Visual and Textual Feature Fusion for Document Analysis

Patricia Medyna Lauritzen de Lucena Drumond

Supervisor Prof. Dr. Teófilo Emidio de Campos Prof. Dr. Fabricio Ataide Braz

Programa de Pós-Graduação em Informática



11th November 2022

Patricia Medyna Drumond

Visual and Textual Feature

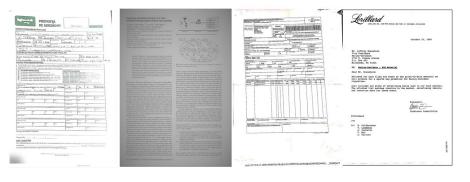
11th November 2022

1/24

- Introduction
- Background
- Proposal
- Experiments
- Results
- Schedule
- Reference

Introduction

Motivation



(a)

Document Images

(c)

(b)

Patricia Medyna Drumond

Visual and Textual Feature

イロト イヨト イヨト イヨト

(d)

Problem

- Document information extraction
- Wide variety of layout
- Visual and textual features

э

Main Goal

• Implement and evaluate document processing methods that combine textual information and layout with low computational cost.

Specific Goals

- to propose the joint of textual and layout features for information extraction.
- to evaluate this approach for document classification and page segmentation.
- to compare the models with baselines.

Contributions

- A novel approach to fuse textual and layout information.
- The simple yet effective model.
- The source code of our library, which is available from https://github.com/patriciamedyna/LayoutQT

7 / 24

Table: Document AI: models, modality, backbone and datasets.

Models	Modality	Backbone	Datasets	
Asim et al. (2019) [1]	T + I	InceptionV3	Tobacco-3482	
Asini et al. (2019) [1]	1 + 1	Multi-channel CNN	RVL-CDIP	
Audebert et al. (2020) [2]	T + I	Multimodal	Tobacco-3482	
		Neural Network	RVL-CDIP	
		Transformer	FUNSD	
LayoutLM (2020) [8]	T + L	BERT	SROIE	
		DENT	RVL-CDIP	
Wiedemann and	T + I	CNN (VGG16)	Tobacco800	
Heyer (2021) [7]	1 + 1	MLP	German datase	
Braz et al. (2021) [3]	I only	CNN (VGG16)	Tobacco800	
Diaz et al. (2021) [5]	TOnly	EfficientNet	AI.Lab.Splitter	
			FUNSD	
	T + L + I		SROIE	
Lavout Mu2 (2021)[0]		Transformer	CORD	
LayoutLMv2 (2021)[9]			Kleister-NDA	
			RVL-CDIP	
			DocVQA	
		< □ ▶ ·	(四) (き) (き)	
Maduna Duumand	Manal and Ta		11th Neurophan 20	

LayoutQT



Preprocessing

Processing Language Model

æ

9/24

イロト イヨト イヨト イヨト

LayoutQT

Expression	Description
xxbob	begin-of-block
xxeob	end-of-block
xxbcet	begin-of-center-text
xxecet	end-of-center-text
xxQhi_vj	Quadrant tag, line hi and column vj
xxPk	Page tag, number k

xxQ00_00

	xxQ00_00	xxQ00_01		xxQ00_m-1		
	Univ	ersal Language Model Fir	e-tuning for Text Classificatio	ia i		P901
	-					(211_00 website 2018
		Jeremy Howard"	Schooling Rader			
		fastai	Insight Centre, NUI Galway			scu
		Inversity of San Francisco	Aylies Ltd., Dublin			Wir 1001.0614645
		jýfast.ai	sebastian@roder_ic			00_HOD+x deem
	0	0		0.00		
	xxQ01_00	xxQ01_01		xxQ01_m-1		P901
		Abstract				Q30:00 obshi Universal Language Model Fine-tuning for Text Classifica
			While Deep Learning models have state-of-the-art on many NLP tasks, the	e actorsed		web xx200_03
-	Frankley In	aroler learning has greatly in-	are trained from scatch, requiring large	re datarets.		P001
		peter vision, but existing ap-	and days to converge. Research in NL	P focused		(30.00 sobob lerenty Heward"
R		and training from scratch.	recordy on successfully transfer that	ser et al.		a si
	We propos	Universal Language Model	2007). For inductive transfer, fire-transfer, fire-	at 2011		elementary of San Francisco
10	Fine-tuning	CLMPIT), an effective trans-	a simple transfer todinique that only	i terpota a		Hertel
23 May .	for learning	method that can be applied to N.P. and introduce techniques	raodel's first layer, has had a large impo	act in prac-		web x+001_61
		v fer fire-turing a language	tice and is used in prost state-of-the-a Recert approaches that concutinate m	et models.		
	model. Ou	r method significantly corper-	derived from other tasks with the input of			9901
н		are of the art on six text clas-	lavers (Peters et al., 2017; McCann et	ul., 2007;		Q31_01 website websiter Abstract weekent
	NEGATION IN	ills, reducing the error by 18- mainties of datasets. Further	Peters et al., 2018) still train the main t			web xxQ01_61
		only 100 labeled examples, if	from control and more protocical onto found pursuancers, limiting their confulne			ann
		performance of training from	In fight of the bracity of pretraining			Q31_00 soboti sobort inductive transfer learning has greatly invision or
	senitch on	100 < more data. We open- tertrained models and code ¹ .	et al., 2010s, we should be able to do			hert pacted computer vision, but existing ap-samet.
Fł.	PERSONAL COMP.	and a second second second	randomly initializing the remaining par- per models. However, inductive transfe	tameters of		bert prooches in NLP still require task-specific specific
5	I Introduct		taning has been unsuccessful for NLP (or van sine-		locet modifications and baining from scratch, specel
R		Contract has had a large strengt	2016). Dai and Le (2015) first prog			doot We propose Universal Language Model assoct
1801.06146v5	inductive training	en tearning has had a targe impact ion (CV). Applied CV models (in-	taning a language model (LM) has requi	ine millions		doot Fine faming (ULMFIT), an effective trans-oxecet
E	chaling object	detection, elassification, and sog-	of in-domain documents to achieve go manner, which severely limits in applica			beet for learning method that can be applied to secon bent any task in NEP, and introduce techniques secont
arXiv I	incitation y are t	arely trained from weratche but in-	We show that not the idea of LM fine			dent day take in NDP, and introduce schwigues takent dent that are key for first-baring a language second
12		and from models that have been	per lack of knowledge of how to train	n them ef-		heat model. Our method script with notices, special
1		Reprint et al., 2004; Lone et al.,	feetinely has been hindering wher adopt			act forms the state of the art on six text clas execut
	2015a; He et al.	2016; Houng et al., 2017).	everit to small datasets and surfered to forgetting when fine-tured within classi	atastrophic		bort sification tasks, reducing the error by \$8-xmoot
		at on in a category of Natural Lan- te (NLP) tasks with real-world an-	rared to CV. NLP models are perically			bort 24N on the majority of datasets. Further-specif
		ig (NLP) tasks with real-world ap- as spars, fraud, and bet detection	kow and thus require different fire-tunin,	g methods.	20	dicet more, with only 200 labeled examples, it second
	Findal and Lin.	2007; Neai et al., 2011; Chu et al.,	We propose a new method, Universal	Language		locet matches the performance of training from sciences
	2012), emorpos	es response (Catagea et al., 2011),	Model Fine-tuning (ULMFIT) that addr issues and enables robest induction tran	tions these		doort scratch on 100x more data. We open-seecet
_		Accurces classification, such as	INCOMENDATION AND A DESCRIPTION OF THE PARTY	TELEVILL		surce our pretrained models and codel.
		erv (Reitblat et al., 2010).	models: The same 3-layer LSTM and	hitecture-		ineb xxC03_01
	"Margor//slg "Egent contribu	- Fast	with the same hyperparameters and			P901
	engotical and state	ATTACC, MEANING REACT OF BUILD	tions other than tuned deepost Reperper outperforms highly engineered models.	and trans-		PBC; D31 02 which 1 introduction
	And and a second second second		inference of the other states			000 xxQ03 (1
	xxOn-1 00	xxOn-1 01		xxOn-1 m-1		

イロト イヨト イヨト イヨト

3

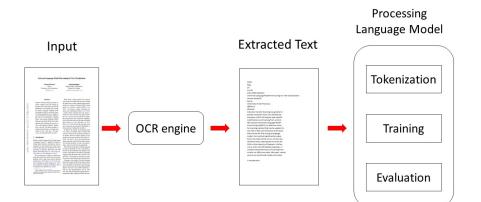
Algorithm 1 LayoutQT Algorithm

Input: multi page document

Output: tokenized text t

1:	for $page = 0, \dots, N-1$ do
2:	t + = add page token (where $+ =$ means insert symbol in string t)
3:	triage each word bounding boxes into line and group
4:	triage groups into coherent page columns
5:	for each group do
6:	t+ = quadrant coordinate of group top left corner
7:	for each text line in this group do
8:	check line centralization w.r.t. its page column center position
9:	if the line is centralized then
10:	t + = centre tag
11:	end if
12:	t + = textual contents of the line
13:	if the line is centralized then
14:	t + = centre tag
15:	end if
16:	end for
17:	t + = quadrant coordinate of group bottom right corner
18:	end for
19:	end for

• Baseline



A B b A B b

< □ > < 同 >

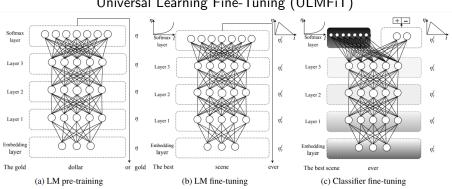
Page Stream Segmentation

- LSTM
- ULMFiT (AWD-LSTM)
- BERT

Document Type Classification

- ULMFiT (AWD-LSTM)
- BFRT

э



Universal Learning Fine-Tuning (ULMFiT)

Source: Howard and Ruder (2018) [5]

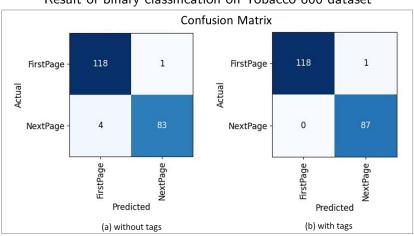
• Dataset:

- Tobacco-800: 1290 images: 831 training, 200 validation, and 259 test images.
- RVL-CDIP: 400,000 images (16 classes with 25,000 images per class): 320,000 training, 40,000 validation, and 40,000 test images.

• Experimental Settings

- The model is trained with a batch size of 128 and a sequence length of 150 for 100 epochs using NVIDIA Tesla V100 32GB GPU.
- LSTM: 256 nodes fully connected with activation "ReLU" and a dropout of 0.3, binary cross-entropy as a loss function with softmax activation and Adam as an optimizer.
- AWD-LSTM [6]: 3 layers, 1152 hidden sizes and 24M parameters.
- BERT [4]: 12 layers, 768 hidden sizes, 12 self-attention heads and 110M parameters.

Results



Result of binary classification on Tobacco 800 dataset

Patricia Medyna Drumond

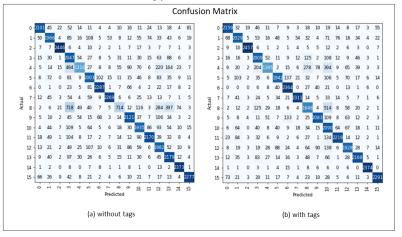
Table: Result of binary classification on Tobacco 800 dataset.

Model	Modality	Backbone	Accuraccy	F1-score
Wiedemann et al. (2019) [7]	text + image	VGG16*	91.1%	90.4%
Braz et al. (2021) [3]	only image	VGG16*	92.0%	91.9%
Braz et al. (2021) [3]	only image	EfficientNet-B0*	83.7%	81.9%
Baseline	only text	LSTM	84.1%	82.9%
LayoutQT	text + layout	LSTM	85.9%	86.1%
Baseline	only text	BERT _{BASE}	92.2%	92.0%
LayoutQT	text + layout	BERT _{BASE}	93.0%	93.0%
Baseline	only text	ULMFiT (AWD-LSTM)	97.5%	97.9%
LayoutQT	text + layout	ULMFiT (AWD-LSTM)	99.5 %	99.1 %

э

Results

Result of the document types classification on RVL-CDIP dataset



< □ > < □ > < □ > < □ > < □ > < □ >

э

Results

Class	Document Type	Baseline AWD-LSTM	LayoutQT AWD-LSTM	Baseline BERT _{BASE}	LayoutQT BERT _{BASE}
0	letter	85.5%	87.8%	83.7%	86.0%
1	form	78.8%	81.3%	77.8%	77.3%
2	email	97.2%	97.6%	93.0%	96.0%
3	handwritten	84.9%	83.3%	63.6%	80.0%
4	advertisement	55.2%	58.5%	66.0%	70.0 %
5	scientific report	80.6%	78.2%	74.8%	80.3%
6	scientific publication	89.1%	92.1%	87.4%	89.0%
7	specification	91.9%	93.6%	90.7%	91.0%
8	file folder	31.9%	73.5%	64.0%	73.8 %
9	news article	86.4%	84.9%	78.8%	82.6%
10	budget	77.8%	84.0%	78.1%	82.3%
11	invoice	87.9%	89.9%	81.4%	85.9%
12	presentation	79.9%	77.8%	70.3%	<mark>81.1</mark> %
13	questionnaire	90.0%	89.5%	83.7%	87.9%
14	resume	93.6%	93.7%	<mark>98.6</mark> %	98.3%
15	memo	91.9%	92.5%	85.4%	90.0%
Average		80.4%	83.6%	80.1%	<mark>84.5</mark> %

Table: F1-score of the document types classification on RVL-CDIP dataset

Patricia Medyna Drumond

11th November 2022

э

20 / 24

 Engineering Applications of Artificial Intelligence – Qualis A1 (in review) Drumond, P. M. and Leite, L. and Campos, T. and Braz, F. "LayoutQT - Novel Preprocessing Combining Visual and Textual Features". (2022)

Table: Summary of research activities planning.

Activity	Nov/22	Dec/22	Jan/23	Feb/23
Experiments	\checkmark	\checkmark		
Baseline on VICTOR dataset	\checkmark			
Current model on VICTOR dataset	\checkmark			
Model parameter adjustment		\checkmark	\checkmark	
Training and validation on chosen datasets		\checkmark		
Thesis Writing	\checkmark	\checkmark		
Background Update + Related Works	\checkmark			
Methodology		\checkmark		
Results and Discussion		\checkmark		
Submit to board			\checkmark	
Wrap Up			\checkmark	\checkmark
Preparing the presentation of the thesis			\checkmark	\checkmark
Thesis defense				\checkmark

э

< A[™]

Reference

Asim, M. et al. (2019) **Two Stream Deep Network for Document Image Classification**. International Conference on Document Analysis and Recognition (ICDAR), 2019, pp. 1410-1416, doi: 10.1109/ICDAR.2019.00227.

Audebert N. et al. (2020) Multimodal Deep Networks for Text and Image-Based Document Classification. In: Cellier P., Driessens K. (eds) Machine Learning and Knowledge Discovery in Databases (ECML/PKDD) Communications in Computer and Information Science, Springer, March, 2020, vol 1167: 427-443.

Braz, F. A. et al. (2021). Leveraging effectiveness and efficiency in Page Stream Deep Segmentation, Engineering Applications of Artificial Intelligence, Volume 105, 2021, ISSN 0952-197, https://doi.org/10.1016/j.engappai.2021.104394.



Devlin, J., Chang, M., Lee, K. and Toutanova, K. (2019) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT (1) 2019: 4171-4186



Howard, Jeremy and Ruder, Sebastian. (2018). Universal Language Model Fine-tuning for Text Classification. 328-339. 10.18653/v1/P18-1031.



Merity, S., Keskar, N. and Socher, R. (2018). **Regularizing and Optimizing LSTM Language Models**. International Conference on Learning Representations, 2018: 1-13.



 Wiedemann, G. and Heyer, G. (2021). Multi-modal page stream segmentation with convolutional neural networks. Language Resources and Evaluation. 55.

 10.1007/s10579-019-09476-2.

Patricia Medyna Drumond

Thanks! Questions? Suggestion?



Patricia Medyna

patriciamedyna@aluno.unb.br