mėn. 31 d.

Visų studijų planas ir jo vykdymo suvestinė

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2021/2022)	2	3
II (2022/2023)	2	1
III (2023/2024)		
IV (2024/2025)		
Iš viso:	4	4

Studijų metai	Dalyvavimas konferencijose				Publikacijos					
	Tarptautinėse		Nacionalinėse		Su citav. rodikliu			Be citav. rodiklio		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Planas	Įvykdyta	Būklė
I (2021/2022)				1						
II (2022/2023)	1	1					Įteikta			
III (2023/2024)					1		Įteikta			
IV (2024/2025)	1				1					
Iš viso:	2	1		1	2					

Vilniaus universitetas

Ataskaitinio pusmečio darbo planas ir jo vykdymo suvestinė

Egzaminai 2023/2024 (I pusmetis)		
Planas	Įvykdyta	Būklė
-	-	-

Dalyvavimas konferencijose 2023/2024 (I pusmetis)		
Planas	Įvykdyta	Konferencijos tipas
-	-	-

Publikacijos 2023/2024 (I pusmetis)			
Planas	Įvykdyta	Būklė	Publikacijos tipas
Egyptian Informatics Journal	R. Gricius, I. Belovas. On the Generation of Synthetic Invoices for Training Machine Learning Models. Journal of King Saud University - Computer and Information Sciences.	Įteikta: 2023-11-30	Žurnalas turi cituojamumo rodiklį (impact factor) CA WoS duomenų bazėje.

Vilniaus universitetas

Informacija apie tarptautinius renginius ir publikacijas, kuriose pateikti pagrindiniai disertacijos rezultatai

Dalyvavimas tarptautinėse konferencijose	
	Aprašas
1.	R. Gricius, I. Belovas "Generation of Synthetic Invoices for the Training of Machine Learning Models". International Conference on Pattern Recognition Applications and Methods (ICPRAM) 2023, Lisabona, Portugalija, 2023-02-22 – 24 d.

	Publikacijos (tik su citavimo rodikliu)	
	Bibliografinis aprašas	Būklė
1.	-	-

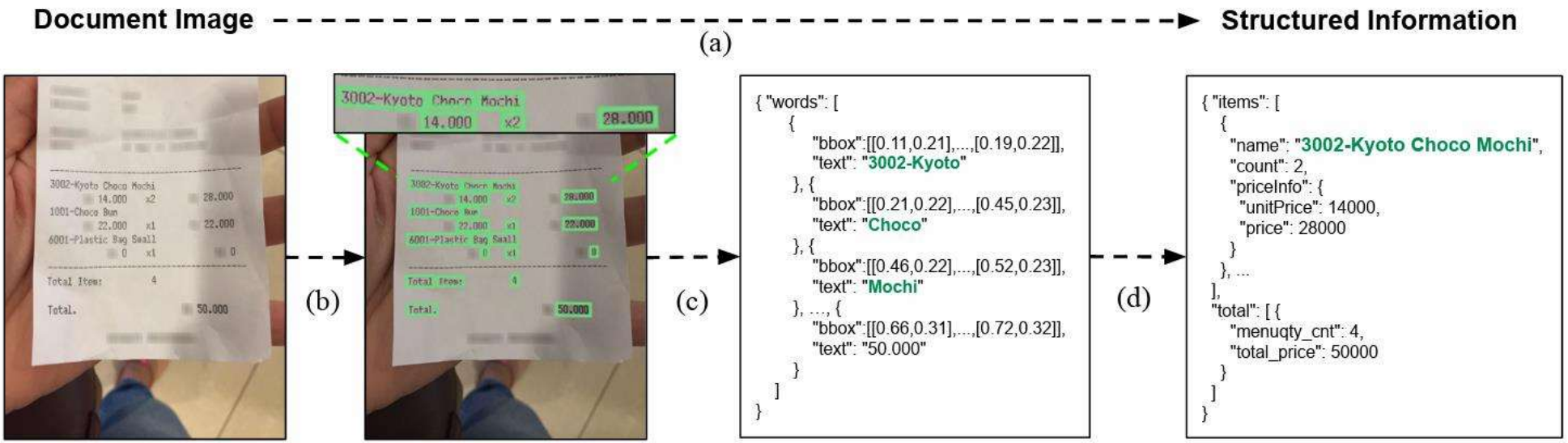
Kita mokslinė ir akademinė veikla

- **Pratybų vedimas MIF bakalauro studentams, kursas "Informacinių sistemų saugos pagrindai", 2023/2024 rudens semestre**
- **Konsultavimas bakalauro studento rašto darbo ruošimo klausimais, tema "*Recognising the contents in digitised financial documents*"**
  - **Svarstoma galimybė pristatyti resultatus LMD ar DAMSS konferencijose**

Tyrimo objektas, tikslas ir uždaviniai

- Tyrimo objektas – teksto esybių atpažinimas pagal tekstą ir jo išdėstymą
- Tikslas – naudojant natūralios kalbos apdorojimo metodus, atpažinti ir ištraukti tolesniam apdorojimui sąskaitos duomenis, reikšmingus:
  - teisėtumui – privalomus pagal teisės aktus duomenis
  - apskaitai – data, pirkėjo ir pardavėjo duomenys, sandorio ir mokesčių sumos
  - sandorio vykdymui – pristatymo duomenys, apmokėjimo detalės
- Uždaviniai – sudaryti duomenų rinkinį tyrimui, atlikti teorinį tyrimą identifikuojant metodus, empirinį tyrimą palyginant jų veikimą ir modifikuoti pritaikant Lietuvos specifikai ir surinktiems duomenims

Klasikinė darbų seka



Vilniaus universitetas

G. Kim et al. OCR-free Document Understanding Transformer. https://doi.org/10.48550/arXiv.2111.15664

Dideli kalbos modeliai (DKM)

- Pastaruosius dvejus metus padidėjo susidomėjimas DKM:
  - Modeliui galima pateikti užduotis ir instrukcijas sklandžia kalba
  - Modeliui galima pateikti dokumento tekstą užduočiai
  - Modelis gali atsakyti sklandžia kalba
  - Yra daugiakalbių modelių
- Iš tikro ne visi modeliai veikia pokalbių (Chat) režimu. Juos galima užklausti specialiu formatu per API

Dideli kalbos modeliai (DKM)

- Modelių išleista gana daug:
  - OpenAI GPT šeima
  - Google T5 šeima
  - Facebook LLaMA šeima
  - Alibaba Qwen šeima
  - Google Gemini šeima
  - ...

Didelių kalbos modelių panaudojimas informacijos ištraukimui

- DKM pateikiamas dokumento tekstas, ir prašoma atsakyti į klausimus apie dokumento turinį
- Yra specializuotų modelių, kurie apmokyti dokumentų tekstais, į mokymo duomenis įtraukiant teksto x, y koordinates dokumente. Pvz. Microsoft LayoutLM modelių šeima
- Jau pasirodė ir multimodalinių modelių (GPT-4V). Jie apmokyti įtraukiant tekstinę ir grafinę informaciją (pikselius) į mokymo duomenis

Atlikti eksperimentai su DKM

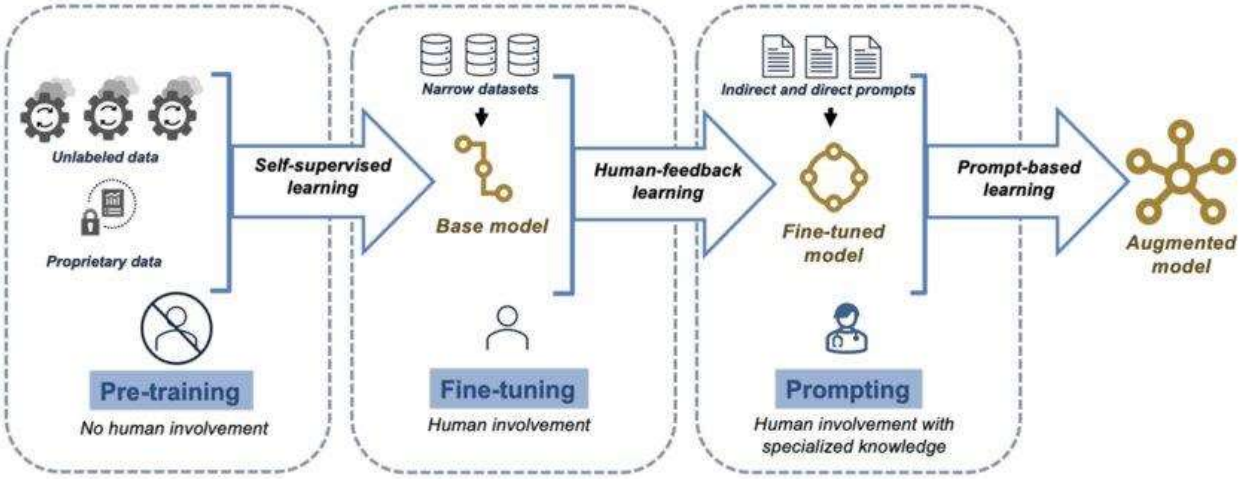
- Modeliai yra įvairių dydžių, maždaug nuo 100 mln parametrų iki 2 trln (10^8 – 2×10^{12})
  - Modelio dydis apytiksliai koreliuoja su naudojamos atminties kiekiu, turint 32GB atminties, galima vykdyti modelius maždaug iki 5-30 bln parametrų. Toliau – naudotas MIF superkompiuteris
- Išbandyti modeliai:
  - T5 modelis supranta tik anglų kalbą. mT5 modifikacija - daugiakalbė.
  - LayoutLM modelis supranta tik anglų kalbą, LayoutXLM modifikacija - daugiakalbė.
  - Multimodalinis Qwen-VL modelis - daugiakalbis.

Atlikti eksperimentai su DKM - rezultatai

- Modelio dydis koreliuoja su rezultatų kokybe
  - Modeliai iki maždaug 10 bln parametrų veikia labai netiksliai - dažniausiai nesugeba atsakyti į klausimą
  - Nuo 30 bln parametrų jau gaunami minimalūs rezultatai, tačiau tai kaip tik riba, kur modelis dar telpa kompiuteryje
  - MIF superkompiuteryje išbandyti didesni modeliai veikė geriau, bet su daug klaidų
  - Parsisiunčiami modeliai yra maždaug iki 500 bln parametrų, didesni modeliai prieinami per API savininkų serveriuose
- Tyrimo kryptis – pagerinti parsisiunčiamų modelių kokybę

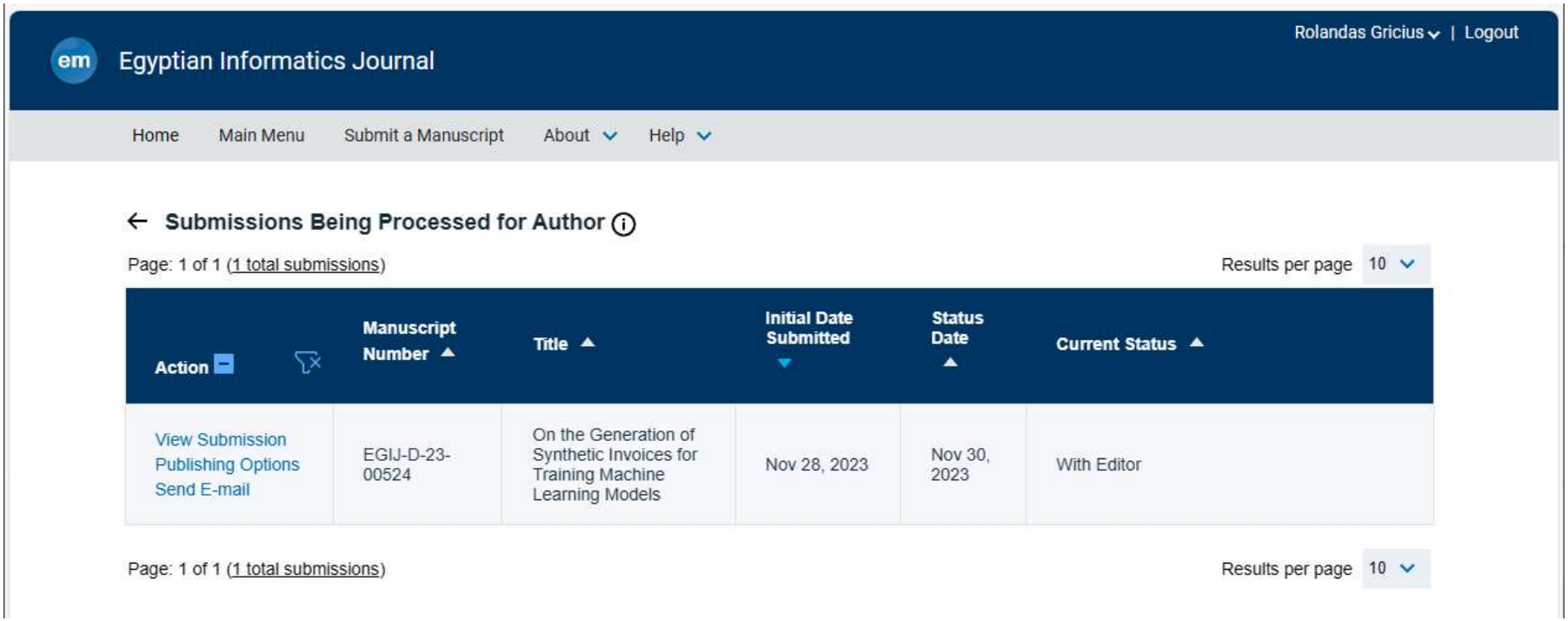
Galimi DKM modifikavimai rezultatams pagerinti

- T5 šeimos modeliams rekomenduojamas papildomas pre-training.
- LayoutLM šeimos modeliams rekomenduojamas fine-tuning.
- Qwen-VL šeimos modeliams rekomenduojamas fine-tuning.



Vilniaus universitetas

Straipsnis - įteikimas



Trumpas per pusmetį gautų mokslinių rezultatų pristatymas

- Išbandytas daugiakalbis Google modelis *mT5*, padarytos išvados jog rezultatams pagerinti tikslinga išbandyti modifikavimą pretraining būdu.
- Išbandyta Microsoft modelių *LayoutLM* šeima, padarytos išvados jog multimodaliniai modeliai veikia geriau nei vien teksto (text-only) modeliai.
- Išbandytas daugiakalbis multimodalinis Alibaba modelis *Qwen-VL*, padarytos išvados jog rezultatams pagerinti tikslinga išbandyti modifikavimą finetuning būdu.
- Įteiktas straipsnis Elsevier leidyklos Web of Science reitinguojamame žurnale *Egyptian Informatics Journal*. R. Gricius, I. Belovas. On the Generation of Synthetic Invoices for Training Machine Learning Models.

Kito pusmečio darbo planas.

1. **Empirinis tyrimas**
   - **Modifikuotų algoritmų eksperimentinis tyrimas analizuojant jų efektyvumą.**
2. **Dalyvavimas 2024-06-11 konferencijoje *Counter Fraud, Cybercrime and Forensic Accounting conference 2024, Portsmouth, UK***
3. **Planuojamas dalyvavimas 2024-09-09 konferencijoje *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD) 2024, Vilniuje***
4. **Pakoreguoti įteiktą straipsnį apie sąskaitų generavimo sprendimą pagal recenzentų atsiliepimus**
5. **Pradėti rengti publikaciją apie sąskaitų atpažinimo sprendimą Web of Science reitinguojamame leidinyje.**

Vilniaus
universitetas

Ačiū už dėmesį

Rolandas Gricius
VU DMSTI doktorantas
rolandas.gricius@mif.stud.vu.lt