Can computers understand PDF documents as humans, or better?

Alexey Subach
Dual Lab

Alexey Subach,
Technical Lead
Dual Lab

Dual Lab

- Service provider company, 50+ engineers
- Experts in PDF. Active at PDF Association
- Lead veraPDF developer
- Specialize in graphic arts, science-intensive solutions
- Apply cutting-edge technologies
- Build dedicated teams for long-term collaboration

Alexey Subach,
Technical Lead
Dual Lab

What we will talk about today

- Recent achievements in AI
- How NNs make it work
- In which areas PDF can benefit from AI
- Designing AI pipelines for PDF to achieve your goals
- Problems remain unsolved

Alexey Subach,
Technical Lead
Dual Lab

Beating the greatest Go player

Alexey Subach,
Technical Lead
Dual Lab

Driving cars

By Grendelkhan - Own work, CC BY-SA 4.0

Alexey Subach,
Technical Lead
Dual Lab

Diagnosing pneumonia better than radiologists

Alexey Subach,
Technical Lead
Dual Lab

Artificial intelligence is here

- Finding computer viruses
- Sorting search results by relevance
- Scanning for fraudulent transactions
- Recommending hotels, films and restaurants
- Looking for your friends on photos
- Discovering new drugs

Alexey Subach,
Technical Lead
Dual Lab

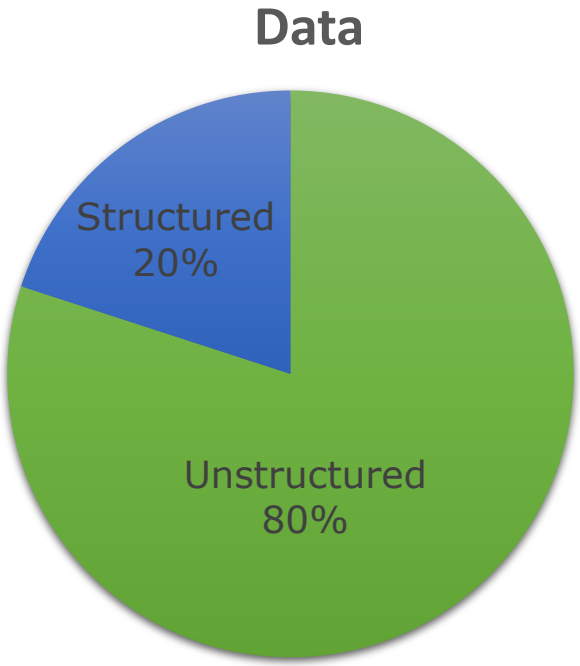
What about documents?

- ICDAR
- Papers dedicated to structure recognition
- Some research is based purely on PDF
- Emerging research on deep learning approaches
- Moderate-sized datasets
- Interesting projects like converting mockup image to HTML + CSS

Alexey Subach,
Technical Lead
Dual Lab

AI is all about unstructured data

- Images: photo, satellite
- Audio: voice recordings, songs, live calls
- Video: movies, vlogs, conferences
- Text: emails, logs, blogposts
- Composite formats, **including PDF**



Alexey Subach,
Technical Lead
Dual Lab

Tagged PDF?

Alexey Subach,
Technical Lead
Dual Lab

Tagged PDF?

- Still not default for most PDF producers

Alexey Subach,
Technical Lead
Dual Lab

Tagged PDF?

- Still not default for most PDF producers
- 3+ trillion files are already out there, 80% untagged*

* Keynote by Leonard Rosenthol, PDF Days Europe 2016

Alexey Subach,
Technical Lead
Dual Lab

Tagged PDF?

- Still not default for most PDF producers
- 3+ trillion files are already out there, 80% untagged*
- Among PDF/A family, only Level A requires tagged structure. But still it does not impose any conditions on how good this structure should be

* Keynote by Leonard Rosenthol, PDF Days Europe 2016

Alexey Subach,
Technical Lead
Dual Lab

Tagged PDF?

- Still not default for most PDF producers
- 3+ trillion files are already out there, 80% untagged*
- Among PDF/A family, only Level A requires tagged structure. But still it does not impose any conditions on how good this structure should be
- Poor quality of some tagged documents

* Keynote by Leonard Rosenthol, PDF Days Europe 2016

Alexey Subach,
Technical Lead
Dual Lab

Tagged PDF?

- Still not default for most PDF producers
- 3+ trillion files are already out there, 80% untagged*
- Among PDF/A family, only Level A requires tagged structure. But still it does not impose any conditions on how good this structure should be
- Poor quality of some tagged documents
- Transition will take some time

* Keynote by Leonard Rosenthol, PDF Days Europe 2016

Alexey Subach,
Technical Lead
Dual Lab

Tagged PDF?

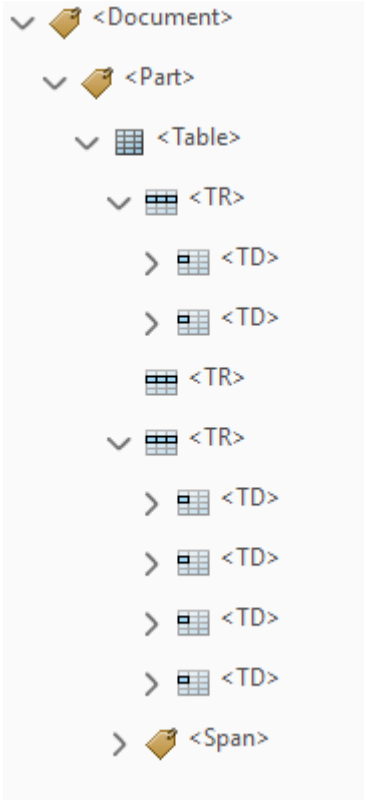
- Still not default for most PDF producers
- 3+ trillion files are already out there, 80% untagged*
- Among PDF/A family, only Level A requires tagged structure. But still it does not impose any conditions on how good this structure should be
- Poor quality of some tagged documents
- Transition will take some time
- Applicable even for tagged PDF (PDF/UA)

* Keynote by Leonard Rosenthol, PDF Days Europe 2016

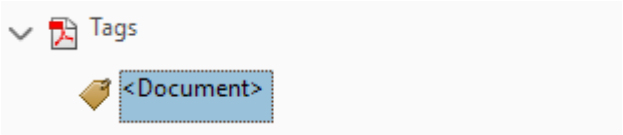
Alexey Subach,
Technical Lead
Dual Lab

Examples of poor tagging

Table cells: different number in rows, no col spans



PDF/A Level A



Alexey Subach,
Technical Lead
Dual Lab

Scientific publications

$$\begin{aligned}\nabla\mathbb{E}[F(x)] &\approx \frac{1}{n\sigma}\sum_{i=1}^{n} F(x+\sigma\epsilon_i)\epsilon_i \\ &= \frac{1}{n/2}\sum_{i=1}^{n/2}\frac{F(x+\sigma\epsilon_i)-F(x-\sigma\epsilon_i)}{2\sigma}\epsilon_i \\ &\approx \frac{1}{n/2}\sum_{i=1}^{n/2} D_{\epsilon_i}(x)\epsilon_i \\ &= \frac{1}{n/2}\sum_{i=1}^{n/2}(\nabla F\cdot\epsilon_i)\epsilon_i\end{aligned}$$

Now, the ϵ_i are effectively randomly drawn Gaussian vectors of size *width* · *height* · *channels*. By a well-known result, these vectors are nearly orthogonal; a formalization of this is in [7], which says that for an n-dimensional space and N randomly sampled Gaussian vectors $v_1 \dots v_N$,

$$N \leq e^{\frac{\delta^2 n}{4}}[-\ln(\theta)]^{\frac{1}{2}} \implies \mathbb{P}\left\{\frac{v_i \cdot v_j}{||v_i||||v_j||} \leq \delta \;\forall\; (i,j)\right\} = \theta$$

Thus, one can "extend" the randomly sampled vectors into a complete basis of the space $[0,1]^n$; then we can perform a basis decomposition on $\nabla F(x)$ to write:

for the top k classes $\{y_1, \dots, y_k\}$. In normal settings, given an image and label (x_i, y), generating an adversarial example (x_{adv}, y_{adv}) for a targeted y_{adv} can be acheived using standard first-order attacks. These are attacks which involve essentially ascending the estimated gradient $\nabla P(y_{adv}|x)$. However, in this case $P(y_{adv}|x_i)$ (and by extension, its gradient) is unavailable to the classifier.

To resolve this, we propose the following algorithm. Rather than beginning with the image x_i, we instead begin with an image x_0 *of the original target class.* Then y_{adv} will be in the top-k classes for x_0. We perform the following iterated optimization:

$$\begin{aligned}\epsilon_t &= \min \epsilon \text{ s.t. } \operatorname{rank}\left(P\left(y_{adv}|\Pi_\epsilon(x_{t-1})\right)\right) < k \\ x_t &= \arg\max_x P(y_{adv}|\Pi_{\epsilon_{t-1}}(x))\end{aligned}$$

where $\Pi_\epsilon(x)$ represents the ℓ_∞ projection of x onto the ϵ-box of x_i. In particular, we concurrently perturb the image to maximize its adversarial probability, while projecting onto ℓ_∞ boxes of decreasing sizes centered at the original image x_i, maintaining that the adversarial class remains within the top-k at all times. In practice, we implement this iterated optimization using backtracking line search to find ϵ_t, and several iterations projected gradient descent (PGD) to find x_t. Alternatingly updating x and ϵ until ϵ reaches the desired value yields an adversarial example that is ϵ-away

dual lab
Alexey Subach,
Technical Lead
Dual Lab

Such different tasks, so similar algorithms

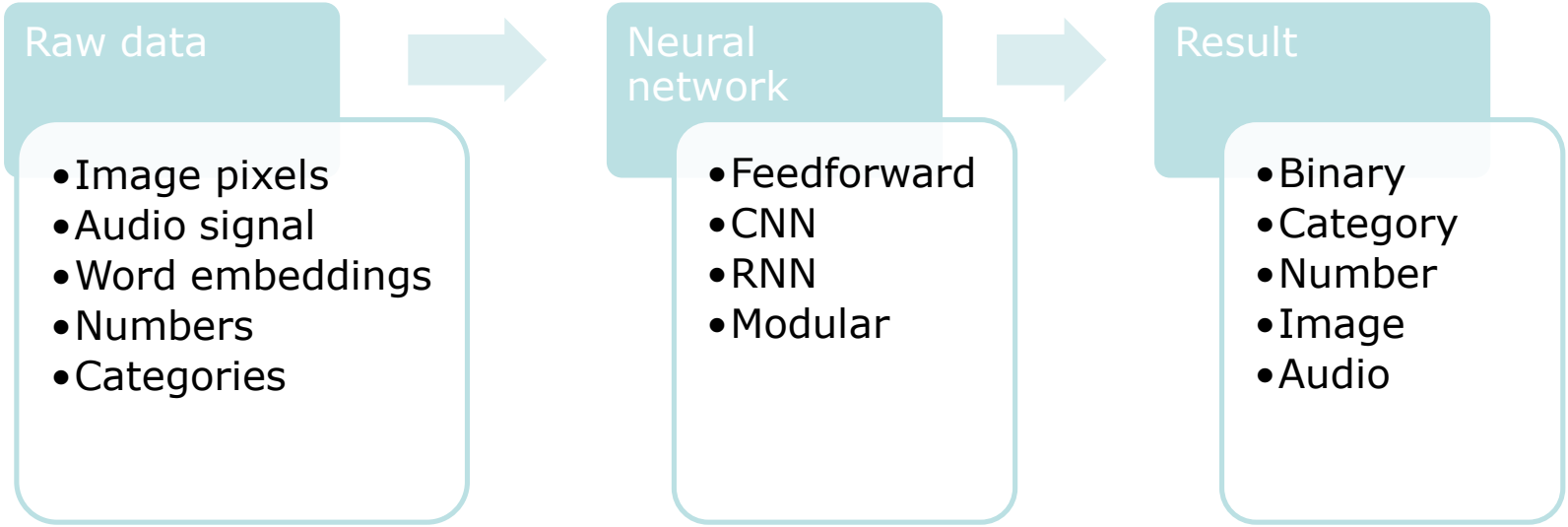
- Deep artificial neural networks (DNNs)
- Invented back in 1940s
- Became widely used in latest 10 years
- Computational power (GPU)
- Large amounts of data

Alexey Subach,
Technical Lead
Dual Lab

Let’s take a closer look…

Alexey Subach,
Technical Lead
Dual Lab

Pipeline examples



Alexey Subach,
Technical Lead
Dual Lab

Neural network architecture

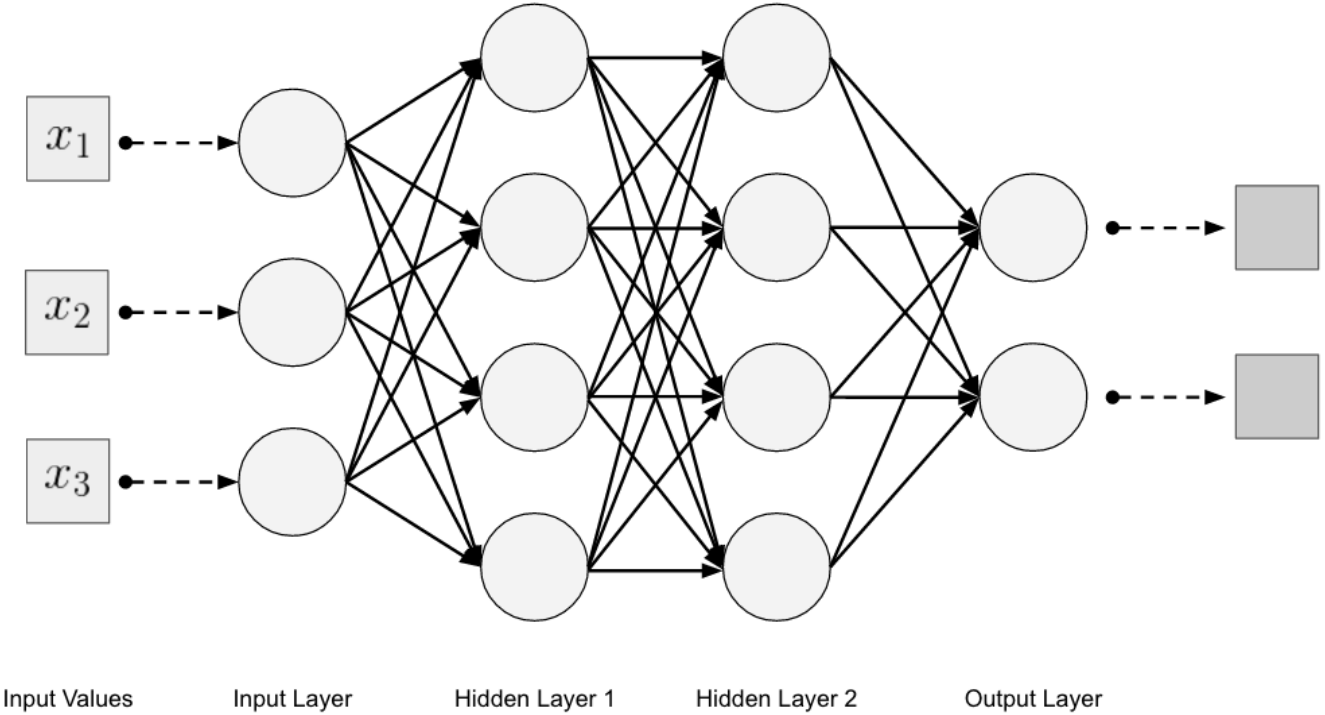
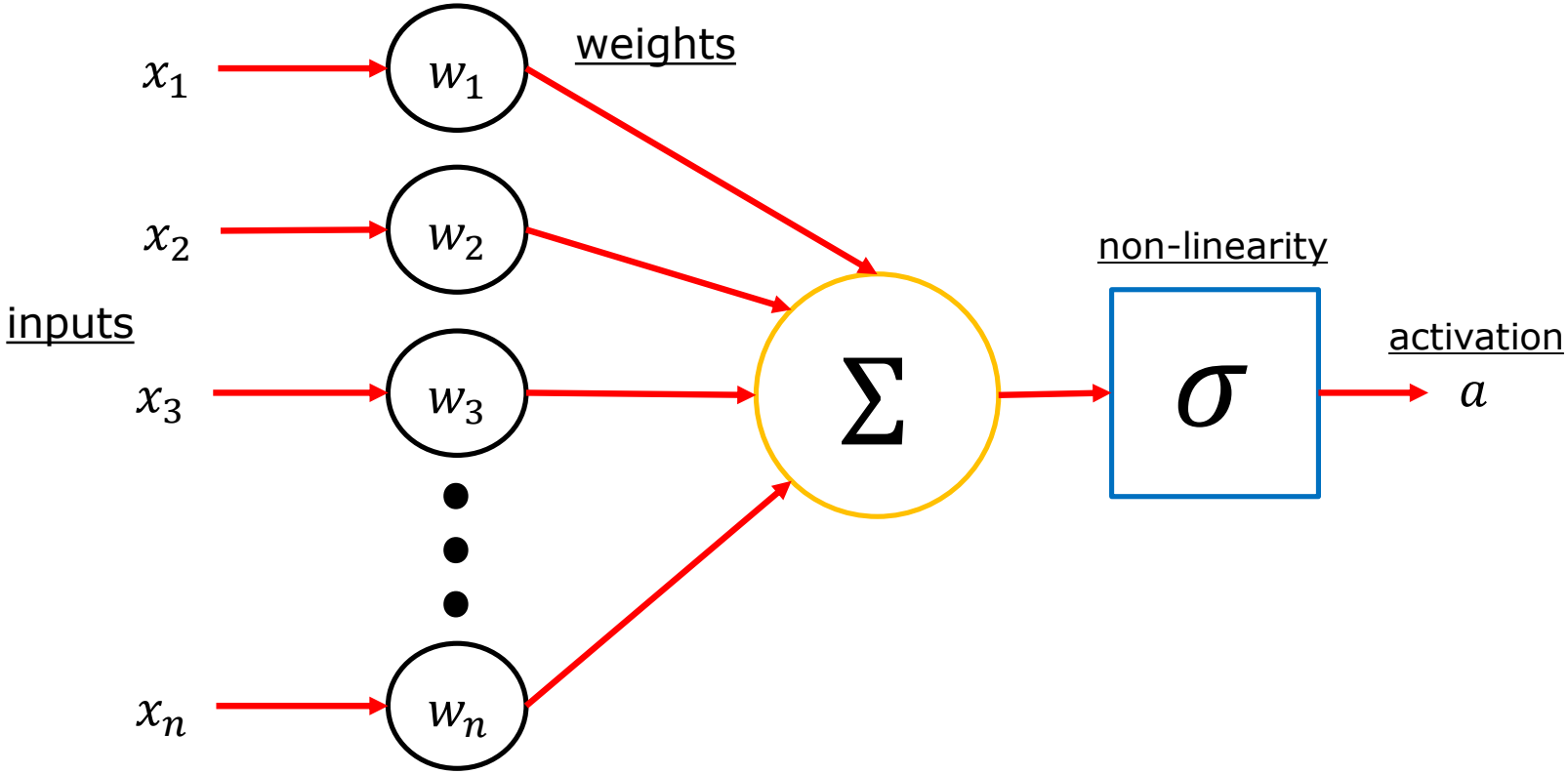


Image credit: Deep Learning by Adam Gibson, Josh Patterson

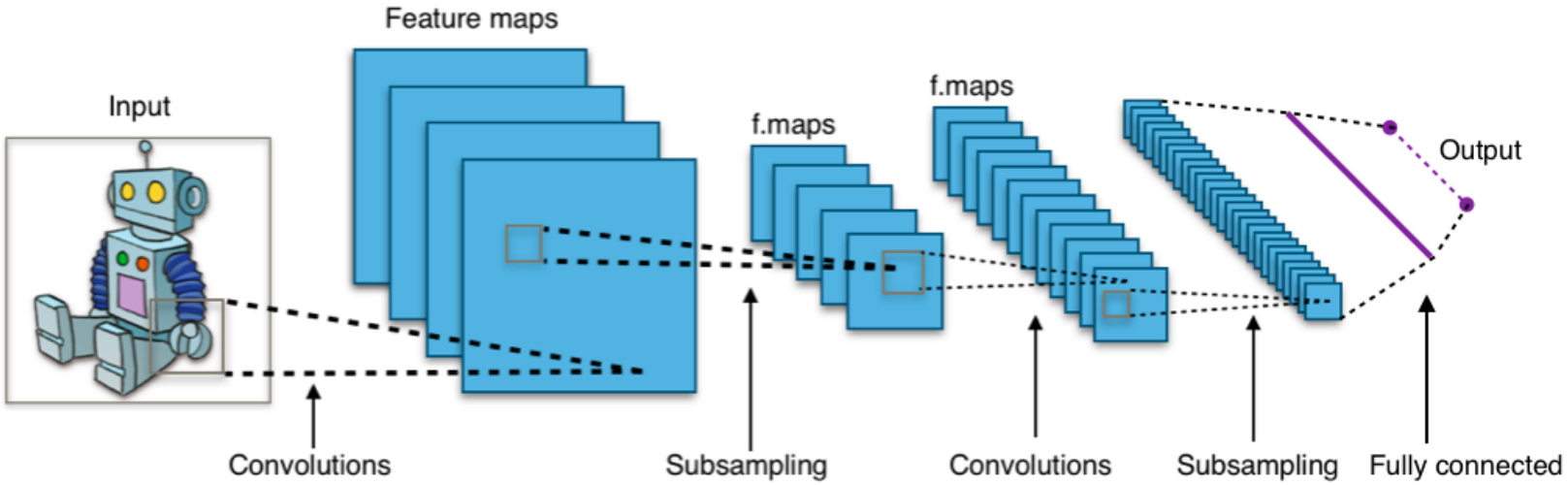
Alexey Subach,
Technical Lead
Dual Lab

Single neuron



Alexey Subach,
Technical Lead
Dual Lab

Deep convolutional neural networks



Alexey Subach,
Technical Lead
Dual Lab

Convolution

Source

22	15	1	3	60
42	5	38	39	7
28	9	4	66	79
0	2	25	12	17
9	14	2	51	3

Filter

0	0	1
0	0	0
1	0	0

Result

29	12	64
38	41	32
13	80	81

Alexey Subach,
Technical Lead
Dual Lab

Convolution

Source

22	15	**1**	3	60
42	5	38	39	7
28	9	4	66	79
0	2	25	12	17
9	14	2	51	3

Filter

0	0	1
0	0	0
1	0	0

Result

29		

Alexey Subach,
Technical Lead
Dual Lab

Convolution

Source

22	15	1	**3**	60
42	5	38	39	7
28	**9**	4	66	79
0	2	25	12	17
9	14	2	51	3

Filter

0	0	1
0	0	0
1	0	0

Result

29	**12**	

Alexey Subach,
Technical Lead
Dual Lab

Edge detection

Horizontal

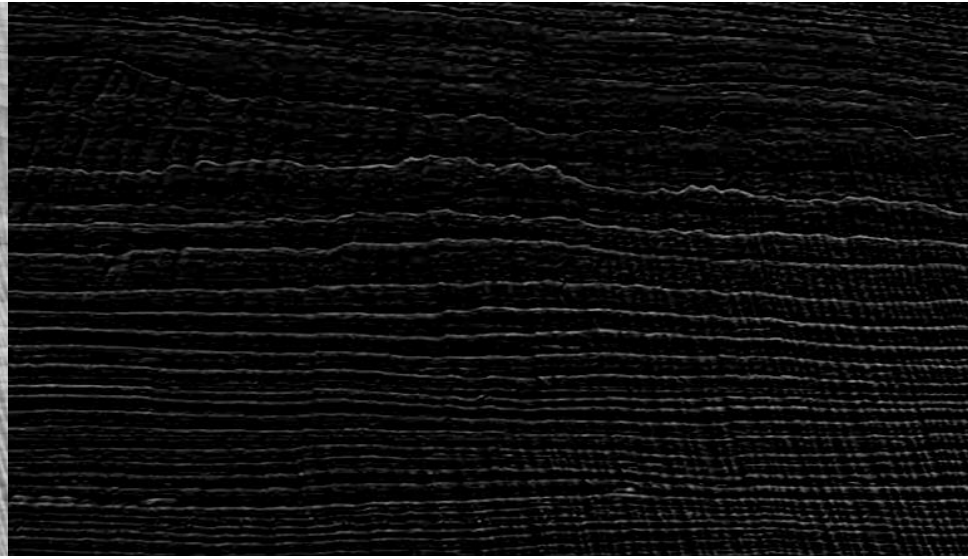
1	1	1
0	0	0
-1	-1	-1

Vertical

1	0	-1
1	0	-1
1	0	-1

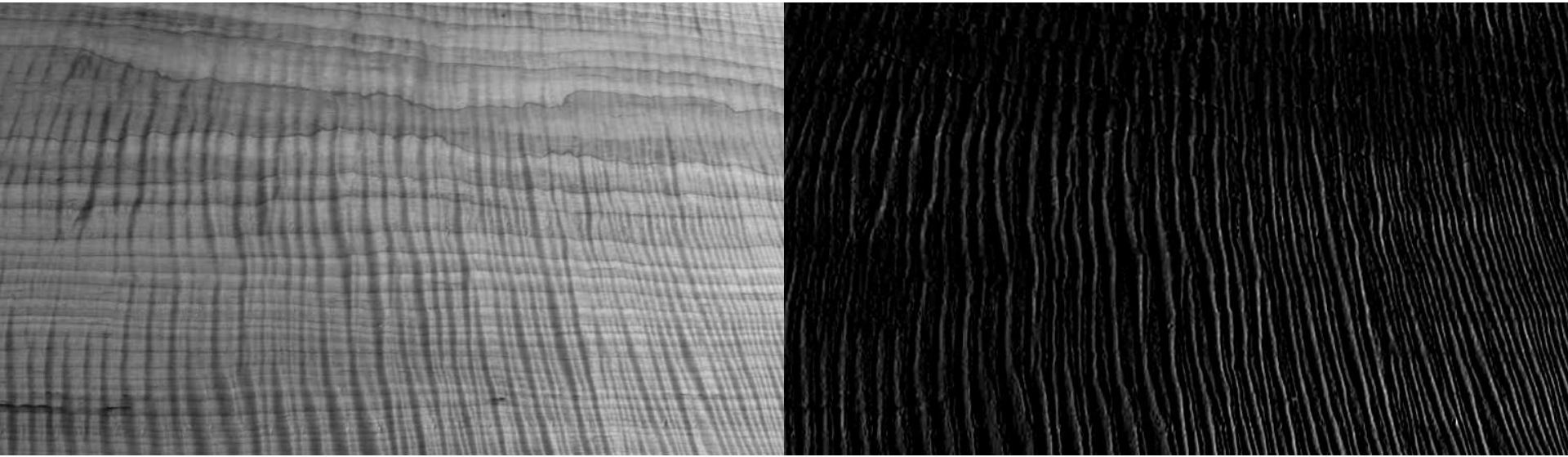
Alexey Subach,
Technical Lead
Dual Lab

Horizontal edges



Alexey Subach,
Technical Lead
Dual Lab

Vertical edges

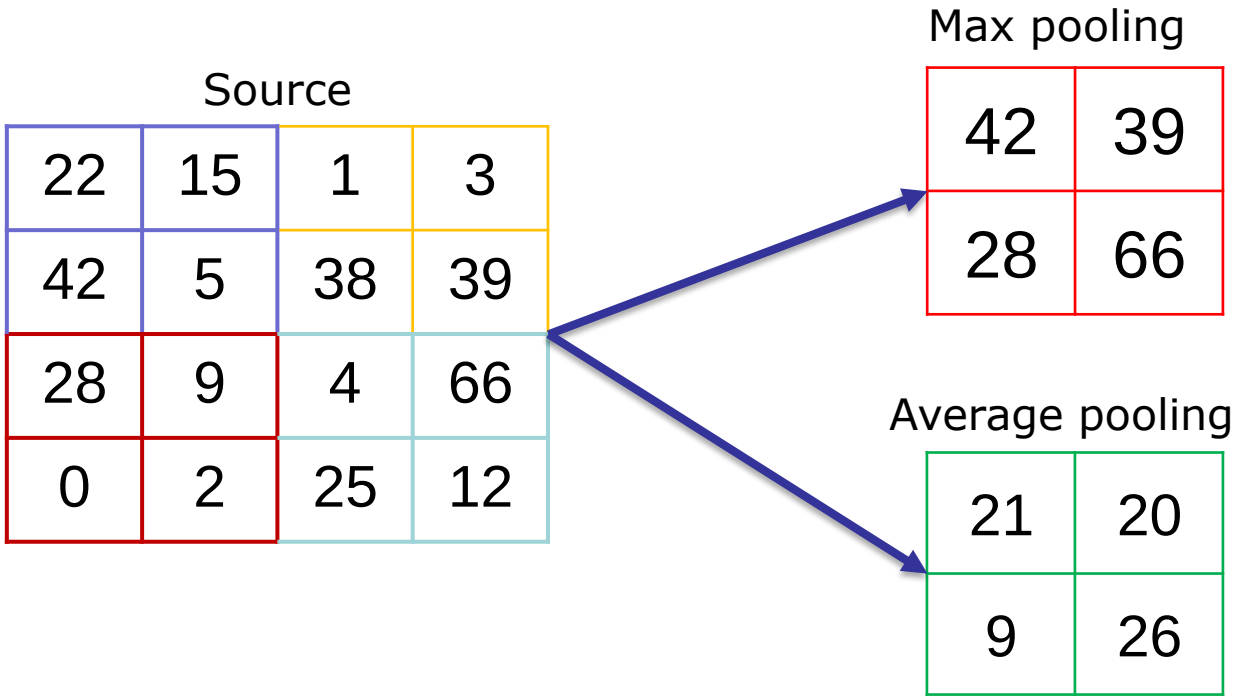


Alexey Subach,
Technical Lead
Dual Lab

One more example

Alexey Subach,
Technical Lead
Dual Lab

Pooling



Alexey Subach,
Technical Lead
Dual Lab

Parameters of the neural network - weights

Filter

0	0	1
0	0	0
1	0	0

Parameters

w_1	w_2	w_3
w_4	w_5	w_6
w_7	w_8	w_9

Alexey Subach,
Technical Lead
Dual Lab

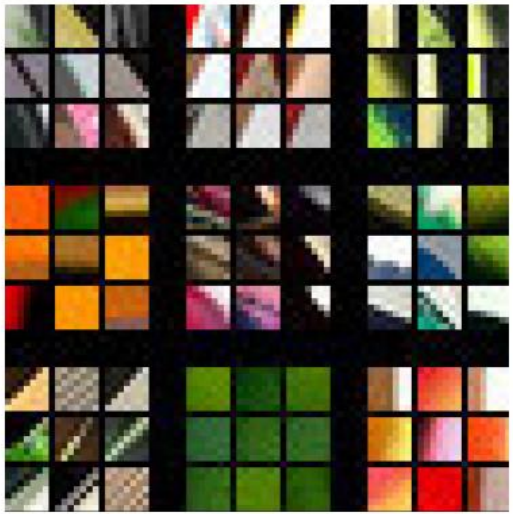
Training

- Last layer outputs the answer (probabilities, bbox etc)
- Loss function given the true answer for a training set item
- Loss optimization problem with respect to weights
- Gradient descent
- Training, validation, test sets
- Large problem requires huge amounts of data
- Trend: More data ↔ Bigger networks ↔ Better accuracy

Alexey Subach,
Technical Lead
Dual Lab

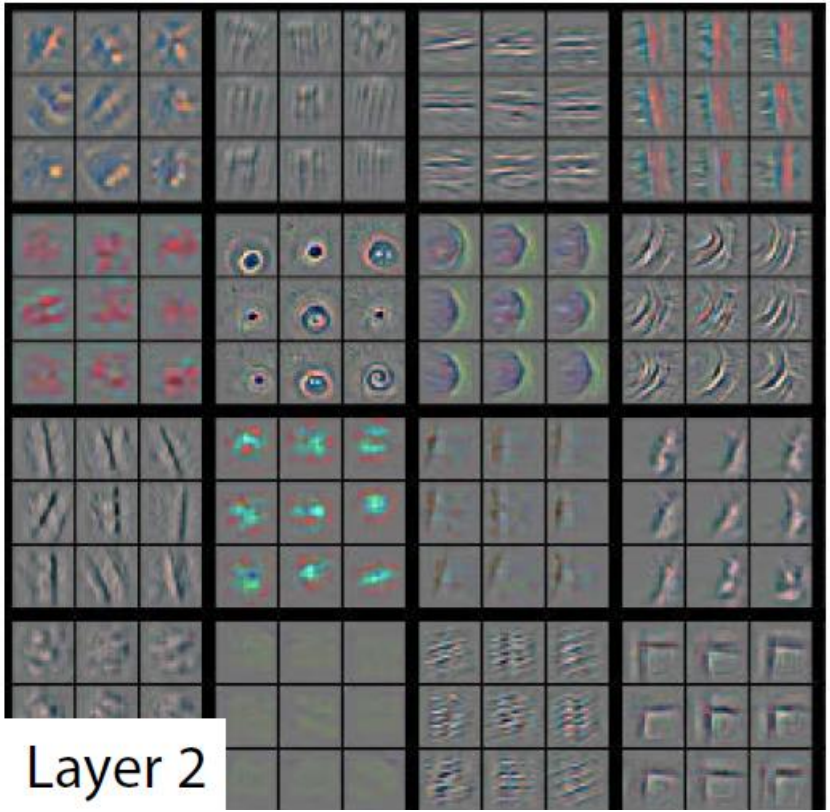
Understanding the impact of layer activations

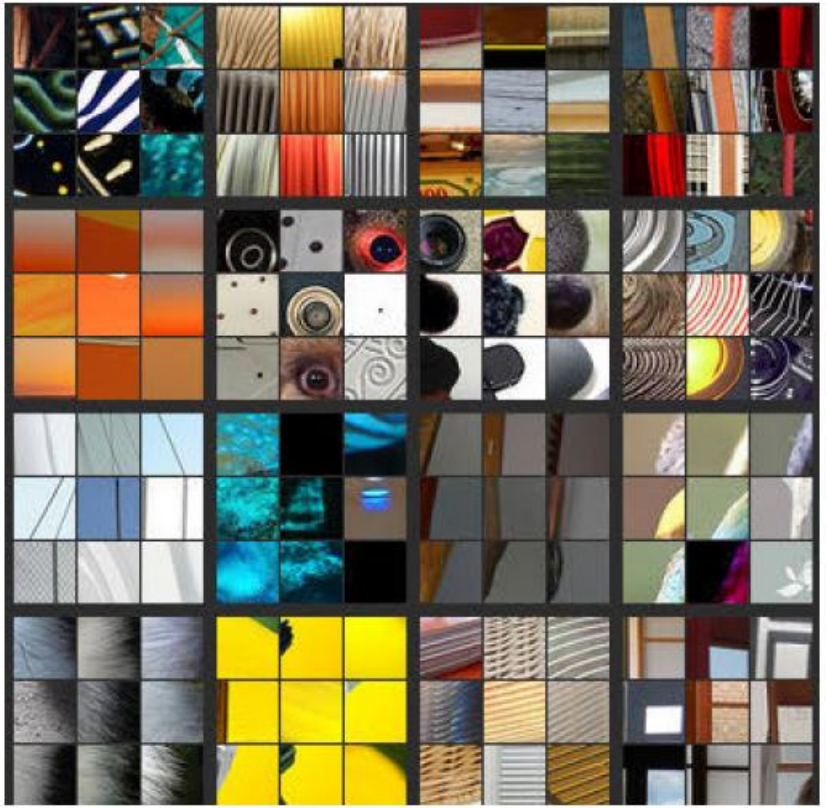
Layer 1



Alexey Subach,
Technical Lead
Dual Lab

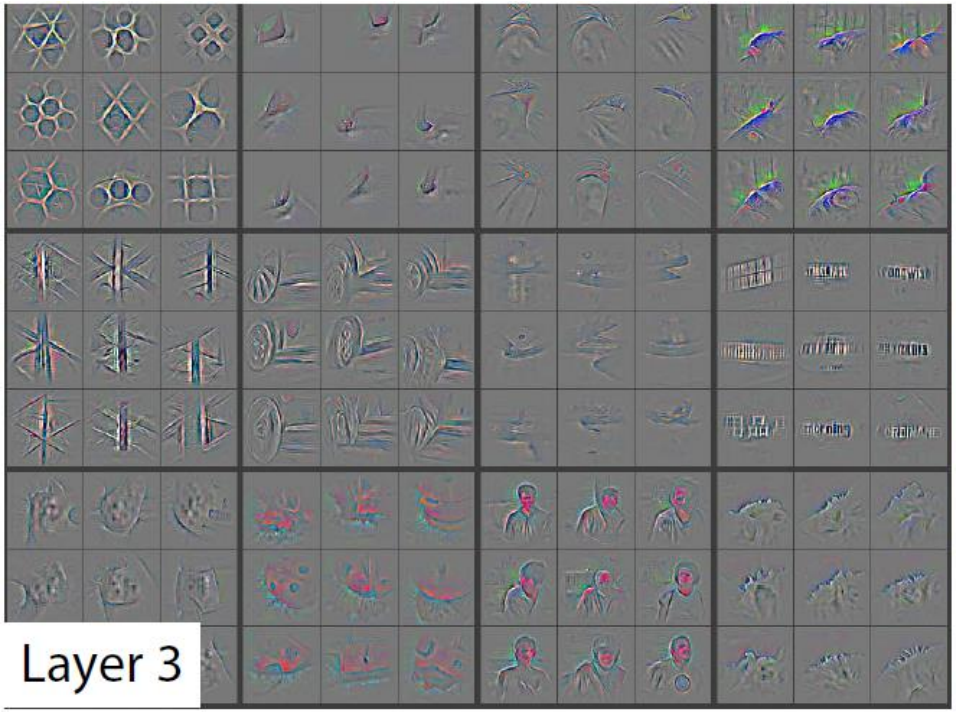
Understanding the impact of layer activations





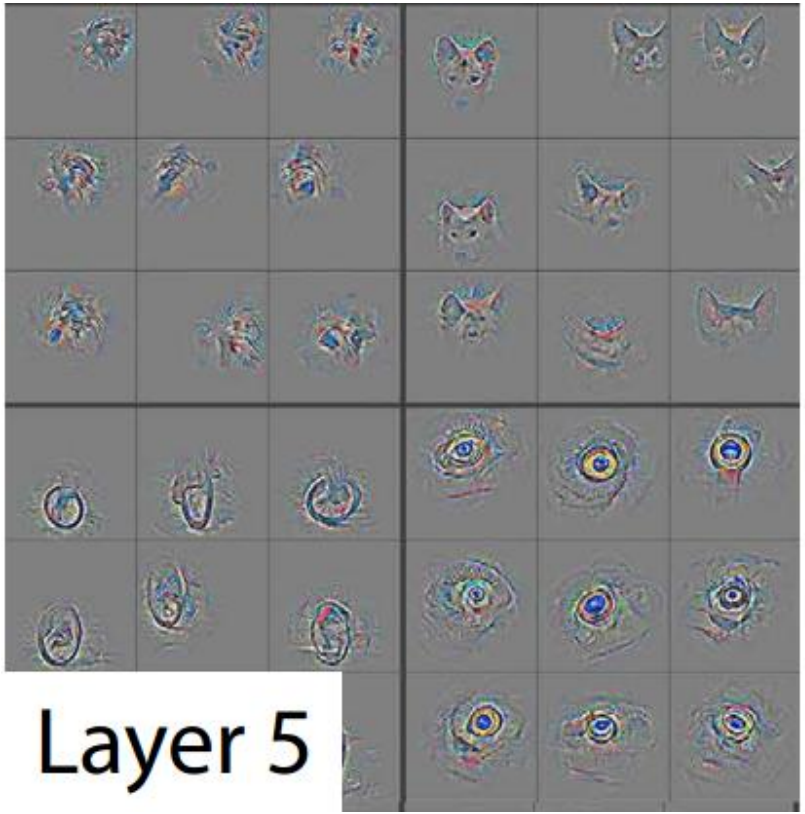
Alexey Subach,
Technical Lead
Dual Lab

Understanding the impact of layer activations



Alexey Subach,
Technical Lead
Dual Lab

Understanding the impact of layer activations



Alexey Subach,
Technical Lead
Dual Lab

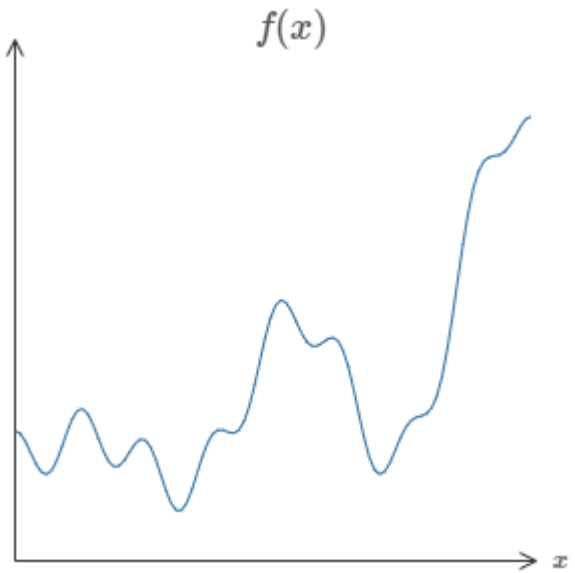
It was only a convolutional NN example

- There are many other types of NNs
- CNN on its own offers only a limited approach
- Recurrent NNs are used for NLP and keep track of the context (e.g. table spanning across several pages)
- What is going on in our brains when we look at a PDF page?
  - Do we remember and use previous pages?
  - Do we consider only visual structure, or read content as well?
  - Modular NNs
- Heuristics in algorithms → heuristics in NN architectures

Alexey Subach,
Technical Lead
Dual Lab

Why do they work so well?

- We don’t know, compared to classical ML algorithms
- Able to approximate any function on the **training set**
- Work well on the **test set** for **real world problems** and examples
- This is not cryptography, humans also make mistakes



Alexey Subach,
Technical Lead
Dual Lab

What about the applications?

Alexey Subach,
Technical Lead
Dual Lab

Structure recognition for untagged PDF

- Accessibility
  - Recognizing headings, tables, paragraphs, …
  - Reading order (two column layout or table columns)

Alexey Subach,
Technical Lead
Dual Lab

Structure recognition for untagged PDF

MULIGHED FOR ET BESKÆFTIGELSESRETTET TILBUD

Udlændinge, der ikke modtager offentlige ydelser, kan også få et beskæftigelsesrettet tilbud. Det er en mulighed for at opnå vigtig erfaring og et ståsted på arbejdsmarkedet.

Ønsker du et beskæftigelsesrettet tilbud, skal du ringe til Integration og Sprog – International House Copenhagen (se kontaktoplysninger på side 8)

MERE INFORMATION

Der er mere information om danskuddannelse, intro-dansk og kursus i danske samfundsforhold og dansk kultur og historie på kommunens hjemmeside:

www.kk.dk/danskuddannelse

POSSIBILITY OF AN EMPLOYMENT ORIENTED OFFER

Foreigners who do not receive social benefits can also receive an employment oriented offer. This is a possibility to gain important experience and a stepping stone to the labour market.

If you would like an employment oriented offer, you can contact The Department of Integration and Language at International House Copenhagen.

FURTHER INFORMATION

There is more information about Danskuddannelse, Intro-dansk and the course in Danish societal conditions, culture and history on the City's website:

www.kk.dk/danskuddannelse

dual lab
Alexey Subach,
Technical Lead
Dual Lab

Structure recognition for untagged PDF

- Accessibility
  - Recognizing headings, tables, paragraphs, …
  - Reading order (two column layout or table columns)
- Repurposing
- Reflow after adding/removing content, **translation**
- Searching / indexing

Alexey Subach,
Technical Lead
Dual Lab

PDF/UA

- *“Content shall be marked in the structure tree with semantically appropriate tags in a logical reading order.”*
- Validating documents to conform PDF/UA standards
- 47/136 Matterhorn Protocol Checkpoints are currently marked as manual
- Can we make a system and mark them as machine, in the next version, maintaining the same level of accuracy?
- Matterhorn protocol: from machine and human to deterministic and non-deterministic

Alexey Subach,
Technical Lead
Dual Lab

Matterhorn protocol examples: Human judgment

- *List is an ordered list, but no value for the ListNumbering attribute is present*
- *Content is a mathematical expression but is not tagged with a Formula tag*
- *Content is tagged as a table for information that is not organized in rows and columns*
- *The structure type and attributes of a structure element are not **semantically** appropriate for the structure element*
- *Tags are not in logical reading order*

Alexey Subach,
Technical Lead
Dual Lab

Is it a table?

$T_c = 0$ (default)	Character
$T_c = 0.25$	Character

Alexey Subach,
Technical Lead
Dual Lab

Is it a table?

$T_c = 0$ (default)	Character
$T_c = 0.25$	C h a r a c t e r

Figure 56: Character spacing in horizontal writing

Alexey Subach,
Technical Lead
Dual Lab

Many more applications

- Categorizing incoming documents, e.g. invoices
- GDPR: Redaction of sensitive information (GAN)
- Generating alternate text (image captioning)
- Repairing broken documents
- OCR (already there)
- Conversion between formats
- Compression, optimization (autoencoders)

Alexey Subach,
Technical Lead
Dual Lab

How world class problems can be solved

- Kaggle data science competitions:
  - Prizes up to several million dollars
  - Real-world challenges by governments
- Might be hard for small businesses to afford the training process
  - Big datasets required
  - Weeks or hundreds of GPUs to train
- Transfer learning to reuse existing models
- Little time, a couple of GPUs, small teams and datasets, good results
- Competitions among big companies drive technology forward not only for them, but also for smaller entities who can use the best models in slightly different applications

Alexey Subach,
Technical Lead
Dual Lab

Software 2.0

- Classical stack – explicit instructions to the computer written by a programmer
- New software – programs written in neural network weights, programmers just define constraints on the behavior
- Large portion of problems share the property that it's easier to collect the data than to write the program
- Software 2.0 is not to replace classical programs, but will take over some areas
- Choice of using a 90% accurate model we understand, or 99% accurate model we don't

Andrej Karpathy
Director of AI at Tesla

Alexey Subach,
Technical Lead
Dual Lab

Why don’t we provide tools for building Software 2.0?

- We are vendors of tools to work with PDF in classical software
- As a community we should provide tools to build PDF-related software 2.0
- Data science community is open, and this drives it forward
- Being more open can attract more attention and help in solving challenges
- Transfer learning will help attracting more people to build Software 2.0 products on top of PDF technology

Alexey Subach,
Technical Lead
Dual Lab

We have to start from something..
How do we build a pipeline for PDF?

Alexey Subach,
Technical Lead
Dual Lab

Training a DNN

- Network architecture (from scratch or using transfer learning)
- **Features**
- **Training set**
- **Loss function**

Alexey Subach,
Technical Lead
Dual Lab

Which features to extract

- Rendered representation
- Glyphs and bboxes, lines and shapes
- Combination
- Special information (e.g. links)
- Make use of the semantics if present (and believed to be accurate)
- Sequence of raw content stream instructions

Alexey Subach,
Technical Lead
Dual Lab

Ground truth DB: Sources

Alexey Subach,
Technical Lead
Dual Lab

Ground truth DB: Sources

- Tagged PDF. Quality evaluation: discrepancy between number of cells in table rows; number of attributes in structure elements; heuristics or existing AI techniques

Alexey Subach,
Technical Lead
Dual Lab

Ground truth DB: Sources

- Tagged PDF. Quality evaluation: discrepancy between number of cells in table rows; number of attributes in structure elements; heuristics or existing AI techniques
- Structured documents: Word, HTML, LaTeX etc

Alexey Subach,
Technical Lead
Dual Lab

Ground truth DB: Sources

- Tagged PDF. Quality evaluation: discrepancy between number of cells in table rows; number of attributes in structure elements; heuristics or existing AI techniques
- Structured documents: Word, HTML, LaTeX etc
- Manual markup

Alexey Subach,
Technical Lead
Dual Lab

Ground truth DB: Sources

- Tagged PDF. Quality evaluation: discrepancy between number of cells in table rows; number of attributes in structure elements; heuristics or existing AI techniques
- Structured documents: Word, HTML, LaTeX etc
- Manual markup
- Semi-automatic markup

Alexey Subach,
Technical Lead
Dual Lab

Ground truth DB: Sources

- Tagged PDF. Quality evaluation: discrepancy between number of cells in table rows; number of attributes in structure elements; heuristics or existing AI techniques
- Structured documents: Word, HTML, LaTeX etc
- Manual markup
- Semi-automatic markup
- Crawling web-pages, using search engines API

Alexey Subach,
Technical Lead
Dual Lab

Ground truth DB: Sources

- Tagged PDF. Quality evaluation: discrepancy between number of cells in table rows; number of attributes in structure elements; heuristics or existing AI techniques
- Structured documents: Word, HTML, LaTeX etc
- Manual markup
- Semi-automatic markup
- Crawling web-pages, using search engines API
- Downloading existing document corpora

Alexey Subach,
Technical Lead
Dual Lab

Ground truth DB: Sources

- Tagged PDF. Quality evaluation: discrepancy between number of cells in table rows; number of attributes in structure elements; heuristics or existing AI techniques
- Structured documents: Word, HTML, LaTeX etc
- Manual markup
- Semi-automatic markup
- Crawling web-pages, using search engines API
- Downloading existing document corpora
- Data augmentation: skew, crop, resize, rotate

Alexey Subach,
Technical Lead
Dual Lab

Tagged PDF: calculating loss function

- Element-wise: object detection and **IOU** (intersection over union). Good for simpler tasks
- Complex tasks → tree-like structure
- Normalize document structure: reduce nesting, use tag subset etc
- Custom tree comparison algorithms
- String representation
  - BLEU (bilingual evaluation understudy)
  - Kernel method
  - Embedding calculation

Alexey Subach,
Technical Lead
Dual Lab

Ground truth DB: Community

- Expert opinion system to reach consensus
- Incentive to participate in this? You get a vote
- Quality measurement system for implementations
- Divided responsibility to store large amounts files
- Penalty for file unavailability
- Might be very open, might be invitation-only

Alexey Subach,
Technical Lead
Dual Lab

With huge benefits come great challenges..

Alexey Subach,
Technical Lead
Dual Lab

Adversarial examples

- Black box system

Skiing	91%
Ski	89%
Piste	86%
Mountain Range	86%
Geological Phenomenon	85%
Glacial Landform	84%
Snow	82%
Winter Sport	78%
Ski Pole	75%

Dog	91%
Dog Like Mammal	87%
Snow	84%
Arctic	70%
Winter	67%
Ice	65%
Fun	60%
Freezing	60%
Glacial Landform	50%

Alexey Subach,
Technical Lead
Dual Lab

Adversarial examples

Alexey Subach,
Technical Lead
Dual Lab

No single answer

Alexey Subach,
Technical Lead
Dual Lab

Other challenges

- PDF specification is very rich
- Have only talked about a subset of PDF – simple text and path drawing instructions
- PDF comprises many other standards and formats.
- Features specifically for interaction with a human: 3D annotations. JavaScript
- Explainability of deep learning models is limited

Alexey Subach,
Technical Lead
Dual Lab

Measuring performance and acceptance testing

- 100% match or threshold of similarity level with ground truth to determine whether test passes
- Additional checks, e.g. no content is missing
- What is the reference acceptable performance?
  - Random people
  - Domain area specialists
  - Author (unrealistic 100% performance)
- Mistakes are inevitable. Voting, ensembling

Alexey Subach,
Technical Lead
Dual Lab

Summary

- It is just the beginning. AI is going to be everywhere
- Sufficient dataset is a vital component
- Openness of the data science community allows rapid progress
- We as PDF Association should provide tools for building Software 2.0
- PDF/UA will evolve and automatic validation is going to become a reality
- The area is pretty challenging and likely cannot be error-free. But we are replacing humans who are also imperfect
- Being open is challenging in a competitive market, but we should try!

Alexey Subach,
Technical Lead
Dual Lab

Thank you! Any questions?

Get in touch: alexey.subach@duallab.com
Web site: duallab.com

Alexey Subach,
Technical Lead
Dual Lab