



Allambiztonsági Szolgálatok Történeti Levéltára



Reviving the Unreadable

AI-Powered OCR for Degraded Documents in Archives

Theory Meets Practice: Harnessing AI for practical implementations in digital archiving

Gábor Kovács Domonkos Czifra Péter Kőrösi-Szabó

7th November 2024



A HUN-REN Alfréd Rényi Institute of Mathematics Artificial Intelligence Group

REVIVING THE UNREADAB



MAGYAR NEMZETI LEVÉLTÁR

AI-POWERED OCR FOR DEGRADED DOCUMENTS IN ARCHIVES

THEORY MEETS PRACTICE: HARNESSING AI FOR PRACTICAL IMPLEMENTATIONS IN DIGITAL ARCHIVING

GÁBOR KOVÁCS DOMONKOS CZIFRA PÉTER KŐRÖSI-SZABÓ

7TH NOVEMBER 2024





Rényi AI: Introduction

Alfréd Rényi Institute of Mathematics:

- Artificial Intelligence group

Theoretical Research Areas:

- Computer Vision
- Learning Theory

Applied (Industrial) Projects:

- Time Series Prediction:
 - Energy sector
 - Medical sector
- Natural Language Processing (NLP) and Document AI
 - Hungarian Archives:
 - Historical Archives of the Hungarian State Security
 - National Archives of Hungary





Applied Projects with Archives

Main goal:

- Integrate AI solutions into archival processes

Al solutions:

- Latest advances in AI technology:
 - Large Language Models (LLMs)

Applications:

- Retrieval Augmented Generation (RAG)
- Sensitive Data Removal
 - GDPR, inquiry, research
- Named Entity Recognition
 - Database, Knowledge Graph





Tasks & Challenges





- Data format:
 - We have: scanned document images
 - We need: text-based data
- Data quality:
 - highly degraded documents





- Data format:
 - We have: scanned document images
 - We need: text-based data
- Data quality:
 - highly degraded documents



- Data format:
 - We have: scanned document images
 - We need: text-based data
- Data quality:
 - highly degraded documents







- Data format:
 - We have: scanned document images
 - We need: text-based data
- Data quality:
 - highly degraded documents









- Data format:
 - We have: scanned document images
 - We need: text-based data
- Data quality:
 - highly degraded documents



Archive specific challenges:

- Data format:
 - We have: scanned document images
 - We need: text-based data
- Data quality:
 - highly degraded documents



Utodállamok. Magyarország.

•A Balkán az európai Aiplomácia közöppontjúban* elmmel a Honitor /13/ belgráfi tudósítója felsorolja az utóbbi heteknek politikai eseményeit, melyek Furópa figyelmét a Dólkelet felő terelik "zek között első helyen Grandi olasz külügyministernek albániai utját és egy olasz tábornoknak az albán hadsereg élőre állítását. A Monitor ebben a fascista szupremacia európai megerősödését váli látni. Minthogy a Balkánnak ez a fiplomáciai tevőkenysége nagy órdekellentőteket rejt magában,-irja a lap, es aligha fogja az általános balkáni bőke üg ét szolgálni, ha apróbb surlódási okokat, ta-

	Kádár Róbertné: Tisztelt Búdapesti Tanács! Ok-
~	töber 22én Budapest iskossága eleget tett hazafias kötelességé-
	nek és szavazatával eldöntötte, hogy a hatalóm gyskorlása a dolgo-
	zó nép kezébe kerüljön,
~	A választás eredménye bebizonyitotta, hogy Budep
	lekossága felzárkózott a Bápfront zászlaja mögött és egyuttal
	retet tett og ötéves terv negvilésitésérs, a béke megvédésére.
	A Budapesti Választási Bizottság megállapította, hogy
	a választás mindenütt a legnagyobb rendben, a törvényes előirá-
~	soknak megfelelően zajlott le.

A DIB magyar lapszelét kizil a Göbbels beszídéhez fizütt komentárokból és ismerteti a Fercer Lloy, Ut bit y rada, hagyarag és ősestbartis ostkatt. A insz Guerdanu Caitum, a háborban álló megyarországról irva het sulyozan, norv a magyarte nyugalomul és bizalomal várják eselett aravanal fojlményeit. Magyerországon nincs ese izgelőmi son hintária, vestdeutochar Beobachter a magyar nagybirtek kérdésérő irva hingaulyöss, hogy sgösz Burópában nar ösek ingvarorszison vinnet nagybirtokolt, s norv a háboru után a nagybirtokoknek feltítlenül el kell túnnök.

A Stoold olns Tidningen budapest, tudó itó jának jolentését közli a Budapest ellon intizett logatóval lésitánudáról.

<u>A szövetséges földkesi tandari ada "Baseurorendaré"</u> erkező hirek szerint" ugy értesül, bogy <u>follar Kazur</u> volt ainiszternek unokašopsét internálták, tert több zsidőt szöktetett át Jslovákiába.

Good

Moderate





Demo

How SOTA methods struggle to extract text





Demo: SOTA off-the-shelf OCR systems

Original Document



117 ORSZÁGNEI EVELTÁN



Demo: SOTA off-the-shelf OCR systems

Original Document

Zimmer Ferene fösserkasstö A kindskri falds-
C. Made Anto W MAGVAR TAVIRATI IRODA (U 1990)
BIZALMAS KULPULITIKAI SZEMLE
.51. szán,
1. oldel. 1944. juliu: 29.
Magjarorsság.
A 2008 internet internet internet in a gonder in benefather Histiff konnetsparked i Startrett i Steady i U.S. 1998 internet Ingranuds de Cherchartis etkekt, A linst <u>Gungtongu Intern</u> et historian <u>Gin Lagranowskippi intern</u> har har <u>hortan</u> startret
(ydgalostal % diakaansa worku e wittet hervers indinesses hervertersetten ander ander ander ander ander ander ander ander generative state ander ander ander ander ander ander ander ander generative state ander ander ander ander ander ander ander ander generative state ander ander ander ander ander ander ander ander a generative state ander ander ander ander ander ander ander ander ander state ander and
nikk. <u>A Sicold alta Tidnine m</u> budap svi tidi itojársk ja- lentésik kozdi a Rydapiet alton iti izete jajutóvel léjtiniké- róli
drkand hisk attribute interactions into a state of a st
dátat, horr beta tilter i ter andre andre andre i ter Borrier i 1999 - Ministri i terret, ittel i a shirar tilt der andre i terret andre i terret - Ministri i terret, ittel i a shirar tilt der andre i terret - Ministri i terret, i terret i ber i terret i terret andre i terret - Ministri i terret, i terret i ber i terret i terret i terret i terret - Ministri i terret i terret - Ministri i terret i te
ast filities. Some of the state
nunkassé stezervesésépől szólva hanyal, osza, hogy is ipuri nunkások bojkottálják az Imrény iltal alkotott punkaszervestet,
Kaloti_erovonel.
Litatia of the Margarian Statistics of the Stati
jointiet, savenisalau, servetis itojihito terijariya, Bolutzia bijederin, vilaine Baya ingelo vanci sileinini. Tovat-ti- costilvod, ilid nicot macon erite a set a in folki partis costilvod, ilid nicot macon erite a set a in folki partis folk scottottek, ito delenje nove huro is folyeni.



1. oldal. 1944. julius 29. Magyarország.

A DNB je l entés k iadja a Göbbels beszédéhez f üzött kommentérokból és ismerteti F e rrer L oy. nagyarország és összetartó orkkét. A linzi Oberdonau Zeitung hábor ban álló Magyarországról írva hangs úlyozza, hogy a magyar mér nyugalonnal és bizalonnal várják e szelle ni áranl a t fejl ődésé t. Magyarországon nincs e zen izgalon, sen his teria. Westdeutscher Beobachter a magyar nagybirtok kérdéséről jírva hangs úly ozza, hogy egész Zurópában csak vagy oni ké rdés. A Sto ol colns Tidning un budapesti tudós ítójának je lentését közli a Budapest cl ion intézet e l oputóel léstin adác A szövetséges földközi tengeri érkező h írek szerint * ugy Cites ul, hogy niszternek unokaje csét inter málták, cart több zsidó s öktet avv át Jel uvákiába. A londoni rádió magyar adás vila r felszólitotta a razdákat, hogy szabotálják, a nogy csak lehet Jurasc k-féle hast gáltatási terv ct, rivel a a boru al ri Ma eyw or az t ozulu t, Magyarország na vár ost Sle tines ni s orsz ugoktól s a par is stságnak szüksége lesz a magyar termé er . Ako a th-rádió Satani aalau el tével kap oson tban azt állitja, hogy Sztanis slau s Kolo lea között hetvenezer ro-Ey er ho rvédet szórt ek szét. A ditó r ondkivül 🗾 hangon 🚽 nad is S stójay miniszterelnököt és a na ryar katonai ver stoat. A noszkvai rádió magyar ad fadban a yar ip eri nunkássá átszervezésé től szólva h engsul os , hogy peri nunkások bojko stálják az Imré ay Altal alkotott runken zervezetet, Kalati sr ovonal. Eshelin pénteken n epiparancsban k osolt o rar Jarószlav, valamint Freszt-litevo z elfog lalts it. i szov Tw inszktől északra és délre, She rli 61/Sch eul on/ daire, ra és nyugatra, valamint Sztaniszlautól ny atra és d el ayn atra további szovjet előrenyom el ést és nagysz enu helység elfoglalását jelentett Sztaniszla n környékén foglalt ar Pere lic tov P olotvin bely ećzeket, v lamint Bere sinszko vasut s állomást. Tees st-Litovs któl nyugatra a szovjet osapat k örül adeta n egy háron had osztályból álló német harc es pertot s azt a su kel oti partja felé szor ították. Itt jelenleg neve n harcok folyan k.



HUN-REN Alfréd Rényi Institute of Mathematics Artificial Intelligence Group

	Correct	
	Small error	
	Large error	

Google Lens

1. oldal. 1944. ju	lius 29.
Magyarorszá 🔨	
A DIB Jagyar 1	laps <mark>wo</mark> lét k <mark>iz ik</mark> a Göbbels beszédéhez
füz <mark>ütt konnen</mark> t	tárokból és ismerteti a Fe tver L oy , seesság,
Lagyarság és Ö	sszetart is owkkét. A linzi Qo rdonau Seitung
hábor ban áll	yarországról irva he sulyos an, hogy a
nyugalommal A	s bizalo nnal várják e keleti ar ovon el f ojl nényeit.
Magyarországon	n nincs se n izgalom, s ou his t eria. v estdeutsch or
Beobachter a ma	agyar nagybirtok kérdéséről irva hangsulyon, holy
egész Európába:	n m ar csak v arote for varet n egybirtoko 1.
8 no vy a hábo	r a után a nagybirtokoke k felt etlenül 31 k al i t vn
niök.	
A S icod ol ne	Vidningen budapesti tudós ítójának j o-
lentését közli s	Bud epest cl ion intizett Jogutóbbi le it inndác
ról.	
A szövetsérges	földközi tan cer "eni ré
érkező hirek s	anint" (rtaciil
niester net und	baie seit inter miltik mert tähb seidi enäk inhett
6+ Tel myériér	
A londoni rédió	neag ar ad ac ile r felszálitatta a
déret home real	abatálják a nogy czak labat Jumaro k- tála baril ba
adltatáci tanva	t wivel a M how a 1 wi May
No ov whome no	a ván pat lelinene ni sonne vzektál s a p vn vs t
cá nak czükcáca	lace a magyan ta ná an
A Vora than 4	tió Catanaa lau al tával kan nenn than
art állitin bon	u Satania ulau ás Kala las köz att hatvanagan na-
ast allitiga, nog	s fat al antit a dijé n andhirul é a hangan an
gy er no rvedet	as ert ex seet a uno ronurivui es hangon
uauja Sacojay	allisterein clot es a na ryar katonal ves stoot.
A HOSZEVEI PEG	io nagyar au fautan a L yar iperi
nunkassa atsz	ervezese toi szoiva nangsul os , nogy ipari
nunkasok bojko	staljak az imre ay ital alkotott punkeng ervezetet,
K al oti pr ovo	nel.
Est chin pent	eken n epiparan esban ko ite Pran 💊
Jar oszlav, vala	mint Preszt- Litevo z elfog lalts it.
A fontes 1 s	zovjt a nijelentes w inszktol eszej
ra es deire, She	rlitol /Schmio n/d I ce, Lublintol eszaknyugat-
ra es nyugatra,	valamint Sztaniszlautol ny atra es d el ay atr e
jelentett, Sat	aniszlan környékén lfoglaltár Peredisztov a Bolotvin
belységeket, va	lamint Bere in ozko vasuti állomást. Tees at-Li-
tovs któl nyug	atra a szovjet <mark>o</mark> sapatok körül ale tak egy három had
osztályból álló	német harccsoportot s azt a iu kel oti partja
felé szor itotté	k. Itt jelenleg neves harcok folynak.



Demo: SOTA off-the-shelf OCR systems

Original Document



117 ORSZÁGOSI EVELTÁS

Gemini

1. oldal. 1944. julius 29. Magyarország. A DNB je l entés k iadja a Göbbels beszédéhez f üzött kommentárokból és ismerteti F e rrer Loy. nagyarország és összetartó orkkét. A linzi Oberdonau Zeitung hábor ban álló Magyarországról írva hangs úlyozza, hogy a magyar né nyugalommal és bizalommal várják e szelle mi áraml a t fejl ődésé t. Magyarországon nincs e zen izgalon, sem his teria. Westdeutscher Beobachter a magyar nagybirtok kérdéséről jírva hangs úly ozza, hogy egész Zurópában csak vagy oni ké rdés. A Sto ol colns Tidning un budapesti tudós itójának je lentését közli a Budapest cl ion intézet e l oputóel léstin adác A szövetséges földközi tengeri érkező h írek szerin<mark>t *</mark>ugy <mark>Ci</mark>tes ul, hogy niszternek unokaje csét inter málták, cart több zsidó s öktet avv át Jel uvákiába. A londoni rádió magyar adás vila r felszólitotta a razdákat, hogy szabotálják, a nogy csak lehet Jurasc k-féle hast gáltatási terv ct, rivel a a boru al 'ri Ma eyw or az t ozulu t Magyarország na vár ost Sle tines ni s orsz ugoktól s a par is stságnak szüksége lesz a magyar termé er . Ako a th-rádió S atani aalau el tével kap oson tban azt állitja, hogy Sztanis slau s Kolo lea között hetvenezer 📪 🗗 Zy er ho rvédet szórt ek szét. A ditó r ondkivül kangon nad ia S atójay miniszterelnököt és a na ryar katonai ver stoat. A noszkvai rádió magyar ad fadban a yar ip eri nunkássá átszervezésé től szólva h engsul os , hogy p eri nunkások bojko stálják az Imré ay Altal alkotott runken zervezetet, Kalati sr ovonal. Eshelin pénteken n epiparancsban k osolt o rar Jarószlav, valamint Freszt-litevo z elfog lalts it. i szov Tw inszktől északra és délre, She rli 61/Sch eul on/ daire, ra és nyugatra, valamint Sztaniszlautól ny atra és d el ayn atra további szovjet előrenyon el ést és nagysz enu helység elfoglalását jelentett Sztaniszla n környékén foglalt ar Pere lic tov P olotvin **CER:** 25.1% bely eczeket, v lamint Bere sinszko vasut s állomást. Tees st-Litovs któl nyugatra a szovjet osapat k örül adeta n egy háron had WER: 56.9% osztályból álló német harc es pertot s azt a su kel oti partja felé szor ították. Itt jelenleg neve n harcok folyan k.



HUN-REN Alfréd Rényi Institute of Mathematics Artificial Intelligence Group

Correct
 Small error
 Large error
 CER: Character Error Rate
 WER: Word Error Rate

Google Lens

1. oldal. 1944. julius 29.	
Magyarorszá .	
A DIB Jagyar laps wolét k iz ik a Göbbels beszédéhez	
füz ütt konnentárokból és isnerteti a Fe tver L oy,	ság,
Lagyarság és Összetart is oskkét. A linzi Qo rdonau Se	itung
hábor ban áll yarországról irva he sulyos an, hogy a	
nyugalommal As bizalo nnal várják e keleti ar ovon el f	jl nényeit.
Magyarországon nincs se n izgalom, s ou his t eria. v esto	leutsch or
Beobachter a magyar nagybirtok kérdéséről irva hangsulyo	, ho y
egész Európában m ar csak v arote for varet n egy	birtoko 1,
8 no vy a hábor a után a nagybirtokoke k felt etlenül 3	k al i t wn
niök.	
A S icod ol ne Tidningen budapesti tudós ítójának j o-	
lentését közli a Bud epest cl ion int izett Jogutóbbi le	it inndác
ról.	
A szövetsérges földközi t an cer "eni ré	
érkező hirek s terint" C rtesül,	
niszter nek unokaje csét inter málták, mert több zsidó 🛛 🕫	pök inbatt
át Jel uvákiába.	
A londoni rádió usag ar ad as ila r felszólitotta a	-
dákat, ho ry szabotálják, a nogy csak lehet Jurasc k- téle	hasil be
gáltatási tervet, rivel a M bor a 1'ri Mau	t,
Na cy urors nág a vár nat lelinero mi s orsz ugoktó	l e a p ur us t-
sá nak szüksége lesz a magyar te mé er .	
A Kora th-r étió S atanaa lau al tével kap oson tban	
azt állitja, hogy Sztanis ulau és Kolo lea köz ott hetveneze	r re-
gy er ho rvédet az ért ek szét. A diló rondkiv ul é s han	gon
uadja S atójay minisztereln clöt és a na ryar katonai ves	stoot.
A noszky ei rádió nagyar ad fadban a L yar ip eri	
nunkássá átszervezésé től szólva hangsul os , hogy jips	ri
nunkások bojko stálják az Inré ay Ital alkotott punkeng	ervezetet,
K al oti provonal.	
Est chin pénteken n epiparan esban kö lte Pran 🛰	
Jar oszlav, valamint Preszt-litevo z elfog lalts it.	
A fonte s i szovj t a nijelentés w inszktől észej	
ra és délre, She rlitól /Schmio n/d l ce, Lublintól észab	nyugat-
ra és nyugatra, valamint Sztaniszlautól ny atra és d el ay	atre
and the second se	
jelentett, S ataniszla n környékén lfoglaltá r Perediszto	v a Bolotvin
belységeket, valamint Bere g in o zko vasuti állomást. Te es	at-Li-
tovs któl nyugatra a szovjet <mark>o</mark> sapatok körül sale tak egy h	áron had
osztályból álló német harccsoportot s azt a iu keloti par	tja
felé szor ították. Itt jelenleg neves harcok folynak.	

CER: 22.9% WER: 60.5%





Error Analysis

Reasons behind the poor performance:

- Text Detection:

- challenging to detect text lines
- some part of the text is already lost
- Text Recognition:
 - challenging to read out text



Error Analysis

Reasons behind the poor performance:

Text Detection: —

- challenging to detect text lines _
- some part of the text is already lost _
- Text Recognition: _
 - challenging to read out text -



nunkassá átszerveléséről szólve hungell esse, hogy te ipurt nunkások bojkottálják az Imrédy által alkotott nunknazerveletet

manesba.8

n nonteken noni

elentitt Szteniszlen kögyékén "lfoglaltát" Lységeketly lanint Bere 85ozko vasuti ál tovsattól nyugatra a szovjeg Gesepatak körüleg esztályból álló nénet harcesserrtot s azt a

lé szoritották. Itt jelenleg neves hercok fely

52

Kaloti grevonal.

92

100

false positive overlapping boxes

75

97

- 1 102

103 ORSZÁGOR = 103

inconsistent size





Error Analysis

Reasons behind the poor performance:

- Text Detection:
 - challenging to detect text lines
 - some part of the text is already lost
- Text Recognition:
 - challenging to read out text

A szövetséges földközi tenderi ado "ingravorardaro
--

- (GT) A szövetséges földközi tengeri adó "Magyarországról
- (OCR) A szövetsé<mark>r</mark>ges földközi t an cer "emi ré
- gyar horvédet szírtek szét. A mádió rendkivül éles hangon tá-
- (GT) gyar honvédet szórtak szét. A rádió rendkivül éles hangon tá-
- (OCR) gy er ho rvédet az ért ek szét. A diió r ondkiv ul é s hangon –





Our Solution





Approaches

Enhancing OCR systems directly:

- Data perspective:
 - archival data domain
- Architecture perspective:
 - larger models
 - change model type
 - e.g. Transformers

Preprocessing for OCR:

- Document Image Enhancement (DIE)
 - improving the quality of the input image



Approaches

Enhancing OCR systems directly:

- Data perspective:
 - archival data domain
- Architecture perspective:
 - larger models
 - change model type
 - e.g. Transformers

Preprocessing for OCR:

- Document Image Enhancement (DIE)
 - improving the quality of the input image









Approaches

Enhancing OCR systems directly:

- Data perspective:
 - archival data domain
- Architecture perspective:
 - larger models
 - change model type
 - e.g. Transformers

Preprocessing for OCR:

- Document Image Enhancement (DIE)
 - improving the quality of the input image



Zimmer Ference fösserkesstö
Hézer Zoltán CE MAGYAR TÁVIRATI IRODA DE Telefon: 145-510
6300 Minden jog jenntartázával Kéziratnak tekintendő Házi vok zorovitáv 0060
BIZALMAS KÜLPOLITIKAL SZEMLE
51. száu.
1. oldel.
Magyarorsság.
A DIB pagyar lapszellét ktall a Göbbele bezvédéhez
hagyarság és Castartis e kkét. A linzi Querdonau Saitun.
nadordan airt ingerorszardi irva nel sulydzza, hofy a hage ret nyugalonnul és bizalonnal várják a heldti arovonal fajlanényeit.
Magyarországon mines sem izgalom, sel hilstéria. Westdeutscher Beobachter a magyar nagybirtek kérdéséről irva hangsulyóla , holy
 s nogy a háborn után a nagybirtokoknek feltítlenül el kell tún- /
niök. A Stooldolna Tidningen budapesud tudésítétának ja-
lentését közli a Budepest ellen intízett legutórel légitimlár- ról.
A SZOVUTSÁRES TÖLÜKAZI tenderi adu "da nurorendard" érkező hirek azerint" ugy ditesül, hogy dállas Kazar volt di-
niszternek unckejensét internálták, mert több zaudól emőktetett
A Londoni rádjó negyar adásában felozólitotta a rez-
<u>Altatici terret, pivel ha e hibera altri Magaronatici terretti.</u>
sá nak szülszige lesz a negyer ternészie.
azt Allitja, hogy Sztenischer (s Kolo.ca között hetvenezer re-
uadia Estőjay miniszterelnölöt és a paryar katonai vezetőst.
hunkásad: átazorvetőséről szölve hungsul cone, hogy on iperi
nunkasok bojkottáljék ez imrédy fital alkotott runkrozervezetet,
Kalpti erovonel.
Jarószlav, valavint Bregzt-Fitovszk elfoglalisit.
R és délre. Shevlitól /Schemlen/ delle. Lablintól északevtet-
ra és nyugatra, valamint Sztaniszlautól nyu atra és dólnyugatra - további szovjet előrenvenelést és ha szzinu helysés alforlalását
jelentit Szteniszleu környékén ilőglalták Perejigztov, a Bolotyin belységeket, vilujnt Benejingko visti állondat, logart-ld-
tovsaktól nyugatra a ezovjet csapatak körülzéctas egy három had- osztályból álló német henen somtat s egt a fela kelati nertin
felé szoritották. Itt jelchleg neves harcok folyanh.
112- 0000/0000000000
1 7 OPES/Aldres 1 How 2 MM





Solution

Document Image Enhancement (DIE):

- Cleaning document images:
 - removing unnecessary noise while retaining text
- Model architecture:
 - convolutional based U-Net
 - U-Net: image transformation, removing noise
 - Convolution operation: filtering image
 - Skip connection: better training properties





Solution

Document Image Enhancement (DIE):

- Cleaning document images:
 - removing unnecessary noise while retaining text
- Model architecture:
 - convolutional based U-Net
 - U-Net: image transformation, removing noise
 - Convolution operation: filtering image
 - Skip connection: better training properties







Solution

Document Image Enhancement (DIE):

- Cleaning document images:
 - removing unnecessary noise while retaining text
- Model architecture:
 - convolutional based U-Net
 - U-Net: image transformation, removing noise
 - Convolution operation: filtering image
 - Skip connection: better training properties









Demo





Demo: Document Image Enhancement (DIE)











Machine Learning perspective:

- Machine Learning (ML) paradigm:
 - solving the task by learning from examples / data
- ML model needs data:
 - <u>Text Detection</u>: bounding boxes around the text lines
 - <u>Text Recognition</u>: text in the document
 - <u>Document Image Enhancement</u>: clean version (black text on white background)

Supervised Learning:

- Human labeling:
 - Text Detection: cumbersome
 - Text Recognition: cumbersome
 - Document Image Enhancement: extreme hard
- Synthetic generation:
 - Text Detection: easy
 - Text Recognition: easy
 - Document Image Enhancement: moderate





Machine Learning perspective:

- Machine Learning (ML) paradigm:
 - solving the task by learning from examples / data
- ML model needs data:
 - <u>Text Detection</u>: bounding boxes around the text lines
 - <u>Text Recognition</u>: text in the document
 - <u>Document Image Enhancement</u>: clean version (black text on white background)

Supervised Learning:

- Human labeling:
 - Text Detection: cumbersome
 - Text Recognition: cumbersome
 - Document Image Enhancement: extreme hard
- Synthetic generation:
 - Text Detection: easy
 - Text Recognition: easy
 - Document Image Enhancement: moderate





Machine Learning perspective:

- Machine Learning (ML) paradigm:
 - solving the task by learning from examples / data
- ML model needs data:
 - <u>Text Detection</u>: bounding boxes around the text lines
 - <u>Text Recognition</u>: text in the document
 - <u>Document Image Enhancement</u>: clean version (black text on white background)

Supervised Learning:

- Human labeling:
 - Text Detection: cumbersome
 - Text Recognition: cumbersome
 - Document Image Enhancement: extreme hard
- Synthetic generation:
 - Text Detection: easy
 - Text Recognition: easy
 - Document Image Enhancement: moderate





Synthetic Data Generation framework









Text:

- language
- layout
- fonts

Background:

- paper type
- overbleed

Perturbations:

- 20 noise types
- aged document

Geometric:

- rotation
- warp





Demo: Synthetic Data Generation

-				and the second
Sec. 3			· h	Lesanth Linga acet inga winet-main
			······ No.	sont as level and fuller to be a serie of the series of th
				- Treas
1	6.1%		27	23 St control and a short y polatical all a canadala
Sec. 1	124184	and the		And the second se
		-		
and and	1			et dele wege egisconette mittale watenershite?nx
		1		The second se
		1	1000	
1.			1	s change have and set a set a set a set a set
	-	Contraction of	the state	THE REAL PROPERTY OF THE PROPE
	1226		1 10	di regi tattel, o della a svervezičk bitatikusta sta
	100	1	1	a los de lamp destricted a de la contra
	100	1	1 2	in the station of the second to set the
4		1	1 00	discription haloning a stranger to a taken areas
		11		r to be possible use to be of the cape of the one able \
1.16		1	1	as the service of the service and the service of th
12.24%	11.1			
1.5	in f	1	10	o na user reality special error and participation in a party
100	1.00		1 22	to to had constal outputs policitik as Oktobranda,
	-			atily in a sol level of a grave silen of in a galatest
	1 SI		1.1. 0.5.1	and antibage V
City of			1. 1. 1.	et a money service and any the service and any service and the
	1.	-	2	ren e Hazen i Albe began be intizatet provinsia i a. ek. Sand
1	-			and the second s
		-	5	Your Carland, C. Santal.
			10	ding out water source for more to sector eres from the function more no
		1.64	T	apply a perinkue cleane whomewhen a headership of a
1.5	199		624	Like and the second sec
and and	0.00	10.12	1991	a start and the start of the st
S. P.A.			- 15	au ditenti "aga sukufnat, in a Grupanti szervez
1. 1. 1	1	1 6.8	2.0	with cell 4, here negaritizeds at Orac to dending t
	14			
	14	1	1	3 ALL OF MORE PERMITED A CALLER & TO BE & A B .
	They have		10	of address.
1	10			autorecides a reasonable of feasible of Operation and
-			3	Soons, a Bakrann e v 2 gazninberral min harantis berbijarini
	14.4		1 4	contraction of the second se
		1	in de	bi river stan Son Caller cantan
Plan's	1.1.1			a he game may but
C. Carl			1 20	train pres 1.510,000 to 11 beaund Cyte, bicatent
	and.			and and mather the best fare legit it it on in
	AND TO			
	100		1 5	and it is the policke back and it is
	25.2		1	gy and apple
	-	1	1 -1	A Cardemanar
-	Contractore and a second come with and			
140	Byar Or	BOOKER	1 1	more maraliants self-laved a sugerment usual is to sederately a n e ra
-	Level	1.00		









How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts

Background images:

- paper types

Noise types:

- stamps, scribbles, textures, ...





How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts

Background images:

- paper types

Noise types:

- stamps, scribbles, textures, ...







How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts

Background images:

- paper types

Noise types:

- stamps, scribbles, textures, ...



Abadi MT Condensed Light Albertus Extra Bold Albertus Medium Antique Olive Arial Arial Black Arial MT




How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts

Background images:

- paper types

Noise types:

- stamps, scribbles, textures, ...



Abadi MT Condensed Light Albertus Extra Bold Albertus Medium Antique Olive Arial Arial Black Arial MT







How to customize the framework?

Corpus:

- different kinds of languages

Text types:

- different kinds of fonts

Background images:

- paper types

Noise types:

- stamps, scribbles, textures, ...



Abadi MT Condensed Light Albertus Extra Bold Albertus Medium Antique Olive Arial Arial Black Arial MT







Our OCR

Our OCR model

Training our custom OCR model:

- TrOCR model
 - Transformer-based model
- Trained on:
 - Hungarian language
 - Degraded archival documents
 - Synthetic Data Generator
 - Human labeled data







OCR Pipeline





































































Evaluation



Performance:

- Character Error Rate (CER)

OCR \ Input	Raw document	DIE-cleaned document
Gemini	25.07% [CER]	6.87% [CER]
Google Lens	22.86% [CER]	5.03% [CER]
Our pipeline	10.59% [CER] *	6.46% [CER]

- Nvidia A-100 server:
 - 8 GPUs (40GB)
 - 2 GPUs:
 - 25 000 document images per day







Performance:

- Character Error Rate (CER)

OCR \ Input	Raw document	DIE-cleaned document
Gemini	25.07% [CER]	6.87% [CER]
Google Lens	22.86% [CER]	5.03% [CER]
Our pipeline	10.59% [CER] *	6.46% [CER]

- Nvidia A-100 server:
 - 8 GPUs (40GB)
 - 2 GPUs:
 - 25 000 document images per day







Performance:

- Character Error Rate (CER)

OCR \ Input	Raw document	DIE-cleaned document
Gemini	25.07% [CER]	6.87% [CER]
Google Lens	22.86% [CER]	5.03% [CER]
Our pipeline	10.59% [CER] *	6.46% [CER]

Local model - data privacy

- Nvidia A-100 server:
 - 8 GPUs (40GB)
 - 2 GPUs:
 - 25 000 document images per day







Performance:

- Character Error Rate (CER)

OCR \ Input	Raw document	DIE-cleaned document
Gemini	25.07% [CER]	6.87% [CER]
Google Lens	22.86% [CER]	5.03% [CER]
Our pipeline	10.59% [CER] *	6.46% [CER]

Local model - data privacy

- Nvidia A-100 server:
 - 8 GPUs (40GB)
 - 2 GPUs:
 - 25 000 document images per day











.

Applications

LLMs in Archival Work:

- Digital Historian:

- Retrieval Augmented Generation (RAG)
- Semantic search in archival database
- Automatic summarization of documents
- Dynamic cross-referencing of sources

- Sensitive Data Removal:

- Sensitive and personal information
- Abstract concepts described in natural language
- Flexible customization for sensitive data needs

- Named Entity Recognition (NER):

- Recognition of names, locations, events, etc.
- Building structured representation (e.g., graphs, tables)
- Distinguishing between similar or overlapping entities



LLMs in Archival Work:

- Digital Historian:

- Retrieval Augmented Generation (RAG)
- Semantic search in archival database
- Automatic summarization of documents
- Dynamic cross-referencing of sources

- Sensitive Data Removal:

- Sensitive and personal information
- Abstract concepts described in natural language
- Flexible customization for sensitive data needs

- Named Entity Recognition (NER):

- Recognition of names, locations, events, etc.
- Building structured representation (e.g., graphs, tables)
- Distinguishing between similar or overlapping entities



HUN-REN Alfréd Rényi Institute of Mathematics Artificial Intelligence Group





LLMs in Archival Work:

- Digital Historian:

- Retrieval Augmented Generation (RAG)
- Semantic search in archival database
- Automatic summarization of documents
- Dynamic cross-referencing of sources

- Sensitive Data Removal:

- Sensitive and personal information
- Abstract concepts described in natural language
- Flexible customization for sensitive data needs

- Named Entity Recognition (NER):

- Recognition of names, locations, events, etc.
- Building structured representation (e.g., graphs, tables)
- Distinguishing between similar or overlapping entities



BIZALMAS

1450 XIT 10.

- 1950.nevenber 3.-in.-

/ az Ulfe az Ujyárosh fzi talfaster fian 16 ára 10 perokor kezőődött./

töber Sk.- én Hudapust iskorséga eleget tett hazarias kötelességének és ezevaratóval eldöntötte, hogy a hatalón gyekori és a solga-

A válaustás orodmánye bobizonyitotta, hogy madge lekasegga felzírközött a dépfront sészinja migött és egynétni rotat tett iz fölvez form regyriódítésére, a böks segrédésére.

A Sudaposti Választási Bizottság megállapította,hogi a választás mindenütt a legnagyobb rendben, a törvényes előirá-

A Budapesti Választési Bizottság nevében kibirdetem

A Budepesti Válazztári Bizottzág munkáját befejezte.

a v'loustési creduényeket: Budspesten a vélaastéara jogosultak széma: 1,246.764. Sbbél szavazott 1,209.457. A hépfrontra adték

A Eudepesti Véresi Tamésanak és minden tegjének mek sikert kivének a Eudapesti Vélamgtési Tirottaég nev Nem na (1 sikk filé

a leadott szavazatok közül érvénytelen volt 15.641.

nagy feladatok elvégzázéhez. / T a p z./

1: szavazatukat 1,186.470-en, s Képfront ellen szevaztak 7.346-en,

K á d á r Rébertné: Fisztelt Súdspesti Tenécs! Ok-

0306/.

zó náp kezébe kerüljön.

soknak megfelelden zajlott 1e.

BUDAPCET FOVAROS LEVELTARA



LLMs in Archival Work:

- Digital Historian:

- Retrieval Augmented Generation (RAG)
- Semantic search in archival database
- Automatic summarization of documents
- Dynamic cross-referencing of sources

- Sensitive Data Removal:

- Sensitive and personal information
- Abstract concepts described in natural language
- Flexible customization for sensitive data needs

- Named Entity Recognition (NER):

- Recognition of names, locations, events, etc.
- Building structured representation (e.g., graphs, tables)
- Distinguishing between similar or overlapping entities



Raw document image

Washington, november 14. (Reuter.) Az

elhunyt conti polgarmester, Micc Swiney ozvenyet és nyolc asszonyt itt elfogtak, akik egy tüntetés alkalmával az angol nagykövetség előtt, az elhunyt polgármestel jelenleg Anglában fogságban tartott növérének szabadonbosátását követetlék (MT.)

§ Az özságos magyar gyűjtenényegyden taihássa alakuló üléset hovarhatt. 3- an deltatná 5 órakor tartja a Migyar Tudományos Akadémia Uléstermében A kázoktatásúgyi míniszter, aki a förvény értelmében a tanácsnak a munkálatok megindulása tidejében elnőke, ez alkalommal levéltrainki, könyvtáralnik és muzeumaink helyzetére kiterjeszkedő megnyító beszédet fog mondani.

§ Róma, <mark>novembor 15</mark>. (Magyar Tavirati Iroda..) Romanalli azradas Budapestre utazott.

§ A Mietorotógiai Integral jelentése: A nagy légnyemás északnyugat felé huzódik, maximuma Skócia felett van, minimuma pedig Finnország felett. Szicilia táján is aránylag kicsi a légnyomás. Az idő a kontinens nyugati felében borus, ködös, keleten derültebb. A hömérséklet nem változott tényegsesn. Hazánkban az idő tufnyomóan derült és szárza. Az ígiell minimum az Atlód egyes helyein kisebb a – S foknál. Időprognózis: Nyugatra forduló szelekkel változékony és hűvös idő várható. (ml.)

Proga, november 15. Schuster az situli Schuster Schuster (1998) Schuster Schuster (1998) Schuster (1998)

NER: (persons, locations, dates, ...)



HUN-REN Alfréd Rényi Institute of Mathematics

Artificial Intelligence Group

Graph-based data representation





Thank you for your attention!





© European Union 2020

Unless otherwise noted the reuse of this presentation is authorised under the <u>CC BY 4.0</u> license. For any use or reproduction of elements that are not owned by the EU, permission may need to be sought directly from the respective right holders.

Backup Slides



Text Recognition: TrOCR



U-Net





OCR datasets



Salara Constant	1000	
SWC ENTERPRI	O-VI	BHD
No. 5-7, Jalan t	tahagoni	7/1.
Sekysen 4, Bandar	Utama,	44300
Batang Kali,	Selangor	•
TAX IN	MOLCE	
(GST ID No. : O	02017808	184)
08/01/2018	Cashie	er: 123
11:07:06	No:0100	0080332
Iten/Desc. Oty	Price	Ant.
20X30 BEB 1KB91X30)	
20X30 1K6 1	8.00	8.00
Total Qty :	1	
TOTAL AMOUNT		8.00
CASH		10.00
CHANGE		2.00
G 1644 Included In T	fate	0.45
Thank You ! Pleas	e Come Ag	ain !
Goods Sold Only	Exchange	able

0	99 SPEED HA LOT P.T. 2 TAMS 41150 H 1413- BST ID. H	RT 5/8 (519537- 811, JALAN ANDS N BERKELEY LANG, SELANDOR SETIA ALAM 2 10 : 00018174771	10 A, 2
	INVOICE N	0 1 10222/102/10	
03:	29Ph	582936	20-11-17
06 27 28 43	9 WILD 200 09 ZING HEIM 93 DISNEY DS 885 JOHNSONS	G DAT & WHEAT 112 WIDE NECK PHS.5 21N1 BD	RN4.10 s RN4.10 s RN5.69 s RN5.99 s
T	otal Sales (1 Roundi	nclusive OST) R ing Adjustment R Rounding R CASH R CHANGE I	n 169.78 N .02 N 169.80 N 200.00 M 30.20
	DST Summary 5 = AX	Amount(RN) 160.17	Tax(RH) 9.61

12	12	3	-	PETETHE	R 03.4	al un	Rose	the f		A small
20	25H	ours.	#E440*	at write ILGOID	#1.80708 Gef 11.75	5.16	45	*5		then t
1.5	225	1.04	11	57.115	57.054	12	_	_		First
招			_		_	_				that I
11:22							-	_		for m
12										policy
30.1	-			-	-			-		bellet.
S.			-				_			
1			_	_						Parthe
10				1		-		-		
11		-	-	4	1.4		-	-		
12.			-		A	-				I hav
125	1		18. 1	1000	-		-			which
				iir.	-	-	-	-		t hat
16.	1			1					-	stric
133	-			7		_			-	time
1.			100		_	_		_	8_	
14					_	-	_	_	8	
57.5		-		10.151	111-1	-			22	
4	AAIL		-		_	Cars		-	6-	
			-0	100	100					
			-							
		tarte a		-		11503	16939			8.37
					1000					

a point is that now refined and an

Avapting to Britan, Fulley, et al., the effect of eigenstic method on introduct time in one is not more. The law, therefore and eigens for executants

Report

Dealer, Dervice), and Dis-

offert. If it was emiled re

Letter to the Editor of Personnel Administrator

Blitor.

•

owner of a company with about 100 employees, I found isloon's recent article, "The other side of the anaking controversy," interesting. I'd like to there my own since I've had once experience with the issue.

DRAFT

5/2/83

but vocal group of employees have pressured me to has socking or to segregate sockers and monomolers. ay first approached me I considered the situation but, cal reasons, decided against any restrictions.

implementing smoking policies would have required take action against good implayees who have worked for quite some time. Second, to implement a socking would have discupted my company's work process, since. any offices, coployees with similar skills and impossi-s work ingether.

more, once I took a hard look at the situation, I dis-I that the wast majority of my employees were neither if nor particularly interested in the problem.

not read any of Heis' articles to which Solmon referred a I considered the economic aspect of the argument on non's article was based. But common conse supprite erranging people, changing policies, implementing reons, and disrupting my workforce won't save me money.

Letter -

R. REDACTED MATERIAL

```
What's rest to the Port Port Port Port
                                                                               the revenue a discussion of
                                                                                                               REDACTED
                                                                                                                      PEDACTED
                                                                         District over Monthelia and Mindog v. Sciencelidene Thiopen and Lander (1992)
                                                                               Digition to appendition in Fach Bindington, trainmentation
                                                                               abmountaint over disartering. As an extended that
                                                                                   STATES OF ALL MARKED IN SITE AND AND A STATES
                                                                                   Southernoon with "monorings dipon day suscesses bacad
1642 inc. 10-4
                                                                         Stiffer der für, ihnemeningit im cheinter rittlich
                                                                            stumentate inter function on electric Exclusion
Witchied-
                                                                               Conditioned the Section of material provide and 
transforment for Managirets Charles. 
Conductor Section (2010) Long Section (2010) 
Longers Section 2. Designation
                   the set. Another the conditions of the set o
```

Scene text (IIIT5K, SVT, IC03, IC13, SVHN) Receipts (SROIE)

Documents (RVL-CDIP)

Synthetic Data Generation: Main stages



OCR steps



a kiváló tudósok.



DIE: removing handwriting

F second I
Budapest székesfőváros polgármestere
Előadói io
<u>367433</u> _sam 19y9-170a-a
Egyúttal elintézett számok :
A kindói etastiáson kivill a keelle A kindói etastiáson kivill a keelle hivadinak adoi egyáb etastiáson A kindói etastiáson A kindói etastiáson kivill a keelle A kindói etastiáson A kindói etastiáson A kindói etastiáson A kindói etastiáson A kindói etastiáson Ki Hivat urmat Kindón etacet: Kindón etacet:
johund bound

Budapest székesfőváros polgármestere Előadói ív Előszám : szám 104 fi... Együttal elintézett számok : Tárgy: A kindól utasításon kívül a kezelőhivatalnak adott egyéb utasításoki Kiadta : Letisztázta : Összeegyeztellt : Kiadóba érkezett : Bude peak sathiasi livino hikaleryo mil

Bootstrapping: iterative improvement



Advantage of DIE

Advantage:

- Text Detection:
- Text Recognition

5	4 A rat	Perme forestando Perme forestando Maryar Talurati India rel 5 Solucit folio	6
2	16	HARVAR TAVIRATI IRADA DE 12 There Tavisa 0.0	2
5	220 1	Tinden-jog-jemtartázával-Kéziratnak-tekintéjdőzlázi-zok-zorozitaljecoc	>
		23 BIZALMAS KÜLPOLITIKAI SZEMLE 23	
		24 51. 3761. 24	
7 .	old	A. 1944. juliur 29. 2	27
2 100		wand (1) 90	
, 1111	07410	200 200 magrow The elitter by 20 a first to be define	-
1	31	fusitt komenterokbol és icherjon (100 - 10	-
	35	hioraban alle	1
	35	Harverysztern nines en incloal 38 hi 38 fria. westdeutscher	
-	41	Beobachier a Dagy r sagibirter hersebere 13271 haffeniss. , he.	y
•	43	p nory a haborn uten add gybir49kolork falt tlenul al kelt 112	4
	44	nick. 44 47 A Stockherts Fidning on belon 578 tuddaitoidrak to-	5
	40	lentését közli a Budépent ellon int fett logitorel légitingin.	
	51	50 1 a czyvitaśce: filiki si tenceri un "r 59 wore". 50	-
	52	arkand hirek steringe usy 1996-01, here itil Kash r volt d52	
	53	at Jelováklába, 53	
	55	dákat, hogy szabutalják, anogy csak leheturdal@ -: {la_basiol=	
	64 6	Allateri tervet, pivel and hibrar al ri Mac intole Jotan t	
	02 9	sa max 368: 36; less a payer temesro.	68
-	69	azt filitja, hogy Sztenisalti is kolowea közit63. 70892er ra-	6
-	71	uedja Satójav miniszterolzolát és a uzyar katonzo vébetest.	6
	74	73 A nonzkyci rádio na za oddačnan a 73 zaviperi	
		nunicisok bojkottiljik až Imrédy iltal alkotott runkrazervezetet,	
Ka	loti_	erevonel. 77 * 76	
		78 Catolie winteken puningeneeb78 Matthe Prageral 7	9
	92	Jappez 180, vais int Breezt-Lite voz' alter 8087rit. 937 tipota 93	3
	94	ra da delre, Shull 61 /Scippion i le spiblintol (againguisticate	
	96	ra és Syugata, Syalanin giztaniszlautól nyi gyra és idiny gatrig további szovien szörenvendést és harzszánu selssés sligelulásít	6
	51	jelantett Sztaniszlen aberekán ifosinitát Pére isztov. a Bolotvia	8
10.	99	tovs ktol nyugatra a esovie gesepat ik Mrulagetak egy hárar had	10
1.	18Å	fore szoritották. Itt jalenler noves harok folyash.	10
		102	10
		102/12	

Raw document image

Cleaned document image


OCR results on private dataset (ÁBTL)

