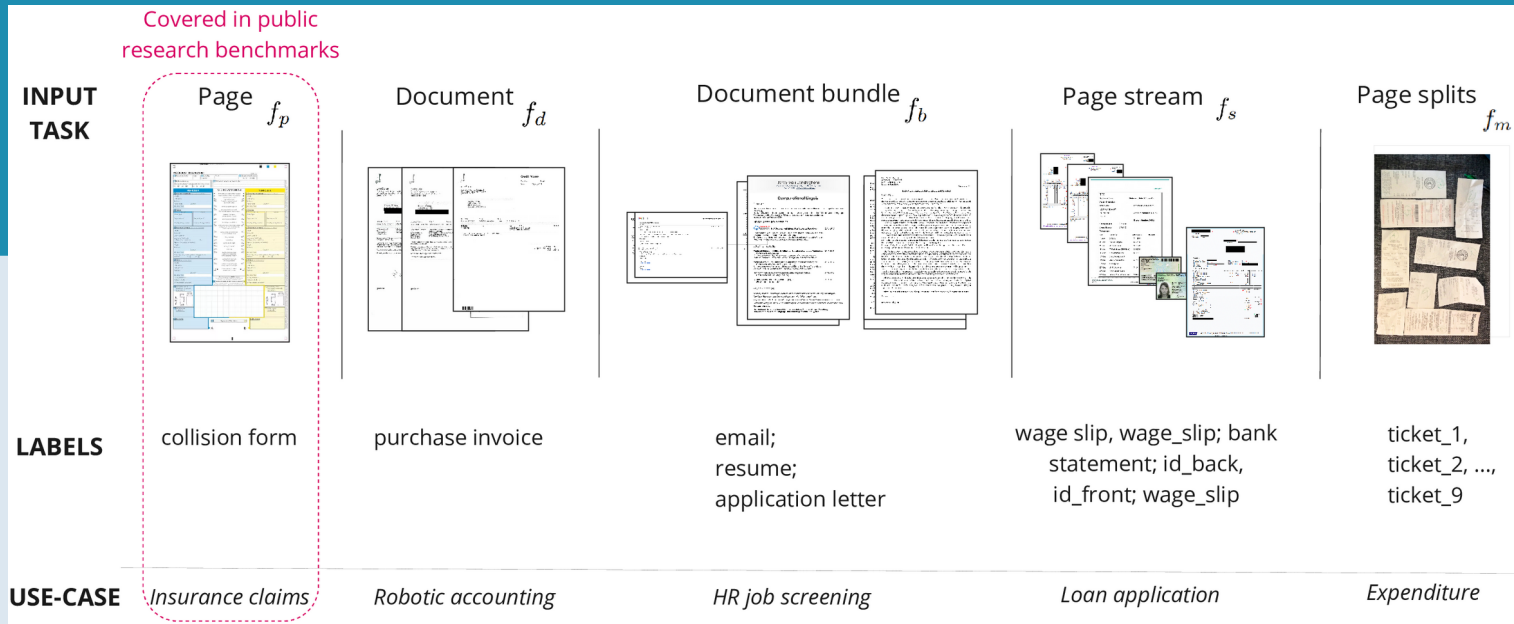
KU LEUVEN contract.fit CVC Centre de Visió per Computador

Beyond Document Page Classification: Design, Datasets, Challenges

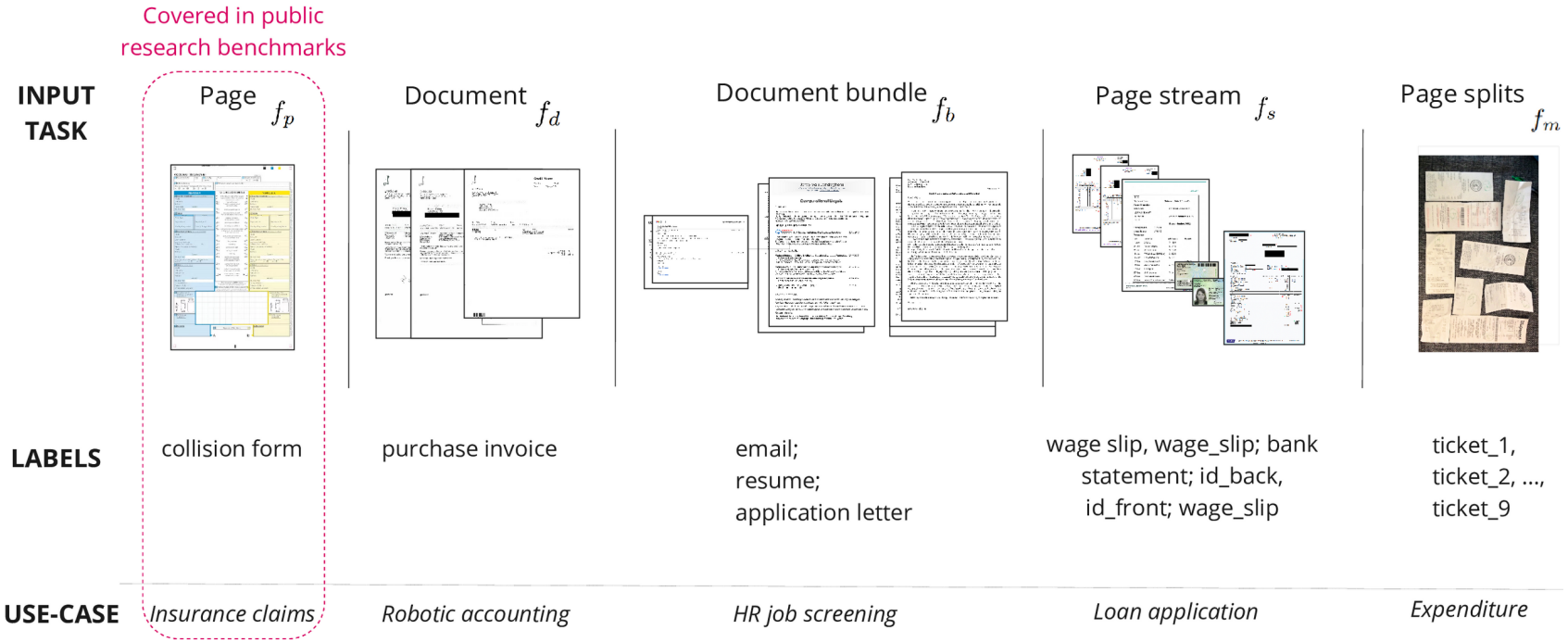
JAN 3-7 WACV 2024 WAIKOLOA · HAWAII

Jordy Van Landeghem, Sanket Biswas, Matthew Blaschko, Marie-Francine Moens



Beyond Document Page Classification: *TLDR;*

Move toward evaluation on <u>complete</u> document inputs, as they prevalently occur (**multi-page** *documents, bundles, page streams, and splits*) across various practical scenarios within real-world DC applications



Beyond Document Page Classification: *scope*

Problem – Observation - Contribution

*The practical task of **long document understanding** is largely underexplored due to several challenges in computation and how to efficiently represent long multimodal input.*

*Document classification should be **evaluated closer to practical applications** and task formulations (documents, bundles & page streams).*

We contribute **novel methodology** and demonstrate that document classification can be considered far from solved.

Beyond Document Page Classification: *goals*

Position paper with main objectives:

I. Benchmark closer to applied document classification scenarios

II. Experimental study on multi-page inference methods

III. Reflect on evaluation practices & moving beyond *iid* test sets

IV. Propose guidelines to foster document dataset construction efforts

Contributions

1. Redesign and formalization of multi-page DC scenarios
2. Construction of two novel datasets **RVL-CDIP_MP** and **RVL-CDIP-N_MP**
3. Comprehensive experimental analysis of the novel datasets
4. Survey of challenges stalling document classification progress

à `huggingface.co/datasets/bdpc/rvl_cdip_mp`

à `huggingface.co/datasets/bdpc/rvl_cdip_n_mp`

I. Problem formulation

Classification tasks

Page classification →

$$f_p : \mathcal{X}_p \rightarrow \mathcal{Y},$$
where $\mathcal{Y} = [C]$ for C mutually exclusive categories.

Document classification →

$$f_d : \mathcal{X}_d \rightarrow \mathcal{Y},$$
where $\mathcal{Y} = [K]$ for K mutually exclusive categories.

Document bundle classification →

$f_b : \mathcal{X}_b \rightarrow \mathcal{Y}$, where $\mathcal{Y}$ is a product space of B documents,
$\mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_B$, with $\{\mathcal{Y}_j = [K] : j \in [B]\}$.

Page stream classification →

$f_s : \mathcal{X}_d \rightarrow \mathcal{Y}$, where $\mathcal{Y}$ is a product space of L pages,
$\mathcal{Y} = \mathcal{Y}_1 \times ... \times \mathcal{Y}_L$, with $\{\mathcal{Y}_j = [C] : j \in [L]\}$.

Page splitting →

$$f_m : \mathcal{X}_p \rightarrow \mathcal{Y}, \text{ where } \mathcal{Y} = \mathbb{Z}^C.$$

II. Balancing research & applications

DU datasets and document sources

Dataset	Size	Data Source	Domain	Task	OCR	Layout
IIT-CDIP [28]	35.5M	UCSF-IDL	Industry	Pretrain	✗	✗
RVL-CDIP [17]	400K	UCSF-IDL	Industry	DC	✗	✗
RVL-CDIP-N [25]	1K	Document Cloud	Industry	DC	✗	✗
TAB [36]	44.8K	UCSF-IDL	Industry	DC	✗	✗
FUNSD [21]	199	UCSF-IDL	Industry	KIE	✓	✗
SP-DocVQA [35]	12K	UCSF-IDL	Industry	QA	✓	✗
OCR-IDL [6]	26M	UCSF-IDL	Industry	Pretrain	✓	✗
FinTabNet [58]	89.7K	Annual Reports S&P	Finance	TSR	✗	✓
Kleister-NDA [47]	3.2K	EDGAR	US NDAs	KIE	✓	✗
Kleister-Charity [47]	61.6K	UK Charity Commission	Legal	KIE	✓	✗
DeepForm [48]	20K	FCC Inspection	Forms broadcast	KIE	✓	✗
TAT-QA [61]	2.8K	Open WorldBank	Finance	QA	✓	✗
PubLayNet [60]	360K	PubMed Central	Scientific	DLA	✗	✓
DocBank [30]	500K	arxiv	Scientific	DLA	✓	✓
PubTabNet [59]	568K	PubMed Central	Scientific	TSR	✗	✓
DUDE [53]	40K	Mixed	Multi-domain	QA	✓	✗
Docile [45]	106K	EDGAR & synthetic	Industry	KIE	✓	✗
CC-PDF [51]	1.1M	Common-Crawl (2010-22)	Multi-domain	Pretrain	✗	✗

DC datasets - lacking variety, scale and multi-page

Proposed multi-page datasets

Dataset	Purpose	#d	#p	$\lvert\mathcal{Y}\rvert$	Language	Color depth
NIST [8]	f_s		5590	20	English	Grayscale
MARG [32]	f_s		1553	2	English	RGB
Tobacco-800 [62]	f_s		800	2	English	Grayscale
TAB [36]	f_s		44.8K	2	English	Grayscale
Tobacco-3482 [23]	f_p		3482	10	English	Grayscale
RVL-CDIP [17]	pre-training, f_p		400K	16	English	Grayscale
RVL-CDIP-N [25]	f_p, OOD		1002	16	English	RGB
RVL-CDIP-O [25]	f_p, OOD		3415	1	English/Mixed	RGB
RVL-CDIP_MP	f_d	±400K	$\mathbb{E}[L] = 5$	16	English	Grayscale
RVL-CDIP-N_MP	f_d, OOD	1002	$\mathbb{E}[L] = 10$	16	English	RGB

III. Multipage inference experiments

Tease some issues and strategies when naively scaling beyond page-level DC.

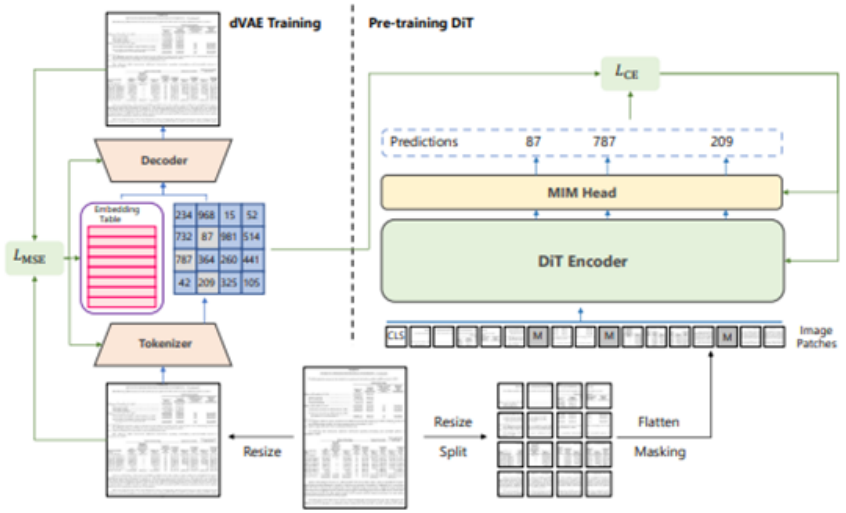
Inference	Strategy	Scope
sample	first	page
	second	page
	last	page
sequence	max confidence	page
	soft voting	page
	hard voting	page
grid	grid	document
document	(not tested)	document

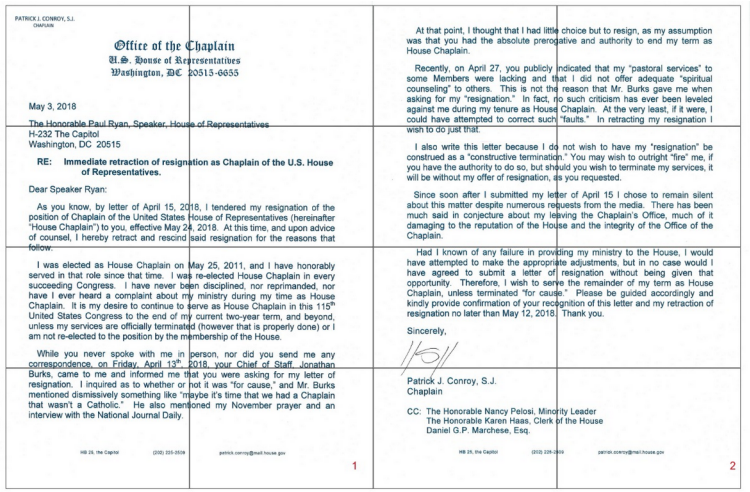
$$\text{MaxConf}(x, y) = \arg\max_{\substack{l \in [L] \\ k \in [K]}} [\tilde{f}_p(x, y)]_k^l$$

$$\text{SoftConf}(x, y) = \arg\max_{k \in [K]} \sum_{l=1}^{L} [\tilde{f}_p(x, y)]^l$$

$$\text{HardVote}(x, y) = \arg\max_{k \in [K]} \sum_{l=1}^{L} e_{\hat{y}^l},$$

Model: Document Image Transformer





Office of the Chaplain
U.S. House of Representatives
Washington, DC 20515-6555

May 3, 2018

The Honorable Paul Ryan, Speaker, House of Representatives
H-232 The Capitol
Washington, DC 20515

RE: Immediate retraction of resignation as Chaplain of the U.S. House of Representatives.

Dear Speaker Ryan:

As you know, by letter of April 15, 2018, I tendered my resignation of the position of Chaplain of the United States House of Representatives (hereinafter "House Chaplain") to you, effective May 24, 2018. At this time, and upon advice of counsel, I hereby retract and rescind said resignation for the reasons that follow.

I was elected as House Chaplain on May 25, 2011, and I have honorably served in that role since that time. I was re-elected House Chaplain in every succeeding Congress. I have never been disciplined, nor reprimanded, nor have I ever heard a complaint about my ministry during my time as House Chaplain. It is my desire to continue to serve as House Chaplain in this 115th United States Congress to the end of my current two-year term, and beyond, unless my services are officially terminated (however that is properly done) or I am not re-elected to the position by the membership of the House.

While you never spoke with me in person, nor did you send me any correspondence, on Friday, April 13th, 2018, your Chief of Staff, Jonathan Burks, came to me and informed me that you were asking for my letter of resignation. I inquired as to whether or not it was "for cause," and Mr. Burks mentioned dismissively something like "maybe it's time that we had a Chaplain that wasn't a Catholic." He also mentioned my November prayer and an interview with the National Journal Daily.

1

At that point, I thought that I had little choice but to resign, as my assumption was that you had the absolute prerogative and authority to end my term as House Chaplain.

Recently, on April 27, you publicly indicated that my "pastoral services" to some Members were lacking and that I did not offer adequate "spiritual counseling" to others. This is not the reason that Mr. Burks gave me when asking for my "resignation." In fact, no such criticism has ever been leveled against me during my tenure as House Chaplain. At the very least, if it were, I could have attempted to correct such "faults." In retracting my resignation I wish to do just that.

I also write this letter because I do not wish to have my "resignation" be construed as a "constructive termination." You may wish to outright "fire" me, if you have the authority to do so, but should you wish to terminate my services, it will be without my offer of resignation, as you requested.

Since soon after I submitted my letter of April 15 I chose to remain silent about this matter despite numerous requests from the media. There has been much said in conjecture about my leaving the Chaplain's Office, much of it damaging to the reputation of the House and the integrity of the Office of the Chaplain.

Had I known of any failure in providing my ministry to the House, I would have attempted to make the appropriate adjustments, but in no case would I have agreed to submit a letter of resignation without being given that opportunity. Therefore, I wish to serve the remainder of my term as House Chaplain, unless terminated "for cause." Please be guided accordingly and kindly provide confirmation of your recognition of this letter and my retraction of resignation no later than May 12, 2018. Thank you.

Sincerely,

Patrick J. Conroy, S.J.
Chaplain

CC: The Honorable Nancy Pelosi, Minority Leader
The Honorable Karen Haas, Clerk of the House
Daniel G.P. Marchese, Esq.

2

III. Results

Proof of concept

Strategy	Acc↑	F1↑	$F1_M$ ↑	ECE↓	AURC↓
f_p\$ [29]	93.345	93.351	93.335	0.075	0.010
first	91.291	91.286	91.271	0.073	0.014
second	87.295	87.305	87.277	**0.070**	0.029
last	85.091	85.060	85.028	0.072	0.038
MaxConf	**91.407**	**91.453**	**91.344**	0.124	0.006
SoftVote	91.220	91.185	91.236	0.134	**0.004**
HardVote	85.995	86.182	85.781	0.085	0.018
grid	72.642	72.045	73.266	0.109	0.042

Table 4. Base classification accuracy of DiT-base [29] (finetuned on RVL-CDIP) evaluated on the test set of `RVL-CDIP_MP` per baseline f_d strategy. Best results per metric are boldfaced. \$ refers to our reproduction of results.

***Hypothesis**: Summary-detail document construction*

Inefficient (L pages) and dependent on calibration of f_p

Dataset	Strategy	Acc↑	Δ
`RVL-CDIP_MP`	first+second$^{(*)}$	93.795	2.504
	first+last$^{(*)}$	93.675	2.384
	second+last$^{(*)}$	89.709	−1.583
	first+second/last$^{(*)}$	**94.454**	3.163
`RVL-CDIP_N_MP`	first+second$^{(*)}$	83.638	4.878
	first+last$^{(*)}$	83.130	4.370
	second+last$^{(*)}$	71.545	−7.215
	first+second/last$^{(*)}$	**84.553**	5.793

Table 6. Best-case classification accuracy indicated with $^{(*)}$ when combining 'knowledge' over different pages. Δ refers to the absolute difference with the first page only.

Multi-page document representations are promising for improving document classification

IV. Challenges and Guidelines

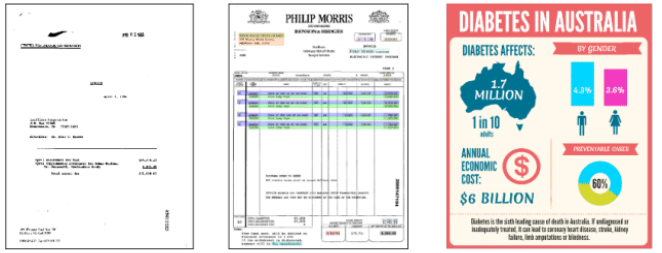
(1) Divergence of tasks

- Only page classification (f_p) research benchmarking
- DC task formulations can help to formulate f in practice

(2) Divergence of label space

- Too simplified label sets in benchmarking
  - K= 16 <-> industry reqs: K ~ 50-400
- Label noise
  - Relabeling campaigns
- MECE principle for construction of doc-level label sets

(3) Divergence of input data

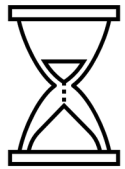


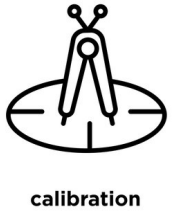
Public document collections
Data synthesis
Anonymization

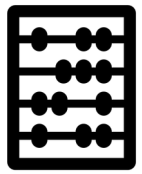
(4) Maturity of evaluation methodology

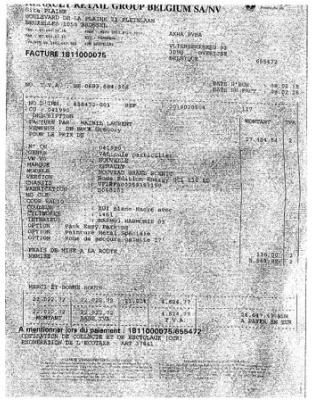
- *lid* assumption is unmaintainable in production

IV. (4) Evaluation beyond accuracy and *iid* setting





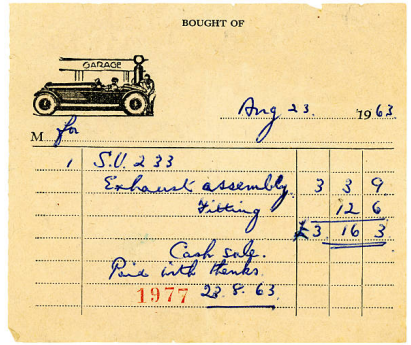




Covariate shifts

Subclass shift

Near OOD



Concept drift

Closing remarks

- Pivotal step forward in establishing methodology for multipage DC
- Covered challenges and limitations hindering progress
- Experimental study shows promise from advancing multi-page document representations and inference
- Recommendations for future DC dataset constructions efforts
    - Type and nature of document data
    - Variety and quality of the classification label set
    - Focus on DC scenarios closer to applications

Questions?
Future collaborations?

See you at WACV 2024!

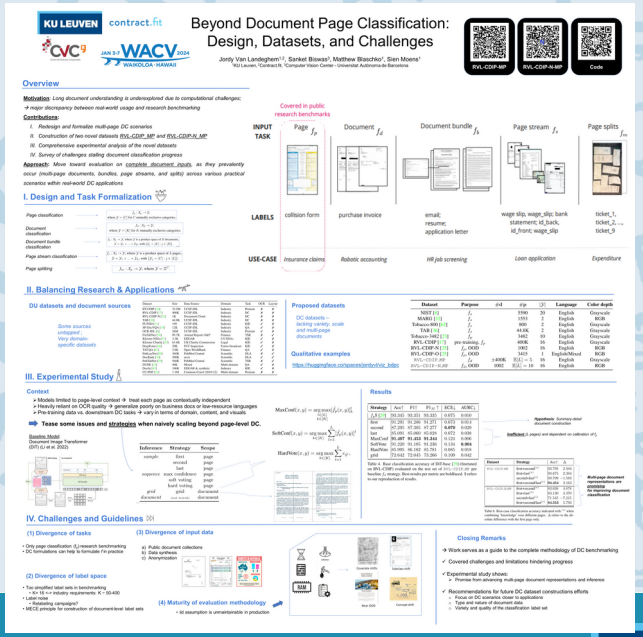
jordy@contract.fit

CODE

huggingface.co/bdpc/src

DATASETS

huggingface.co/datasets/bdpc/rvl_cdip_mp
huggingface.co/datasets/bdpc/rvl_cdip_n_mp



KU LEUVEN