trochę o mnie

uczenie maszynowe, przetwarzanie języka naturalnego, neuronowe modele języka, przetwarzanie tekstów historycznych, ewaluacja systemów uczenia maszynowego, przetwarzanie tekstów prawnych i ekonomicznych, folklorystyka komputerowa, OCR

(Arch) Linux, Bash, Docker, Emacs, git(-annex), Haskell, Jenkins, LaTeX, Make/Shake, Python, PyTorch, Solr

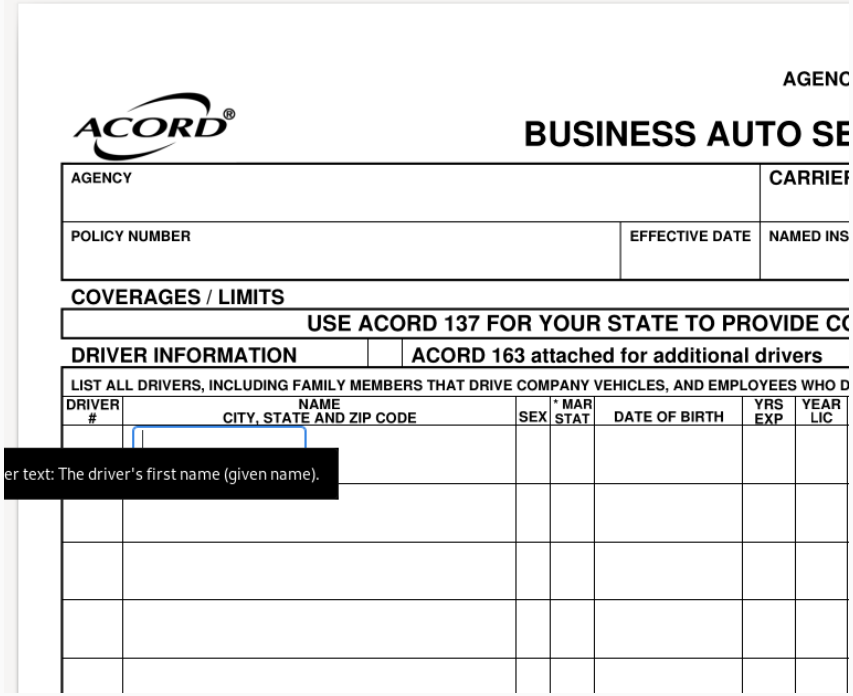
moje projekty otwartoźródłowe: GEval (ewaluacja uczenia maszynowego), Gonito (platforma dla konkursów uczenia maszynowego), Autozoil („sprawdzacz" dla prac w LaTeX-u), Paper Cutter (metaszablon dla prac w LaTeX-u)

była/obecna (współ)praca: PWN.ai, Supermemo, Samsung Electronics R&D, Applica.ai

Propozycje tematów

Uczenie systemów ekstrakcji informacji z formularzy PDF

Pewien odsetek plików PDF-ów zawiera formularze, które mogą być wypełniane elektronicznie. Część z nich to puste formularze, część — wypełnione. Zazwyczaj wypełniane dane są kluczowe z punktu widzenia biznesowego (kwoty, daty, nazwiska itd.), co więcej takie pliki PDF zawierają również przeznaczone dla człowieka etykiety. Można więc wykorzystać te pliki do uczenia systemów wydobywania informacji i odpowiadania na pytania.
Celem projektu jest zebranie zbioru danych i wyuczenie systemu ekstrakcji informacji.



Wydobywanie danych z artykułów naukowych w LaTeX-u

Dla sporej części artykułów naukowych z nauk ścisłych istnieją źródła w formacie LaTeX. Stosunkowo łatwo pozyskać z takich plików dane tabelaryczne. Można w ten sposób uzyskać obszerny zbiór uczący dla wydobywania danych z plików PDF (wygenerowanych z tychże źródeł w LaTeX-u) (1) danych z tabel, (2) danych z tekstu (poprzez wiązanie danych tabelarycznych z ich odniesieniami w tekście).

Model	CORD F1	SROIE F1	DocVQA ANLS	WikiOps Accuracy	RVL-CDIP Accuracy
LayoutLMv2 [57]	96.01	97.81	86.72	—	**95.64**
LAMBERT [12]	96.06	**98.17**	—	—	—
NeOp [2]	—	—	—	59.50	—
TILT-Base	95.11	97.65	83.92	69.16	93.50
TILT-Large	**96.33**	**98.10**	**87.05**	**73.80**	94.02

Table 3. Results of previous state-of-the-art methods in relation to our base and large models. Bold indicates the best score in each category. All results on the test set.

DocVQA. We improved SOTA results on this dataset by 0.33 points. Moreover, detailed results show that model gained the most in table-like categories, i.e., forms (89.5 → 94.6) and tables (87.7 → 89.8), which proved its ability to understand the spatial structure of the document. Besides, we see a vast improvement in the yes/no category (55.2 → 69.0).‡ In such a case, our architecture generates simply *yes* or *no* answer, while sequence labeling based models require additional components such as an extra classification head. We noticed that model achieved lower results in the image/photo category, which can be explained by the low presence of image-rich documents in our datasets.

Co podnieca RoBERTę?

Mając wytrenowany model CNN można próbować wyuczyć obraz wejściowy, który maksymalnie pobudza jakiś fragment sieci. Prowadzi to do interesujących wyników. Z tekstem jest trudniej, ale czy mając pretrenowaną RoBERTę można wymyślić sposób generowania jak najbardziej sensownych tekstów, powodujących pobudzenie określonych warstw?
Efektem projektu byłoby opracowanie systemu generowania tekstów wraz z wizualizacją zachowania sieci neuronowej.

↑ coś jak to, ale dla tekstu

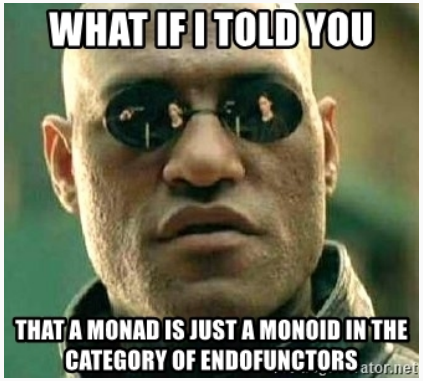
Wykrywanie anomalii / *baseline* dla dowolnego zbioru danych

Dwa pokrewne projekty mające na celu rozwinięcie programu GEval. W obu przypadkach zakłada się implementację sieci GRU lub podobnej działającej na procesorze. Siec ma działać dla szerokiej klasy zbiorów danych na poziomie bajtów, aby być odpornym na wszelkiego rodzaju błędy i zaszumienia w danych.

Uniwersalny detektor prostych anomalii (np. trywialnych błędów w danych).

Uniwersalny *baseline* dla dowolnego wyzwania uczenia maszynowego o charakterze tekstowym.

Wymagana znajomość języka Haskell lub chęć do nauki tego języka.



„Google Image” dla starych gazet i czasopism

W bibliotekach cyfrowych znajdują się olbrzymie zasoby gazet i czasopism. Bardzo przydatny byłby system pozwalający wyszukiwać w nich zdjęcia, rysunki, diagramy itd. W wersji minimum system mógłby być oparty na podpisach i okalających tekstach, lepiej: na rozpoznawaniu obiektów na zdjęciach i rysunkach (taki system można też wyuczyć na podpisach).

Na przykład dla zapytania *pedespedy* chcielibyśmy otrzymać:

siejszemi cyklowemi maszynami (Rysunki №№ 7, 8, 9 i 10) by przeświadczyć się o nieustającej na tem polu pracy i pomysłowości tegoczesnej mechaniki.

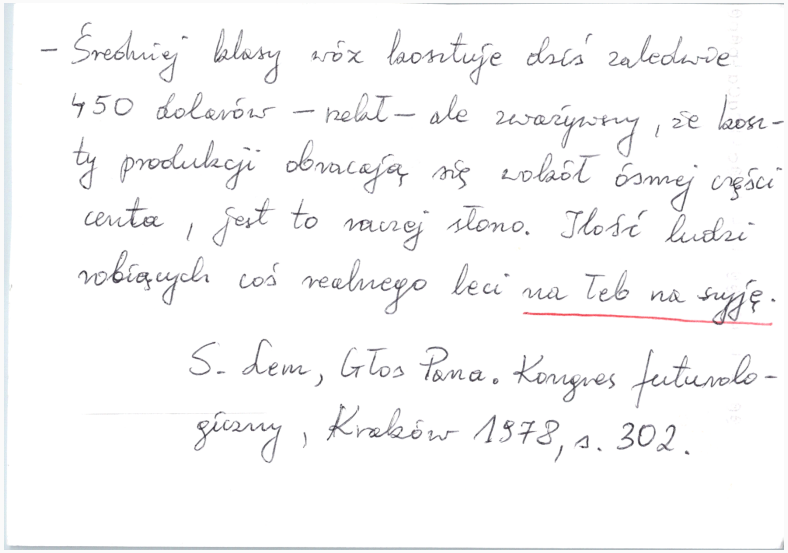
Manja udoskonalania welocypedu, nadania mu lekkości minimalnej, szybkości możliwie największej i t. p. zalet, doprowadziła nawet do pomysłów tak śmiałych i ekscentrycznych, jak „motocykl“ wynalazku włocha Scuri oraz „pedespedy“ (Rysunek 10). Do czego zaś w przyszłości doprowadzić jeszcze może, przesądzać dziś trudno.

Welocypedy systemu Michaux'a pojawiły się

Rys. 10. Pedespedy.

Rozpoznawanie pisma odręcznego w języku polskim

Celem projektu byłoby zamiana pisma odręcznego na tekst za pomocą modeli graficznych Transformer (typu DALL-E).
Osobom realizującym projekt zostanie udostępniony zbiór zeskanowanych fiszek z pismem odręcznych wraz z tekstem wzorcowym.



Wstaw swój projekt

Mile widziane:

- chęć do nauki
- *grit*
- wiersz poleceń i Emacs/vi :), klikanie w przeglądarce :(
- jakiś konik, który można połączyć z uczeniem maszynowym

Każda dziedzina życia w najbliższym czasie zostanie przeryta przez uczenie maszynowe. Chcesz wziąć w tym udział?