NLP SUMMIT

Presented by John Snow LABS

A Unified CV, OCR, and NLP approach for scalable document understanding

Patrick Beukema, Ph.D
Senior Machine Learning Engineer
DocuSign

Misha Chertushkin
Senior Data Scientist
John Snow Labs

Outline

- Problem
- Previous research on Document AI
- Solution [CV, OCR, NLP]
- Data [Acquisition, Infrastructure, Database]
- CV model
- OCR model
- NLP model
- Conclusions and Future Work
- Acknowledgements

N.B. The project is currently pre-GA, and some details have been omitted.

NLP SUMMIT

Problem

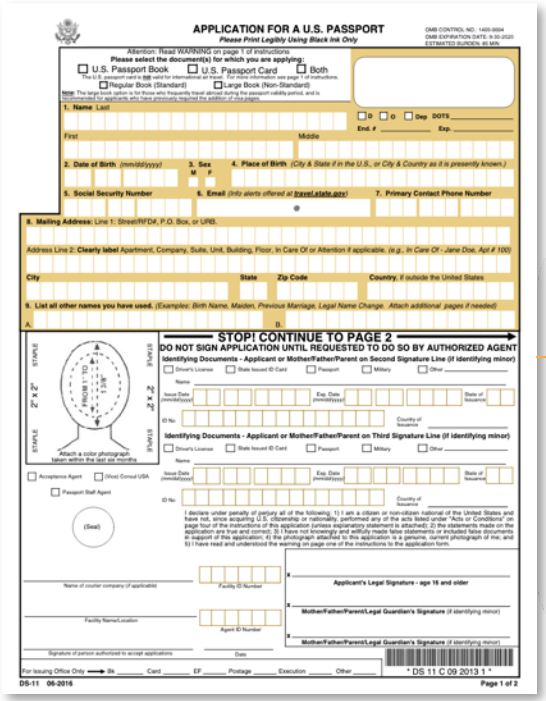
Automate extraction of structured information from document images.

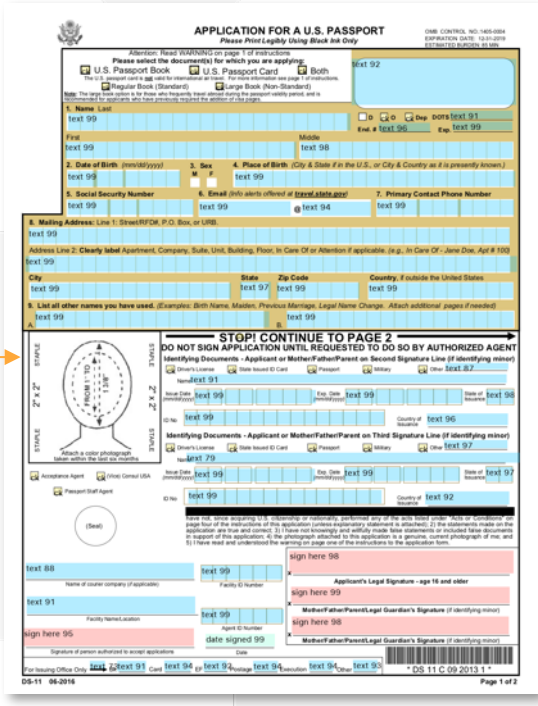
Examples:

- Contracts
- Tax forms
- Passport applications
- Invoices
- etc.

Challenges

- High and growing variation in layout
- Unbounded field type complexity
- Unstructured information





Desired outcome

NLP SUMMIT

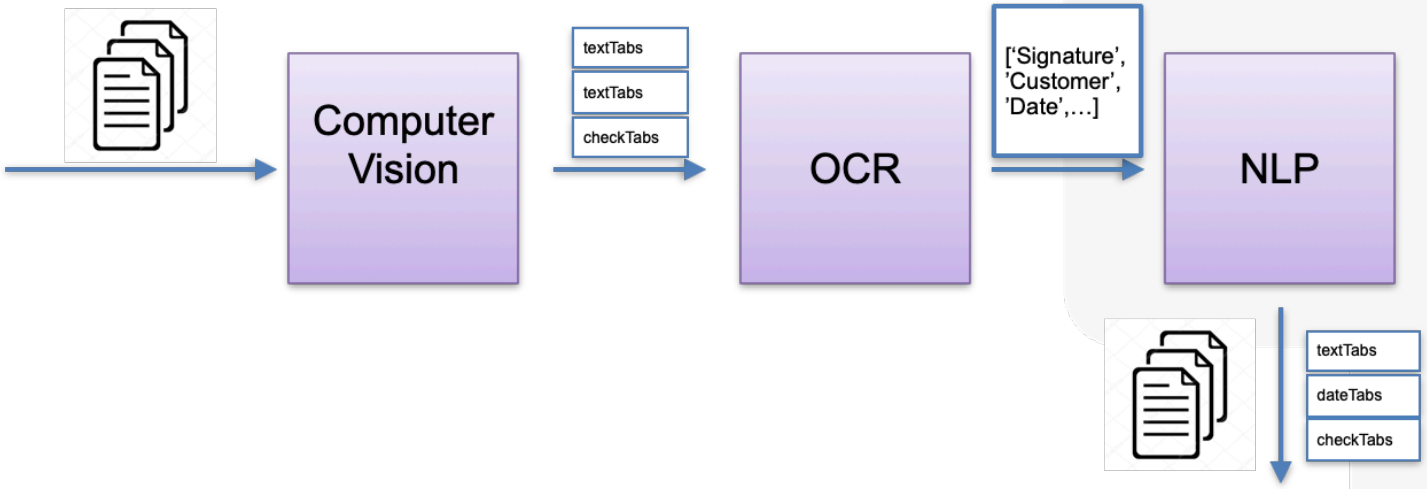
Research in Document AI

- Liu et. al. (2019) – Graph Convolutions
- Katti et. al. (2019) – Chargrid
- Denk et. al. (2019) – BERTGrid
- Majumder et. al. (2020) – Neighborhood Based

Research is increasing but it is still at an early development stage

Solution

- Humans create documents in whatever format best suits their immediate needs.
- Therefore, rules-based engines (template based, position based) will not scale.
- Ideal solution is to learn high level representations from data using AI.



NLP SUMMIT

Data

Acquisition

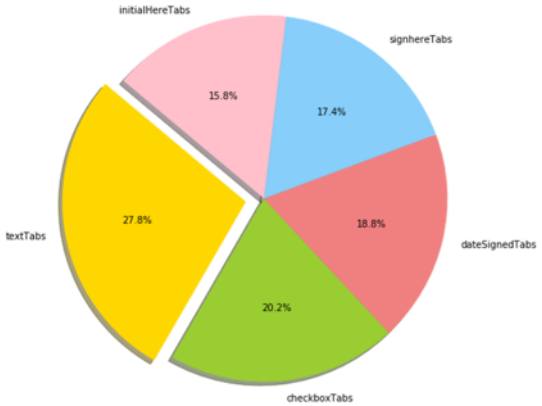
- Human labeled & QA'd with extensive iterations
- Perceptual-hashing for data split integrity and to ensure high variance.

record creation for near-real-time experimentation

- Deep learning framework agnostic

documents

- 10k+ perfectly labeled documents
- 1M+ labeled objects for CV/NLP



Infrastructure

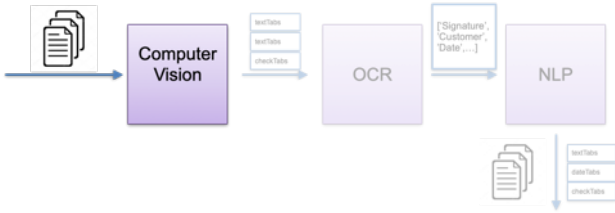
- Supports on-the-fly

Database

- 200k+ weakly labeled

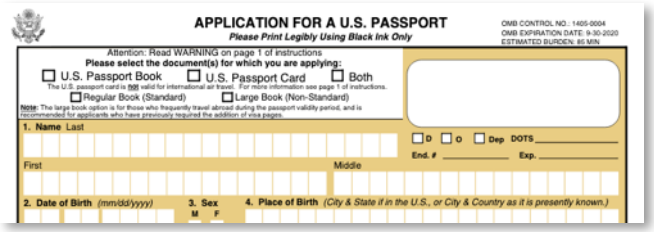
***Only publicly available (non-customer) data was used to build database and to train models.**

NLP SUMMIT

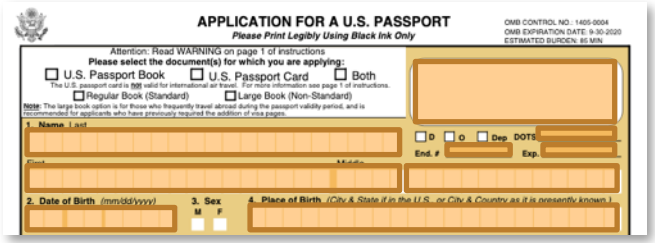


Computer Vision

- Label data as a conventional object detection task

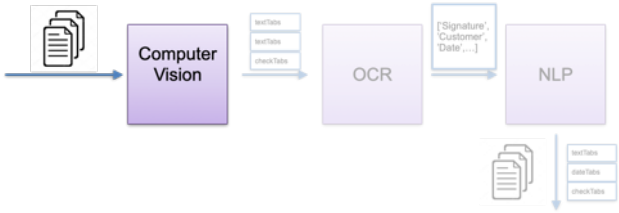


Unlabeled



Labeled

NLP SUMMIT



Unique CV challenges

Size variation

- Objects may be very small, very large or somewhere in between

- Objects can be densely packed or relatively sparse

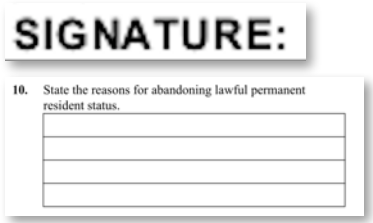
arbitrary aspect ratios

both long and short contextual dependencies

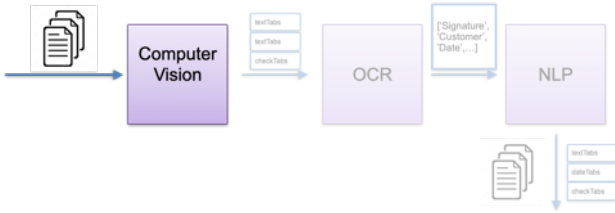
Density

Context

- Objects can exhibit

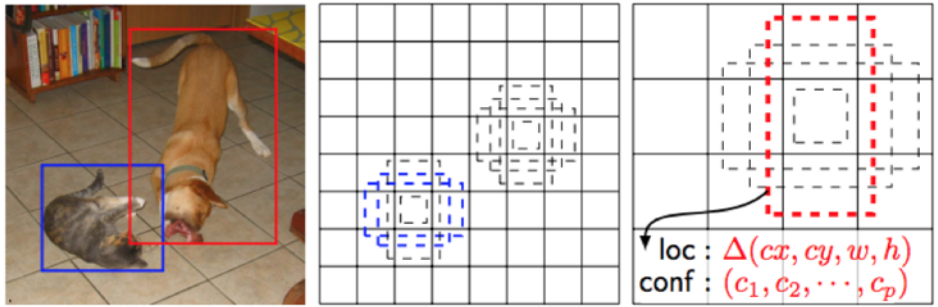


NLP SUMMIT



CV experimentation

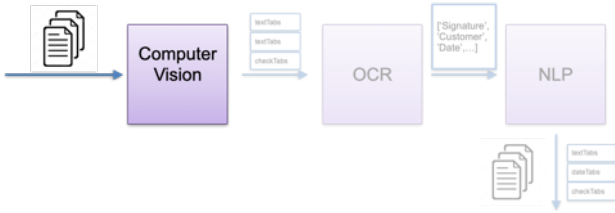
- Evaluating SSDs vs multistage detectors



(a) Image with GT boxes (b) 8 × 8 feature map (c) 4 × 4 feature map

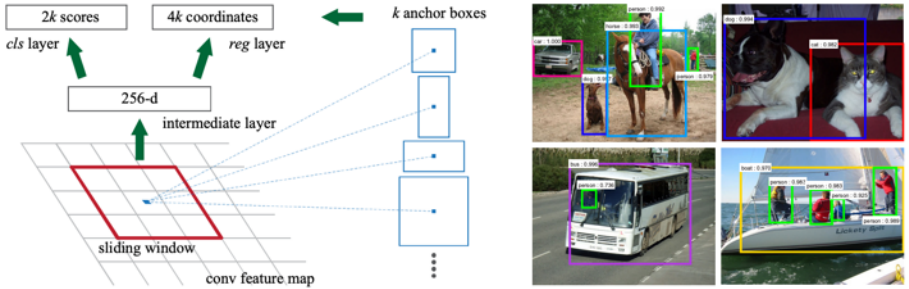
- While fast, single stage detectors do not detect with with sufficient accuracy.

NLP SUMMIT

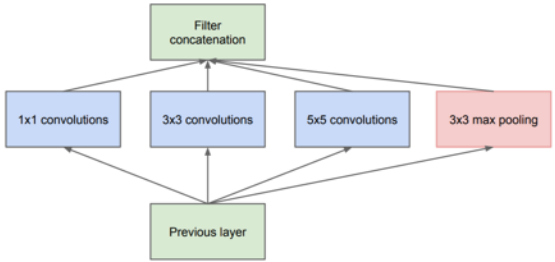


Computer Vision Model

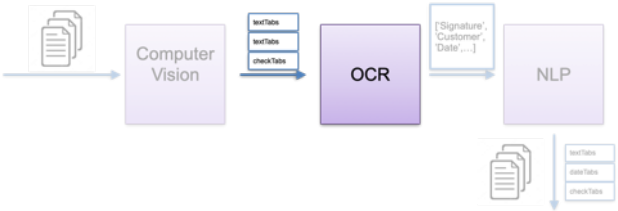
- RPN



- Feature Maps combining Inception and ResNet blocks

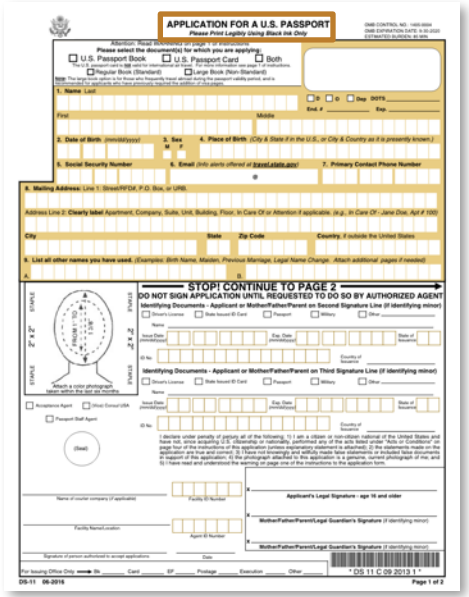


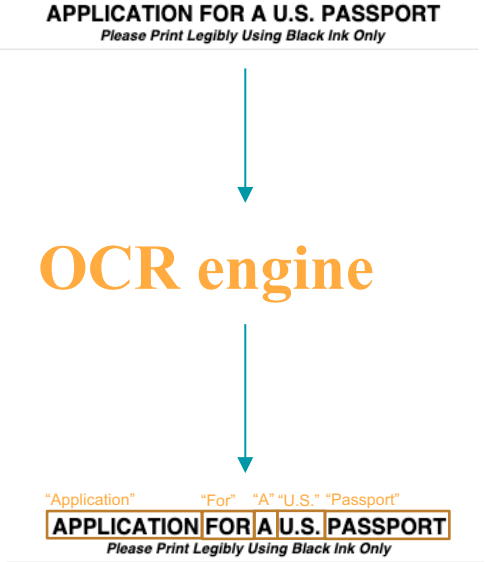
NLP SUMMIT



OCR

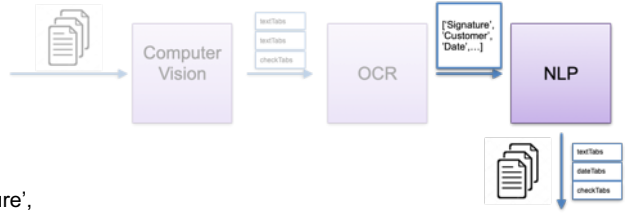
Custom service built on top of Tesseract tuned for speed and accuracy





text	xmin	ymin	xmax	ymax
APPLICATION	436	53	598	71
FOR	608	53	656	71
A	663	54	680	71
U.S.	689	53	726	71
PASSPORT	742	53	872	71
…	966	60	1149	69
Applicant's	757	1321	831	1334
Legal	837	1321	871	1334
Signature	877	1321	940	1334
-	939	1312	949	1341
age	953	1323	977	1334
16	1001	1321	1025	1331
and	1030	1321	1065	1331

NLP SUMMIT

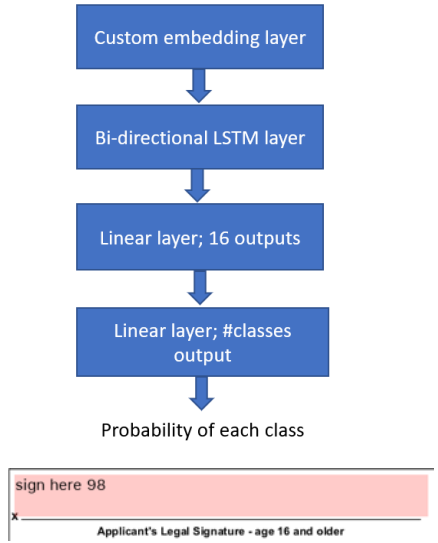


NLP

OCR output

text	xmin	ymin	xmax	ymax
APPLICATION	436	53	598	71
FOR	608	53	656	71
A	663	54	680	71
U.S.	689	53	726	71
PASSPORT	742	53	872	71
...	966	60	1149	69
Applicant's	757	1321	831	1334
Legal	837	1321	871	1334
Signature	877	1321	940	1334
-	939	1312	949	1341
age	953	1323	977	1334
16	1001	1321	1025	1331
and	1030	1321	1065	1331

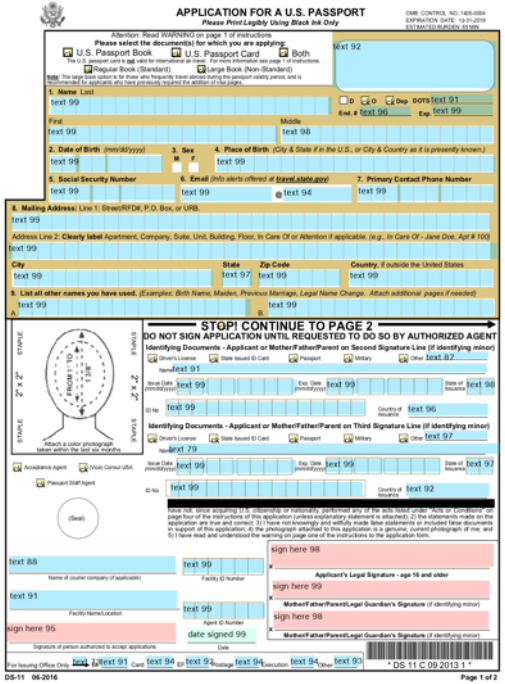
['Applicant's', 'legal', 'Signature', 'age', '16', 'and', Older']



Standard NLP classification model assigns tag type ('signature', 'initial', etc.) to each field.

NLP SUMMIT

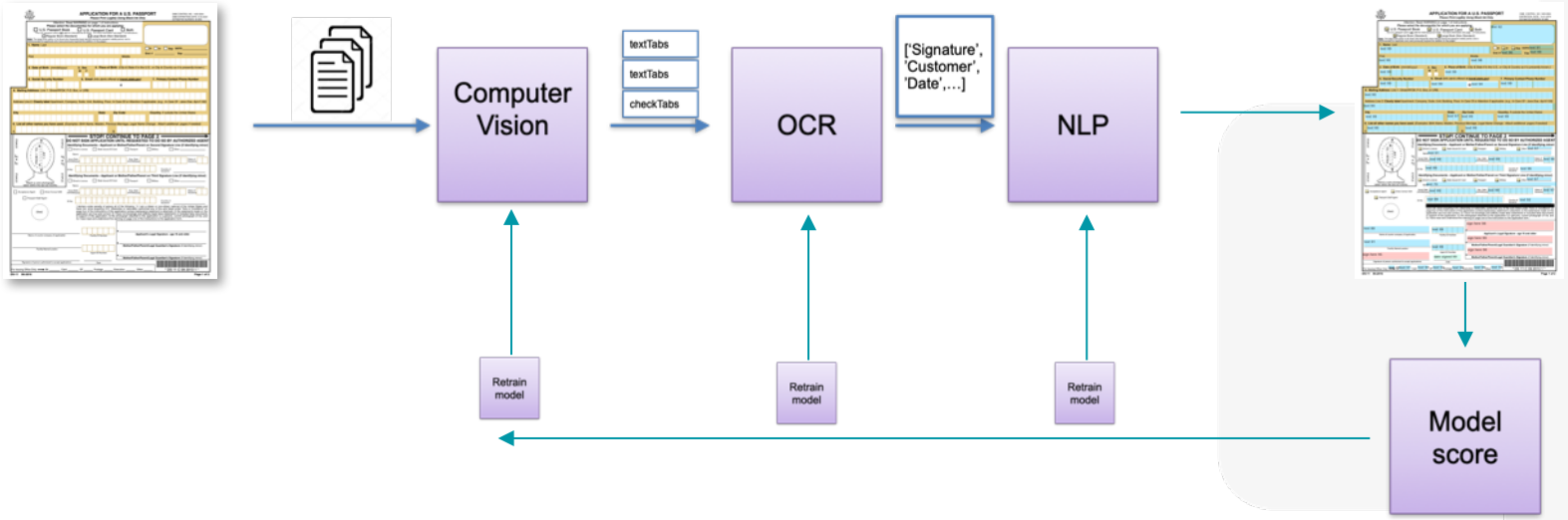
Measuring end-to-end performance



- Extensive testing on as diverse of a dataset as possible
- Score model by combining traditional CV metrics with traditional NLP metrics

NLP SUMMIT

Recap



NLP SUMMIT

Conclusions and Future Work

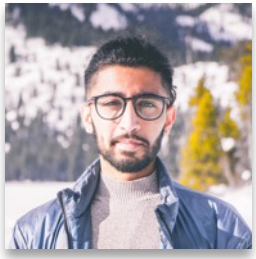
- Document understanding is well suited to an AI solution
- Highly performant models require extensive human-in-the-loop data curation
- SOTA CV models can be adapted to negative space object detection
- Speed accuracy tradeoffs are critical factors at every stage of development
- When designing an ML solution, it is valuable to attempt to emulate how humans perform a task

Acknowledgements

Agreement Intelligence Team

Matthew Braun
S. SDE

Michael Palazzolo
S. DS



Roshan Satish
PM

Vincent Yung
Director

Nathaniel Wichman
SDE

JohnSnowLabs Team

David Talby
CTO

Anju Agarwal
Solutions Lead

Veysel Kocaman
L. DS

NLP SUMMIT

www.nlpsummit.org

NLP SUMMIT

Presented by John Snow LABS