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

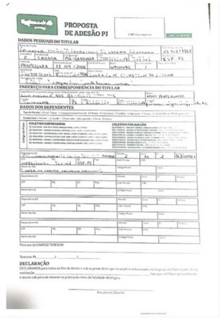
UnB Departamento de Ciência da Computação

11th November 2022

Outline

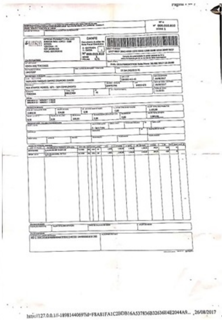
- Introduction
- Background
- Proposal
- Experiments
- Results
- Schedule
- Reference

- **Motivation**



(a)

(b)



(c)

Lorillard
LORILLARD, INC. • ONE PARK AVENUE, NEW YORK, N.Y. 10016-5895 • (212) 545-3000

October 18, 1989

Mr. Jeffrey Bleustein
Vice President
Harley-Davidson
3700 W. Juneau Avenue
P.O. Box 653
Milwaukee, WI 53201

RE: Harley-Davidson - POS Material

Dear Mr. Bleustein:

Enclosed for your files are stats of the point-of-sale material we will produce for a sports bag promotion for Harley-Davidson cigarettes.

Also included are stats of advertising being used in our test markets. The attached list matches creative to the market, advertising vehicle and insertion date for these stats.

Sincerely,

Constance Passuliotis

Attachment

/ec

cc: R. Goldbrenner
V. Lindsley
A. Pasheluk
T. Mau
G. Telford

(d)

Document Images

Introduction

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

Introduction

- **Main Goal**
  - Implement and evaluate document processing methods that combine textual information and layout with low computational cost.

Introduction

- **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.

Introduction

- **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`

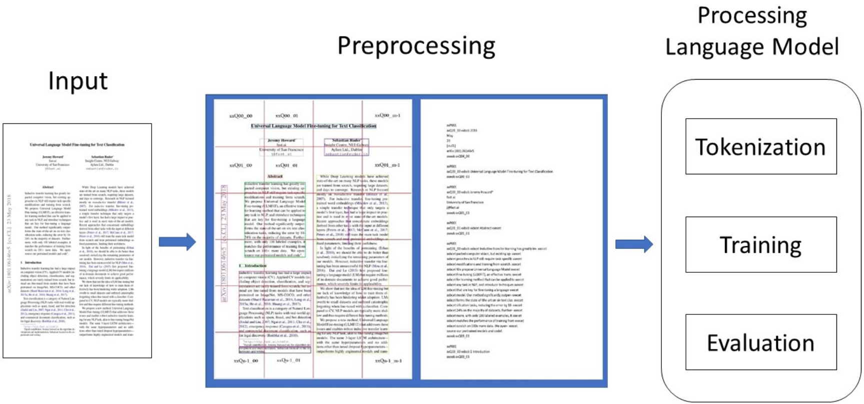
Background

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

Models	Modality	Backbone	Datasets
Asim et al. (2019) [1]	T + I	InceptionV3 Multi-channel CNN	Tobacco-3482 **RVL-CDIP**
Audebert et al. (2020) [2]	T + L	Multimodal Neural Network	Tobacco-3482 **RVL-CDIP**
LayoutLM (2020) [8]	T + L	Transformer BERT	FUNSD SROIE **RVL-CDIP**
Wiedemann and Heyer (2021) [7]	T + I	CNN (VGG16) MLP	**Tobacco800** German dataset
Braz et al. (2021) [3]	I only	CNN (VGG16) EfficientNet	**Tobacco800** AI.Lab.Splitter
LayoutLMv2 (2021)[9]	T + L + I	Transformer	FUNSD SROIE CORD Kleister-NDA **RVL-CDIP** DocVQA

Proposal

- **LayoutQT**

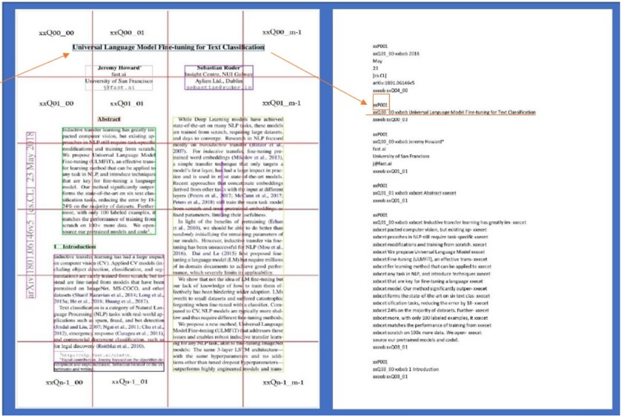


Proposal

- **LayoutQT**

xxQ00_00

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



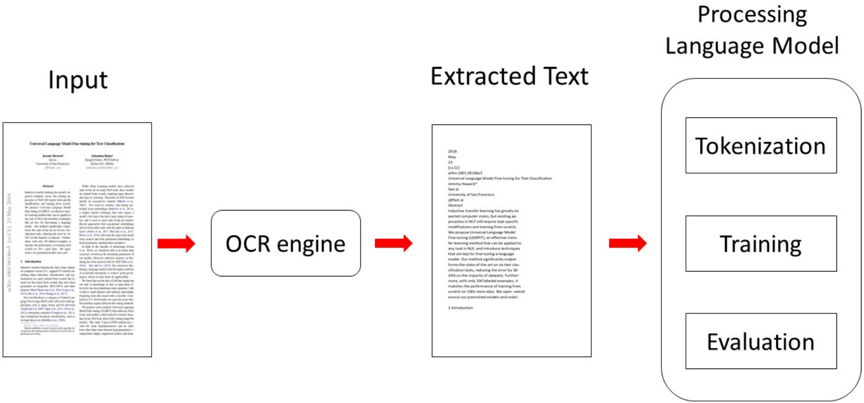
Proposal

Algorithm 1 LayoutQT Algorithm
Input: multi page document
Output: tokenized text t

```
for page = 0, ..., N − 1 do
    t+ = add page token (where + = means insert symbol in string t)
    triage each word bounding boxes into line and group
    triage groups into coherent page columns
    for each group do
        t+ = quadrant coordinate of group top left corner
        for each text line in this group do
            check line centralization w.r.t. its page column center position
            if the line is centralized then
                t+ = centre tag
            end if
            t+ = textual contents of the line
            if the line is centralized then
                t+ = centre tag
            end if
        end for
        t+ = quadrant coordinate of group bottom right corner
    end for
end for
```

Proposal

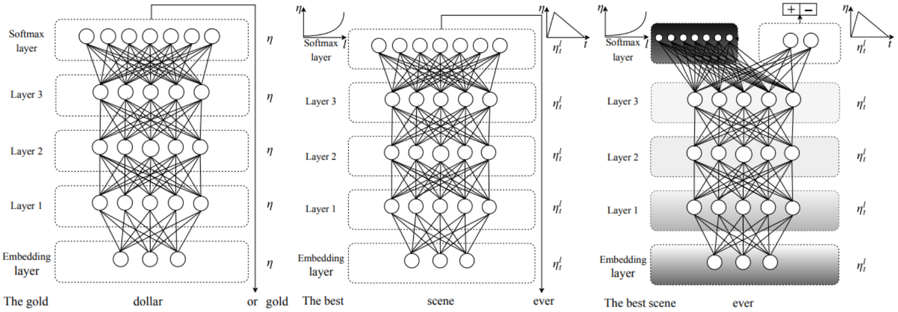
- **Baseline**



Experiments

- **Page Stream Segmentation**
  - LSTM
  - ULMFiT (AWD-LSTM)
  - BERT
- **Document Type Classification**
  - ULMFiT (AWD-LSTM)
  - BERT

Universal Learning Fine-Tuning (ULMFiT)



(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

Source: Howard and Ruder (2018) [5]

Experiments

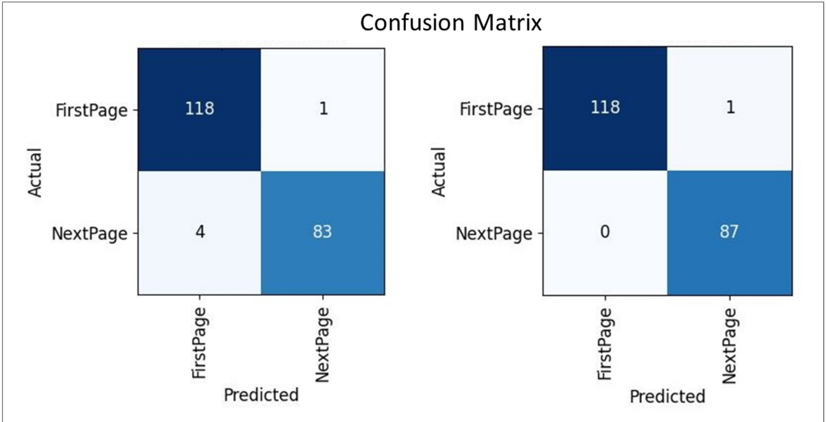
- **Dataset:**
  - Tobacco-800: 1290 images: 831 training, 200 validation, and 259 test images.
  - RVL-CDIP: 400,000 images (16 classes with 25,000 images per class): 320,000 training, 40,000 validation, and 40,000 test images.

Experiments

- **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



(a) without tags (b) with tags

Results

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**%

Results

Result of the document types classification on RVL-CDIP dataset

Confusion Matrix

(a) without tags (rows: Actual, columns: Predicted)

Actual \ Predicted	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	2101	45	22	52	14	11	4	4	10	16	11	24	13	18	4	81
1	50	1966	4	85	16	108	5	53	8	12	55	74	33	43	6	19
2	7	7	2446	6	4	10	2	2	1	7	17	3	7	7	1	3
3	15	30	1	2047	54	27	8	5	31	11	30	15	63	88	6	3
4	5	14	15	494	1316	27	8	8	55	90	70	6	220	164	23	7
5	8	72	0	81	9	2003	102	15	11	15	46	8	83	35	9	11
6	0	1	0	23	5	81	2287	1	7	66	4	2	22	17	8	2
7	12	45	3	54	4	59	9	2269	6	6	25	13	13	7	1	5
8	2	6	21	718	49	40	7	5	714	12	116	3	284	397	74	3
9	5	10	2	45	54	15	68	3	14	2121	37	7	106	34	3	2
10	4	44	7	109	5	64	5	6	18	30	1935	86	93	54	10	15
11	18	49	1	104	8	17	2	7	14	12	90	2170	39	32	8	4
12	13	21	2	49	25	107	10	6	31	86	59	6	1982	52	10	9
13	9	40	2	97	30	26	6	5	15	11	30	6	45	2179	12	4
14	1	2	0	8	0	7	8	1	1	8	1	0	13	2	2373	1
15	66	26	9	42	8	21	2	4	6	10	21	7	17	13	4	2277

(b) with tags (rows: Actual, columns: Predicted)

Actual \ Predicted	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0	2159	32	19	46	11	7	9	3	18	10	19	14	8	17	3	55
1	68	2029	5	53	16	48	5	54	32	4	71	76	16	34	4	22
2	9	10	2457	6	1	2	1	4	5	5	12	2	6	3	0	7
3	16	16	3	2009	52	11	9	12	125	2	108	12	9	46	3	1
4	9	20	2	204	1395	2	15	6	278	78	394	9	65	39	3	3
5	5	103	2	35	6	1942	137	21	32	7	106	5	70	17	6	14
6	0	0	0	6	8	40	2364	0	27	40	21	0	13	1	6	0
7	7	41	3	24	5	34	21	2311	14	5	33	14	5	7	1	6
8	2	12	2	125	29	18	6	4	1646	4	514	8	58	20	2	1
9	5	8	4	11	51	7	133	2	25	2083	109	8	63	12	2	3
10	6	64	0	40	8	40	9	18	34	15	2090	64	67	18	1	11
11	23	84	3	32	6	9	2	6	27	1	134	2219	14	12	2	1
12	8	19	3	19	28	88	24	4	64	90	138	6	1928	28	7	14
13	12	35	3	83	27	14	16	3	48	7	66	1	28	2168	5	1
14	1	1	0	3	1	4	15	1	8	6	6	0	6	0	2374	0
15	73	21	3	20	11	17	7	4	23	10	28	5	6	11	3	2291

Results

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

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%	**81.1**%
13	questionnaire	**90.0%**	89.5%	83.7%	**87.9%**
14	resume	93.6%	**93.7%**	**98.6**%	98.3%
15	memo	91.9%	**92.5%**	85.4%	**90.0%**
Average		80.4%	**83.6%**	80.1%	**84.5**%

Scientific Publication

- 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)

Schedule

Table: Summary of research activities planning.

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

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.

Thanks!
Questions?
Suggestion?

Patricia Medyna

patriciamedyna@aluno.unb.br