CENL

Image-Language Models in the Library
The Case for CLIP

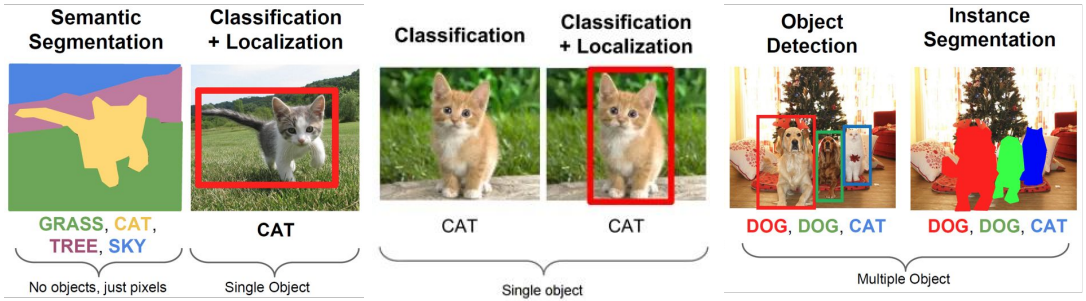
Javier de la Rosa
versae@nb.no
AI-lab
National Library of Norway

Jean-Philippe Moreux
jean-philippe.moreux@bnf.fr
BnF Bibliothèque nationale de France

Horace Lee
horace.lee@eng.ox.ac.uk
UNIVERSITY OF OXFORD

Computer Vision

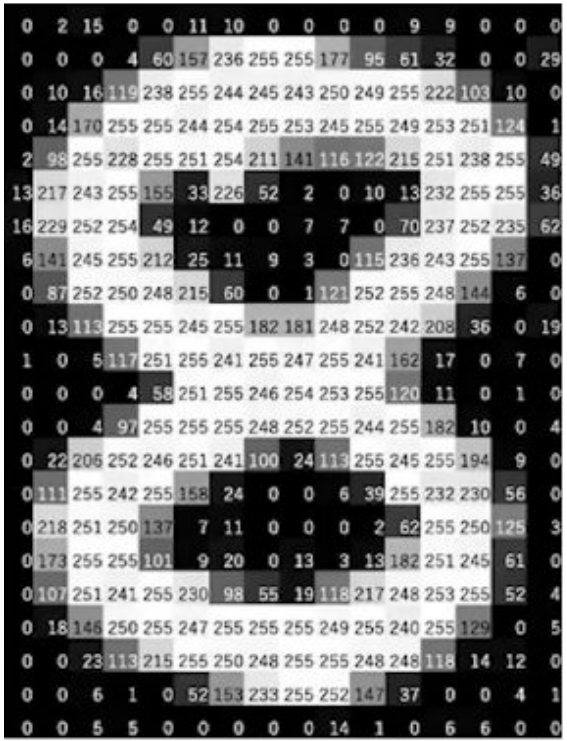
- Ability to "see" an image and understand the content.
- Trivial for a human being, even for small children
  - A person can describe the content of a photograph that he has seen once.
  - A person can summarize a video that he has only seen once.
  - A person can recognize a face that he has only seen once before.



Images are matrices

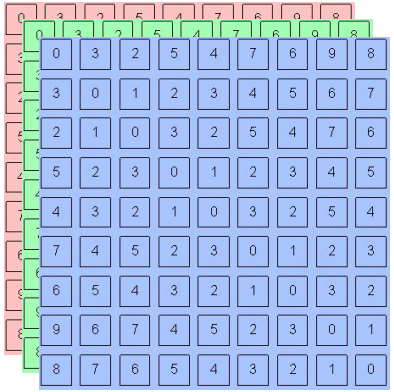
0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

Images are matrices



0	2	15	0	0	11	10	0	0	0	0	9	9	0	0	0
0	0	0	4	60	157	236	255	255	177	95	61	32	0	0	29
0	10	16	119	238	255	244	245	243	250	249	255	222	103	10	0
0	14	170	255	255	244	254	255	253	245	255	249	253	251	124	1
2	98	255	228	255	251	254	211	141	116	122	215	251	238	255	49
13	217	243	255	155	33	226	52	2	0	10	13	232	255	255	36
16	229	252	254	49	12	0	0	7	7	0	70	237	252	235	62
6	141	245	255	212	25	11	9	3	0	115	236	243	255	137	0
0	87	252	250	248	215	60	0	1	121	252	255	248	144	6	0
0	13	113	255	255	245	255	182	181	248	252	242	208	36	0	19
1	0	5	117	251	255	241	255	247	255	241	162	17	0	7	0
0	0	0	4	58	251	255	246	254	253	255	120	11	0	1	0
0	0	4	97	255	255	255	248	252	255	244	255	182	10	0	4
0	22	206	252	246	251	241	100	24	113	255	245	255	194	9	0
0	111	255	242	255	158	24	0	0	6	39	255	232	230	56	0
0	218	251	250	137	7	11	0	0	0	2	62	255	250	125	3
0	173	255	255	101	9	20	0	13	3	13	182	251	245	61	0
0	107	251	241	255	230	98	55	19	118	217	248	253	255	52	4
0	18	146	250	255	247	255	255	255	249	255	240	255	129	0	5
0	0	23	113	215	255	250	248	255	255	248	248	118	14	12	0
0	0	6	1	0	52	153	233	255	252	147	37	0	0	4	1
0	0	5	5	0	0	0	0	0	14	1	0	6	6	0	0

Color images are tensors

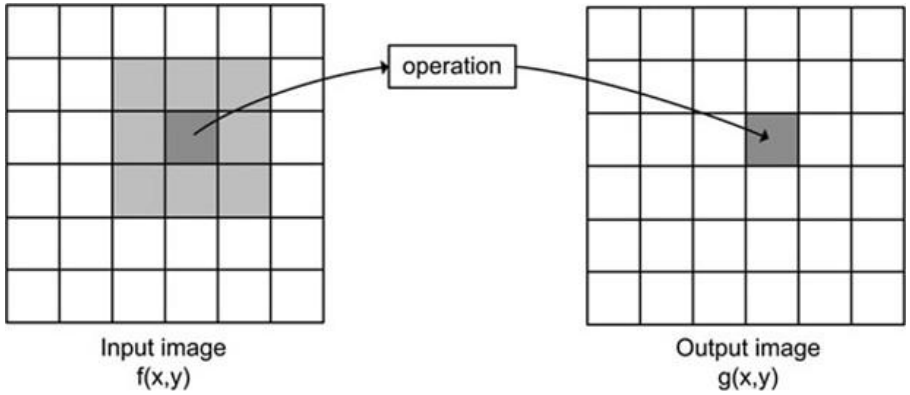


channel x height x width

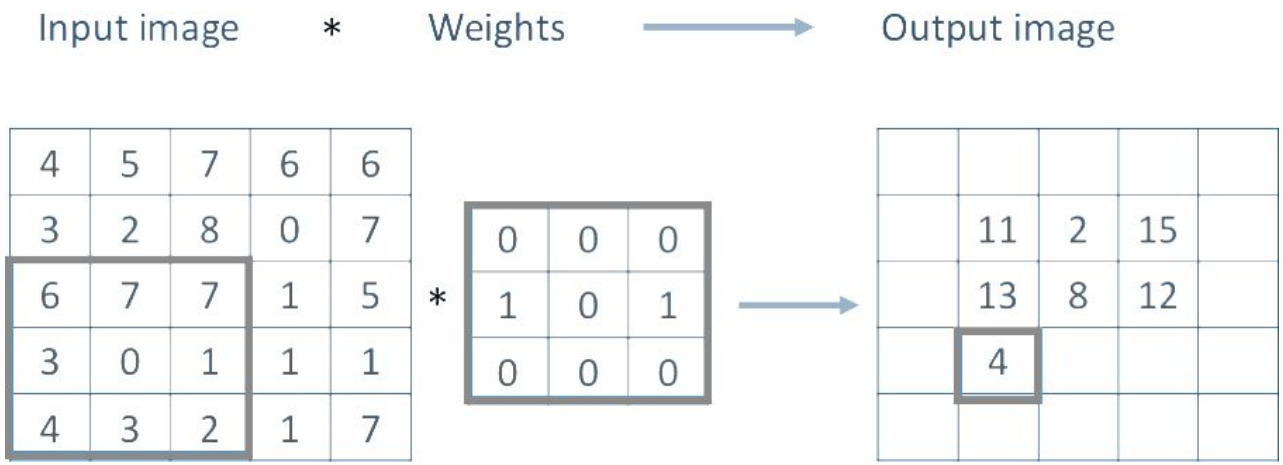
Channels are usually RGB: Red, Green, and Blue

Other color spaces: HSV, HSL, LUV, XYZ, Lab, CMYK, etc.

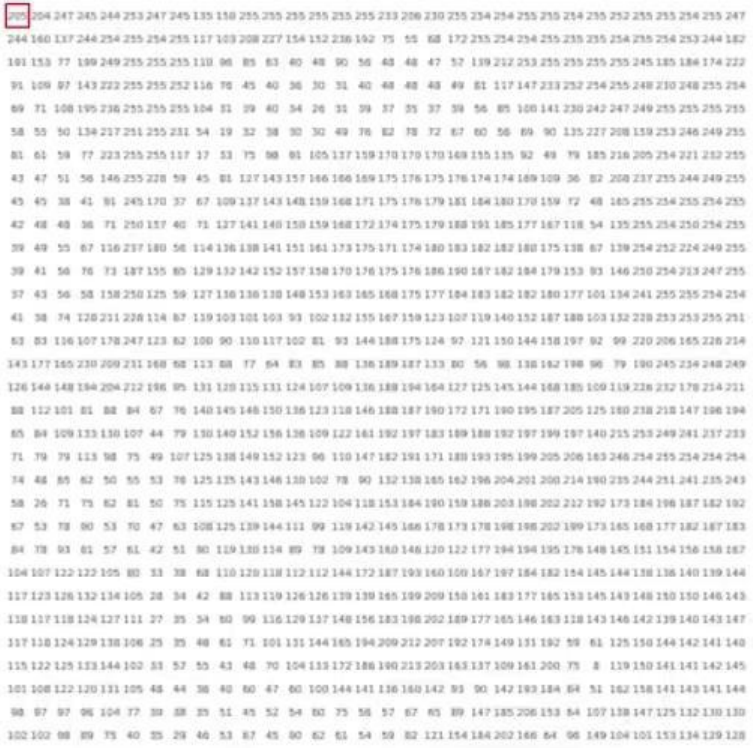
We can operate on them



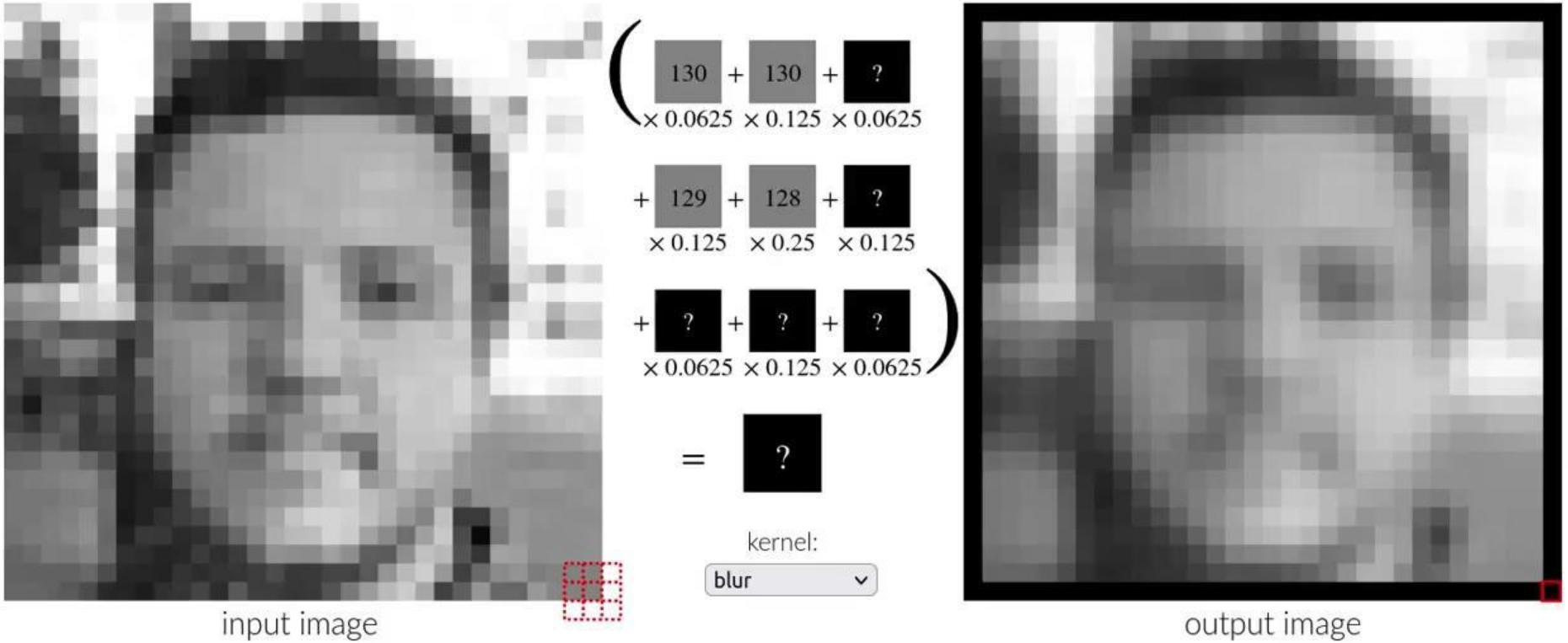
We can operate on them



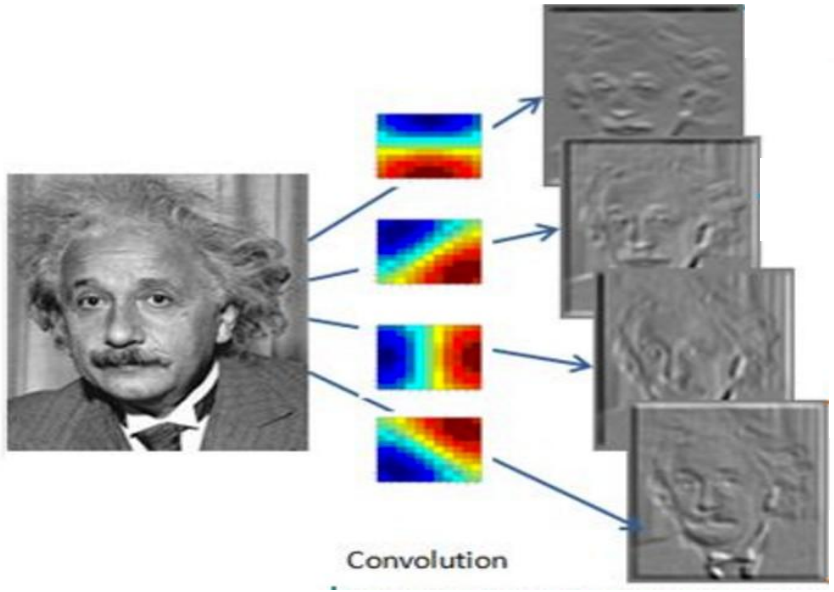
Convolutions



Convolutions



Convolutional Layer



AlexNet

ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

the paper that started the deep learning revolution!

Image classification

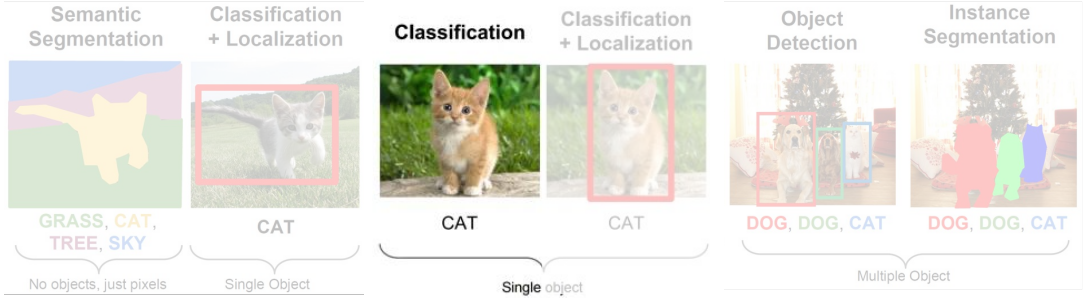


Image classification

Classify an image into **1000** possible classes:

Abyssinian cat, Bulldog, French Terrier, Cormorant, Chickadee, red fox, banjo, barbell, hourglass, knot, maze, viaduct, etc.

Trained on the ImageNet challenge dataset with ~1.2 million images

Image classification

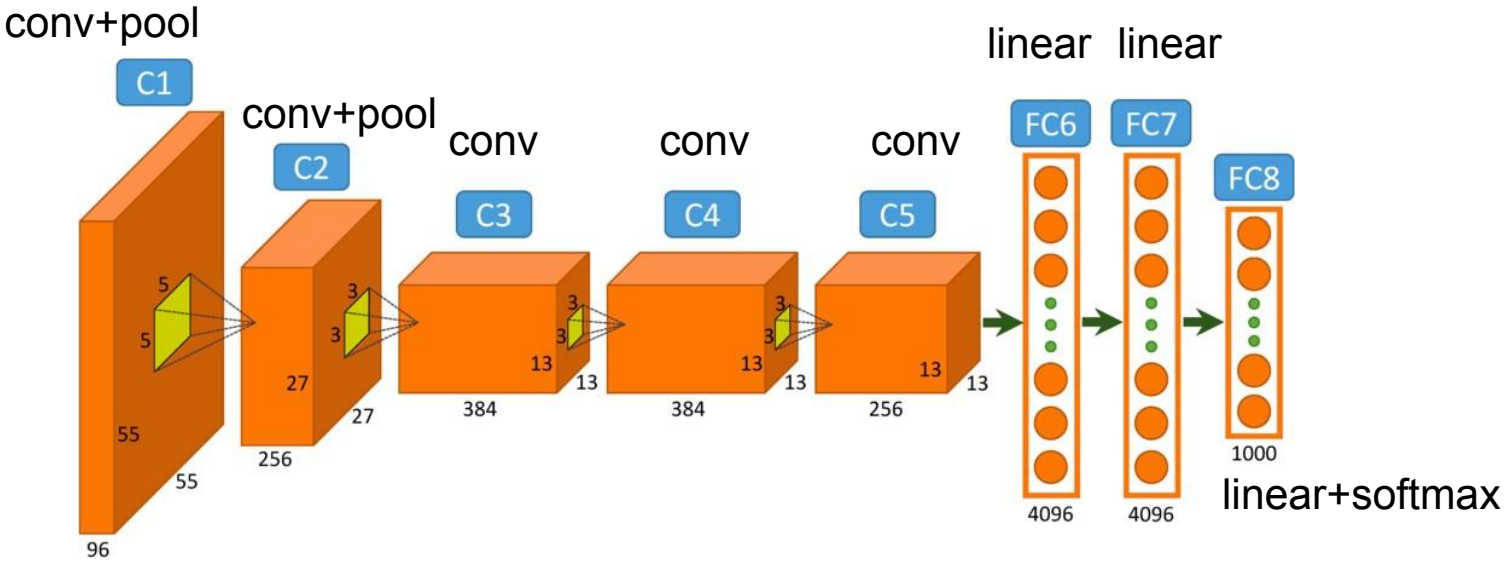
Classify an image into **1000** possible classes:

Abyssinian cat, Bulldog, French Terrier, Cormorant, Chickadee, red fox, banjo, barbell, hourglass, knot, maze, viaduct, etc.

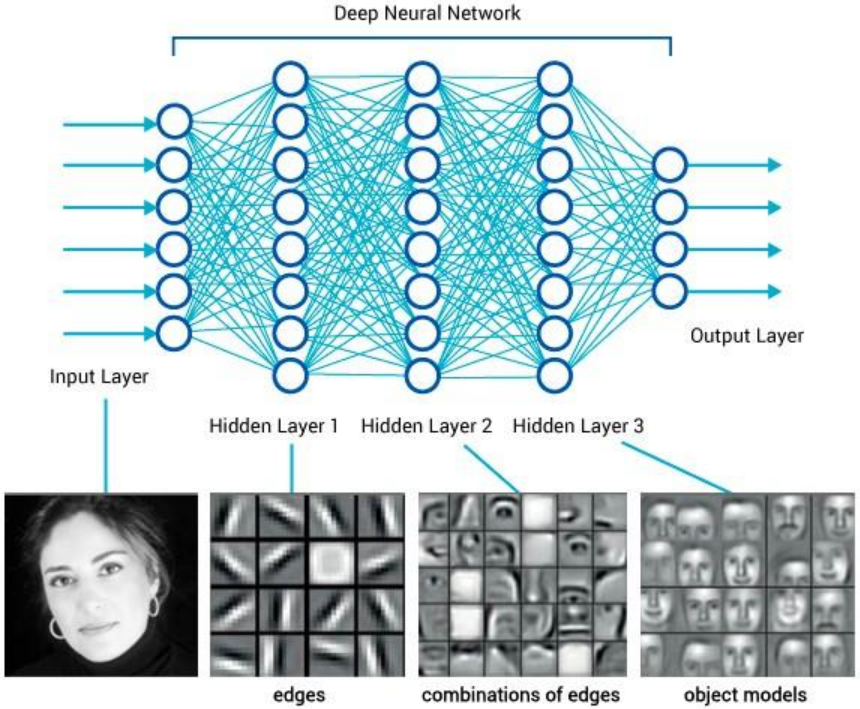
cat, tabby cat (0.71)
Egyptian cat (0.22)
red fox (0.11)

Trained on the ImageNet challenge dataset with ~1.2 million images

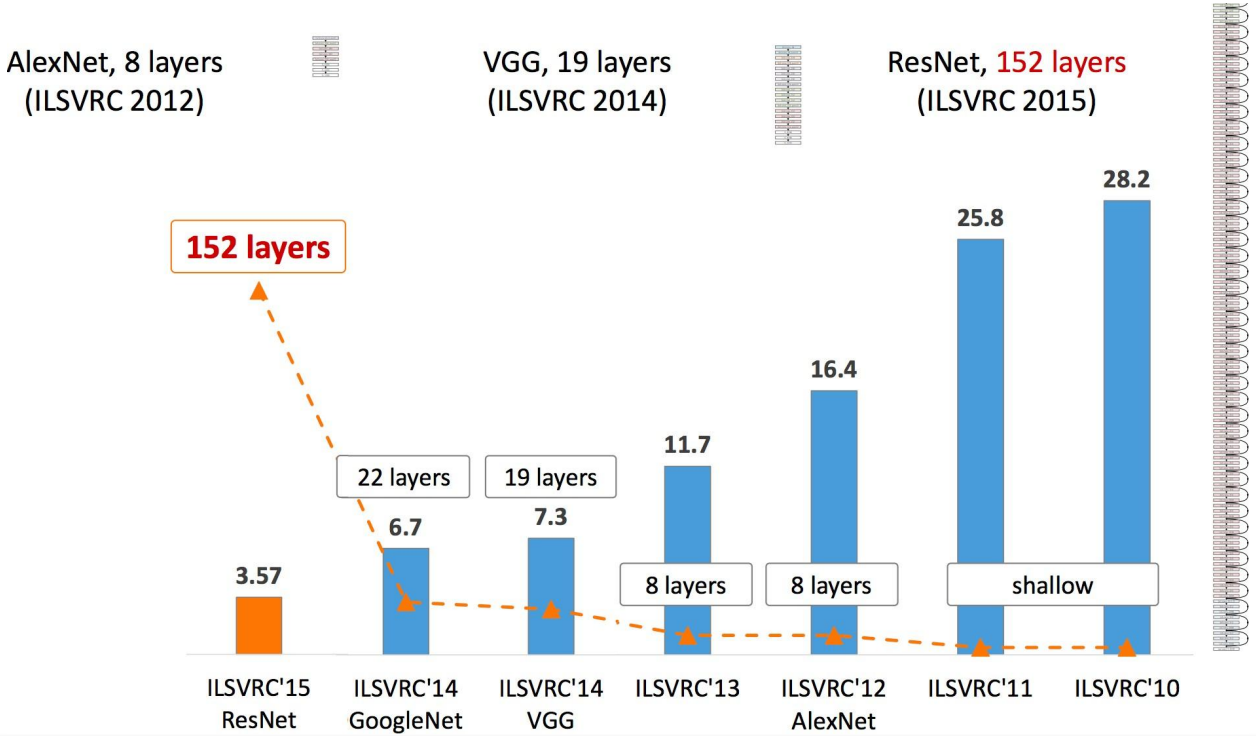
AlexNet


conv+pool
C1
conv+pool
C2
conv
C3
conv
C4
conv
C5
linear
linear
FC6
FC7
FC8
5
5
55
55
96
3
3
27
27
256
3
3
13
13
384
3
3
13
13
384
3
3
13
13
256
4096
4096
1000
linear+softmax

But what is happening?



Deeper networks → more layers → better performance



BERT

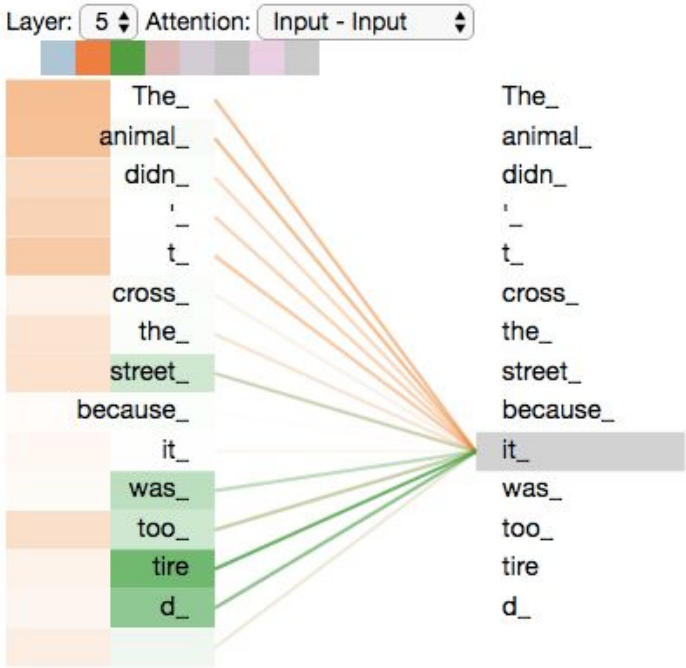
Pre-trained language models (Devlin et al., 2019)

- Transformer-based
- Masked Language Modeling and Next Sentence Prediction
- Self-attention!

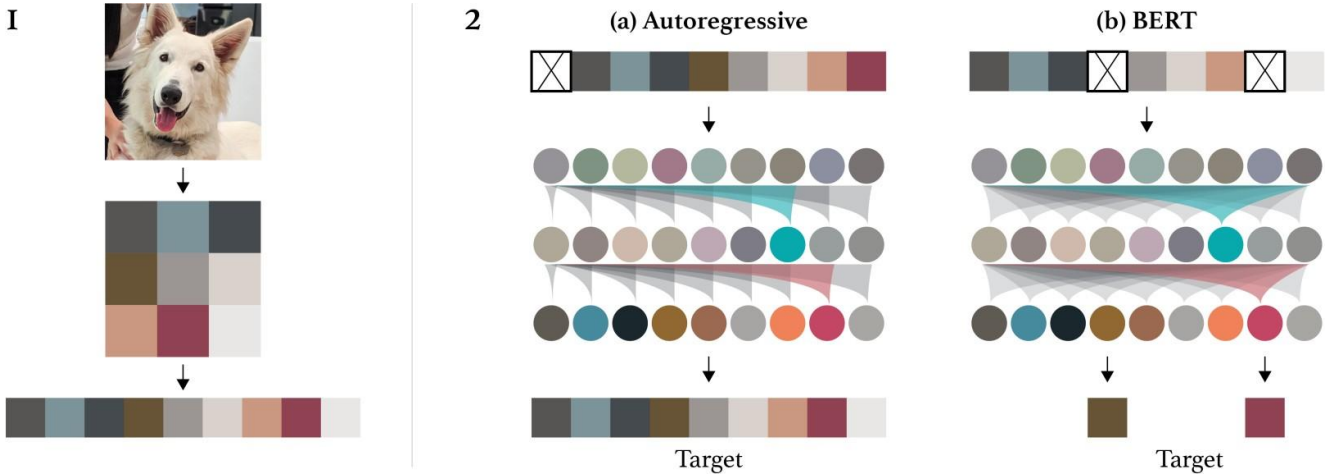
BERT

Pre-trained language models (Devlin et al., 2019)

- Transformer-based
- Masked Language Modeling and Next Sentence Prediction
- **Self-attention!**



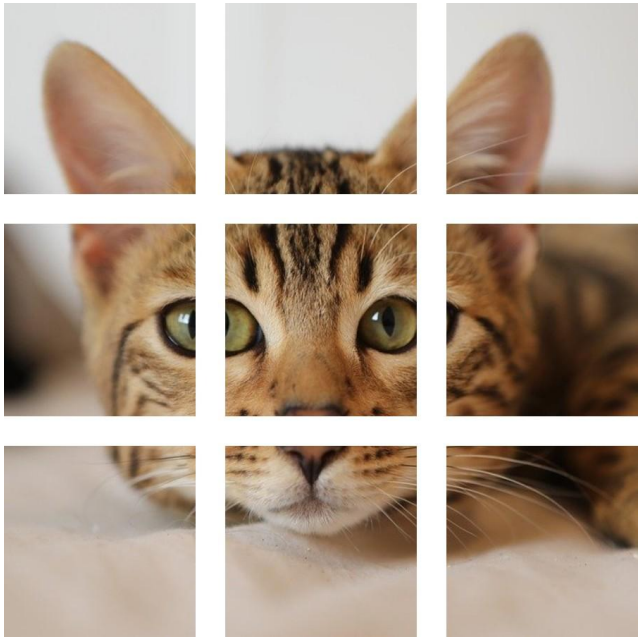
Self-attention on pixels



Self-attention on patches

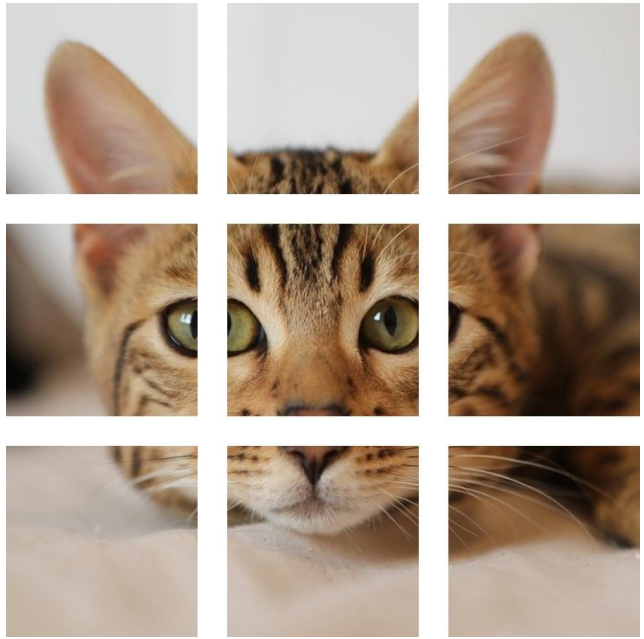
An Image is Worth 16x16 words, Dosovitskiy et al., ICLR 2021

Self-attention on patches

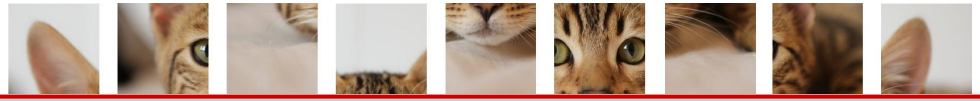


An Image is Worth 16x16 words, Dosovitskiy et al., ICLR 2021

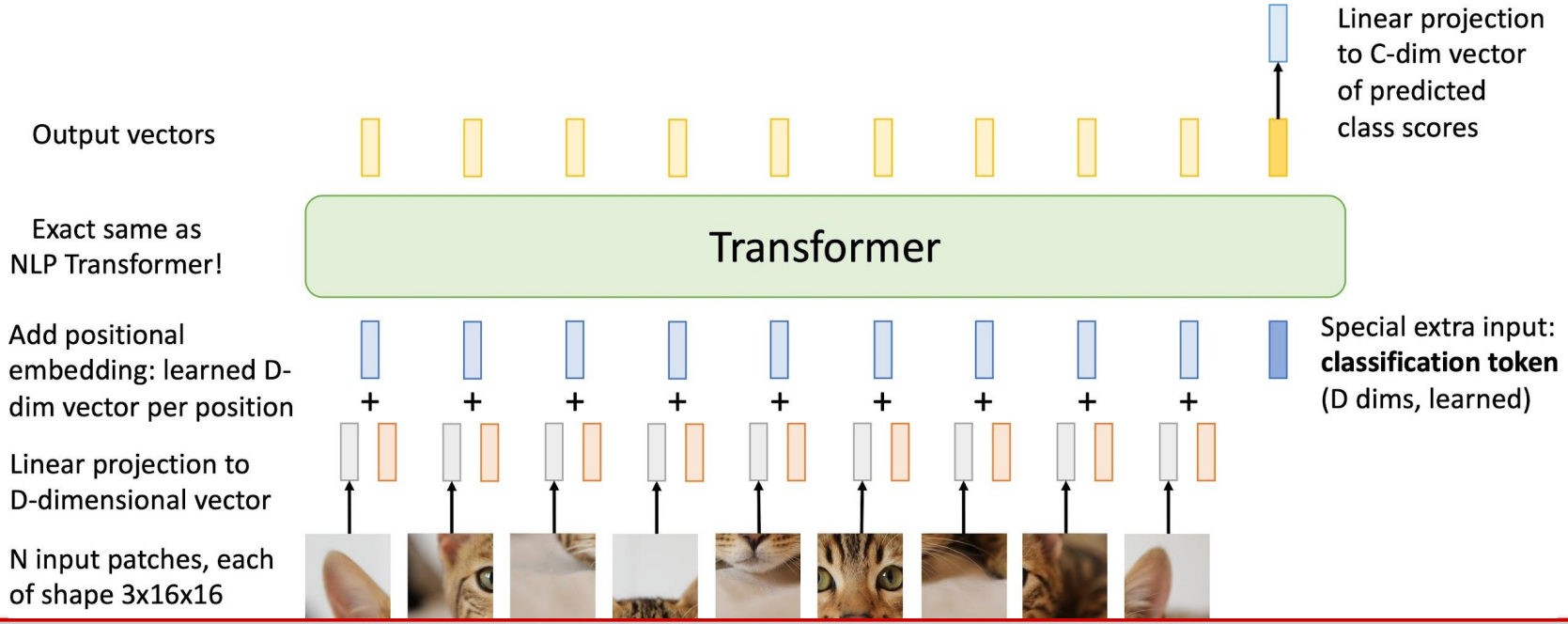
Self-attention on patches



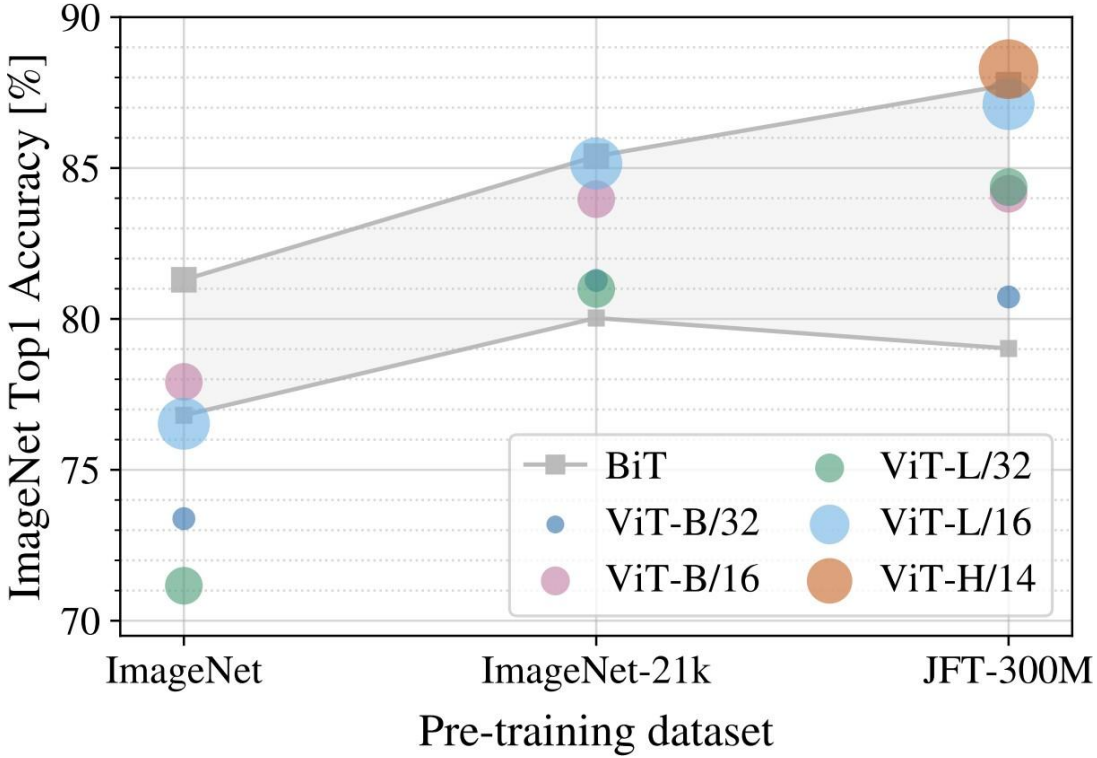
N input patches, each of shape 3x16x16



Self-attention on patches



Vision Transformers



Text and Image Pairs

a red truck is parked on a street lined with trees

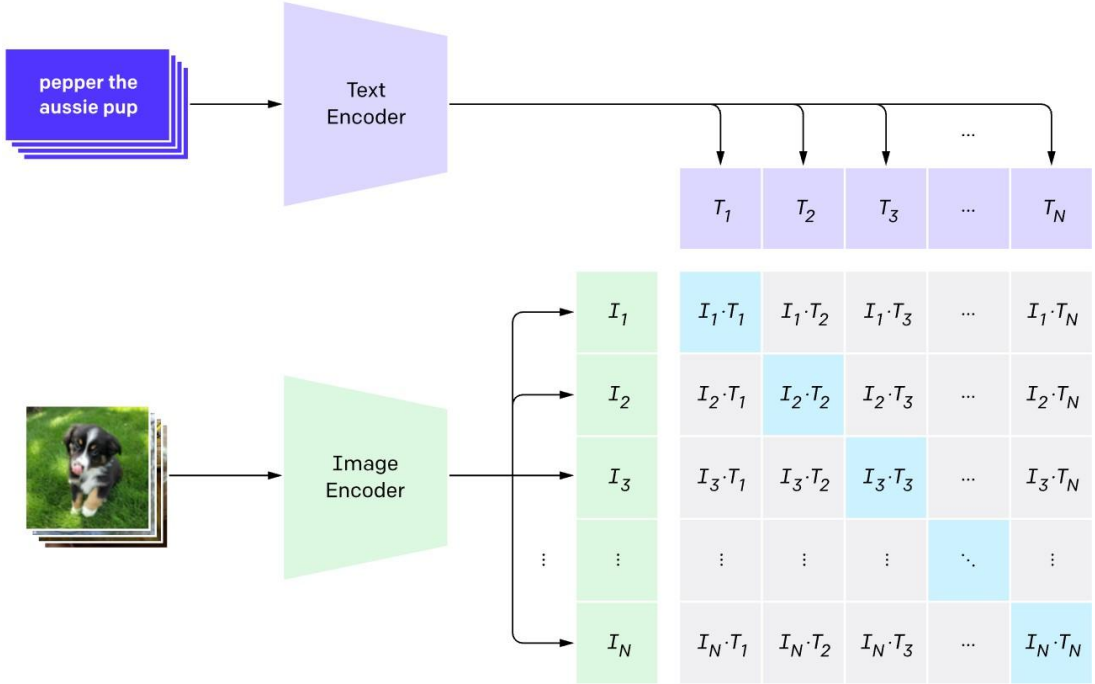
OpenAI's CLIP: Contrastive language-image pretraining

- OpenAI collect 400 million (image, text) pairs from the web

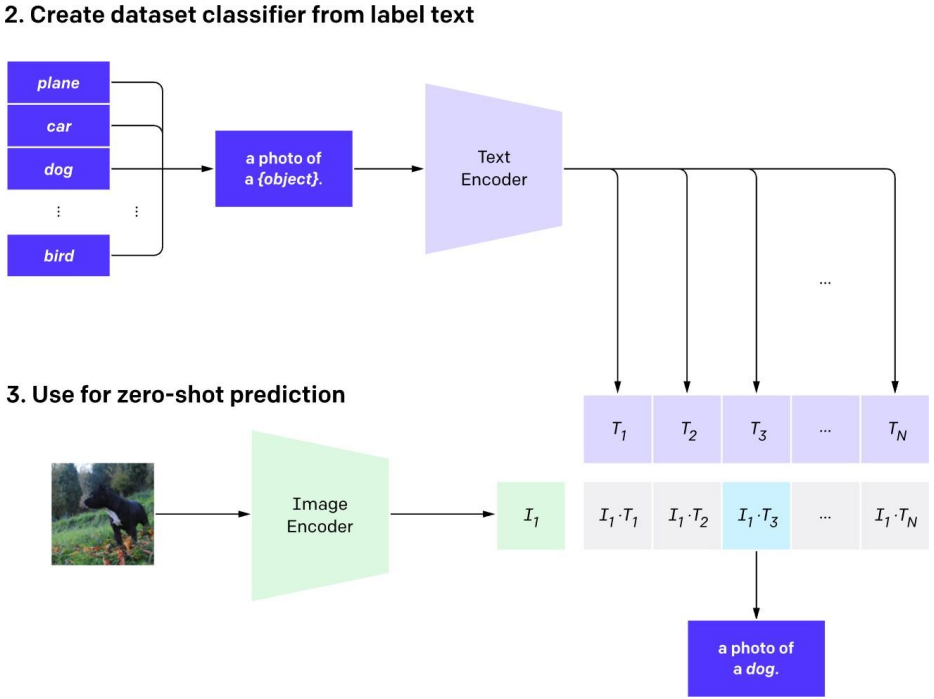
- Image encoder + text encoder with a simple contrastive loss: given a collection of images and text, predict which (image, text) pairs actually occurred in the dataset

OpenAI's CLIP: Contrastive language-image pretraining

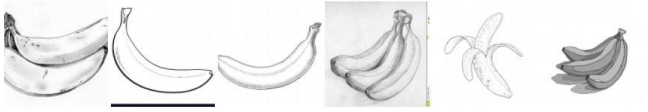
1. Contrastive pre-training



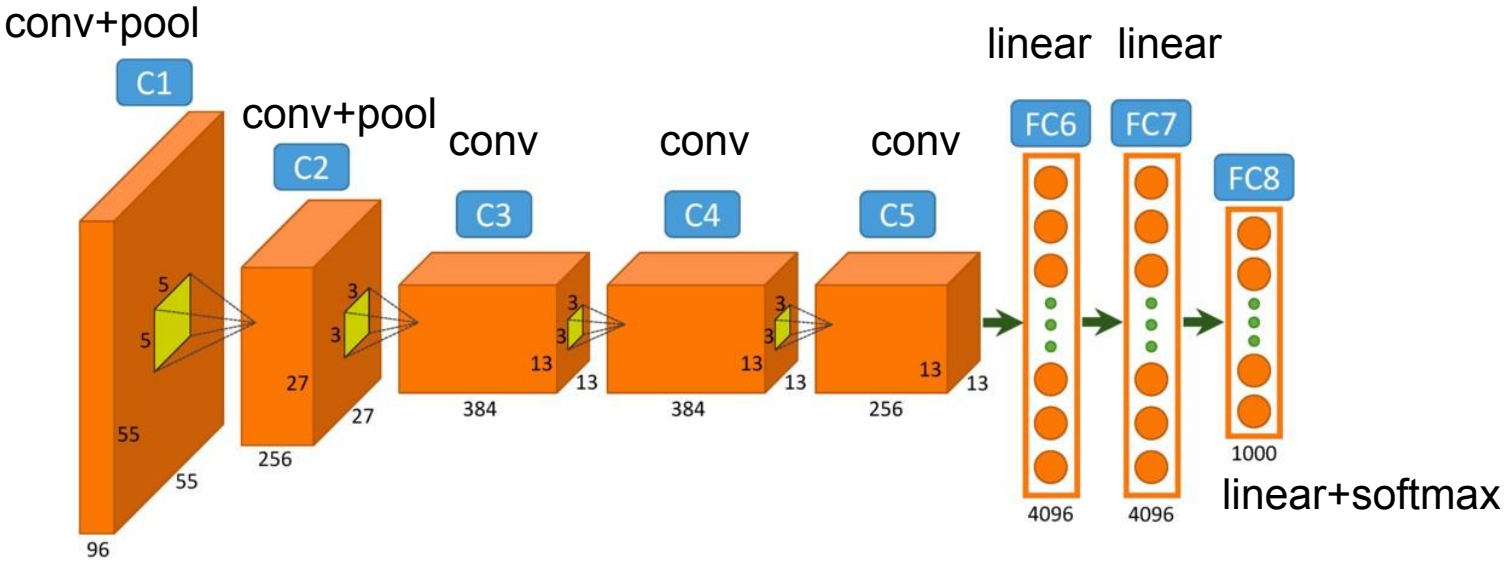
OpenAI's CLIP: Contrastive language-image pretraining



OpenAI's CLIP

DATASET	IMAGENET RESNET101	CLIP VIT-L
 ImageNet	76.2%	76.2%
 ImageNet V2	64.3%	70.1%
 ImageNet Rendition	37.7%	88.9%
 ObjectNet	32.6%	72.3%
 ImageNet Sketch	25.2%	60.2%
 ImageNet Adversarial	2.7%	77.1%

AlexNet


conv+pool
C1
conv+pool
C2
conv
C3
conv
C4
conv
C5
linear
linear
FC6
FC7
FC8
5
5
55
55
96
3
3
27
27
256
3
3
13
13
384
3
3
13
13
384
3
3
13
13
256
4096
4096
1000
linear+softmax

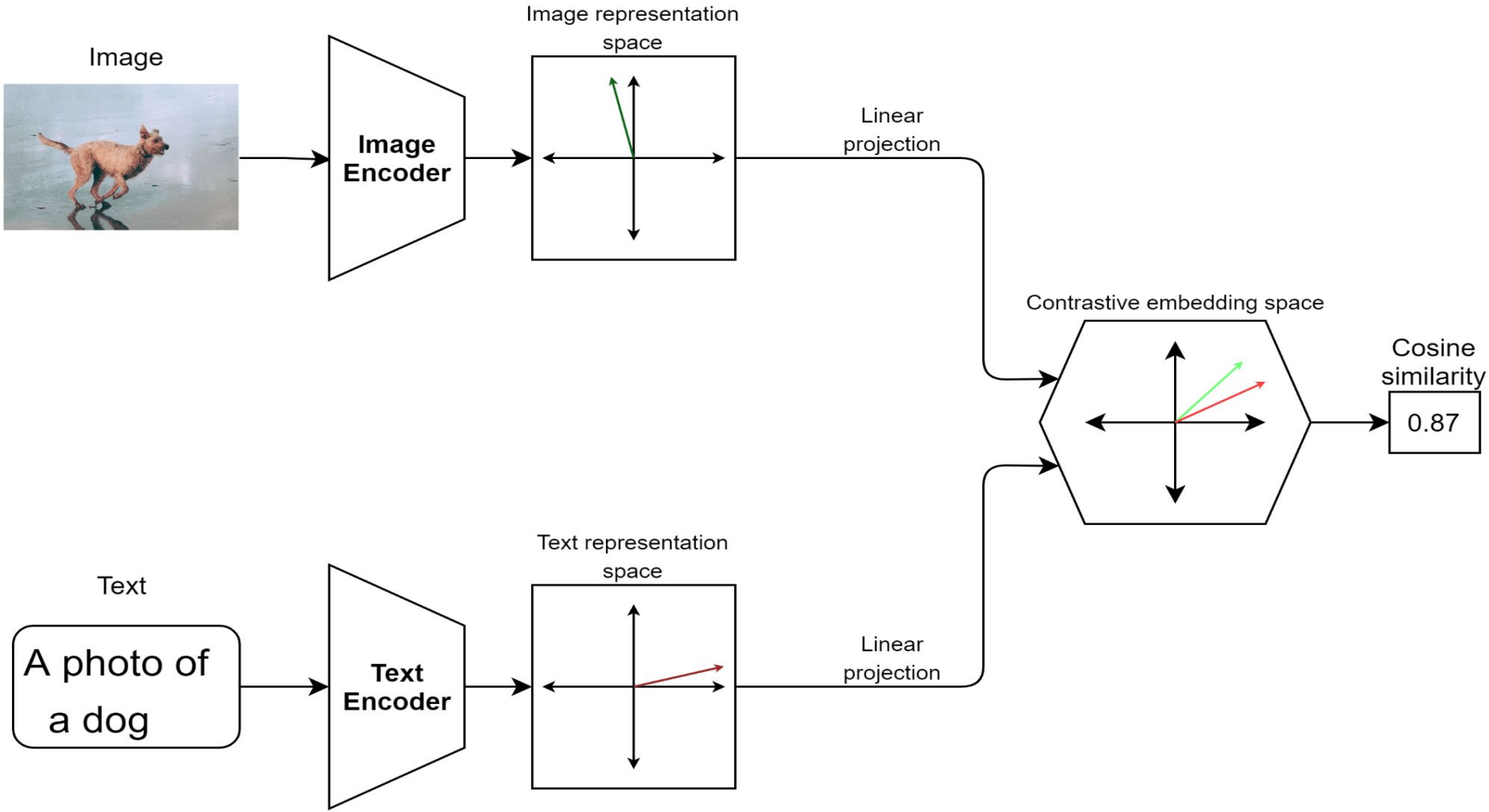
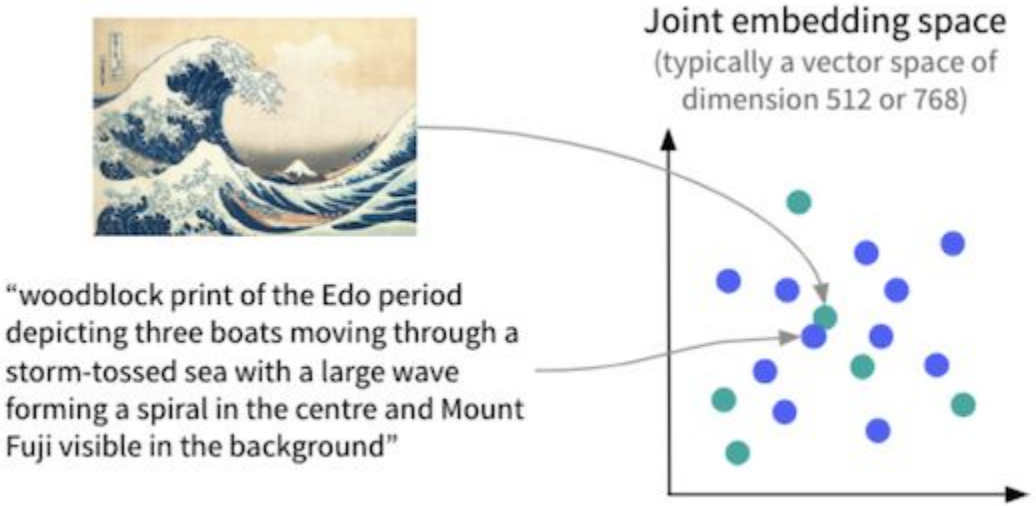


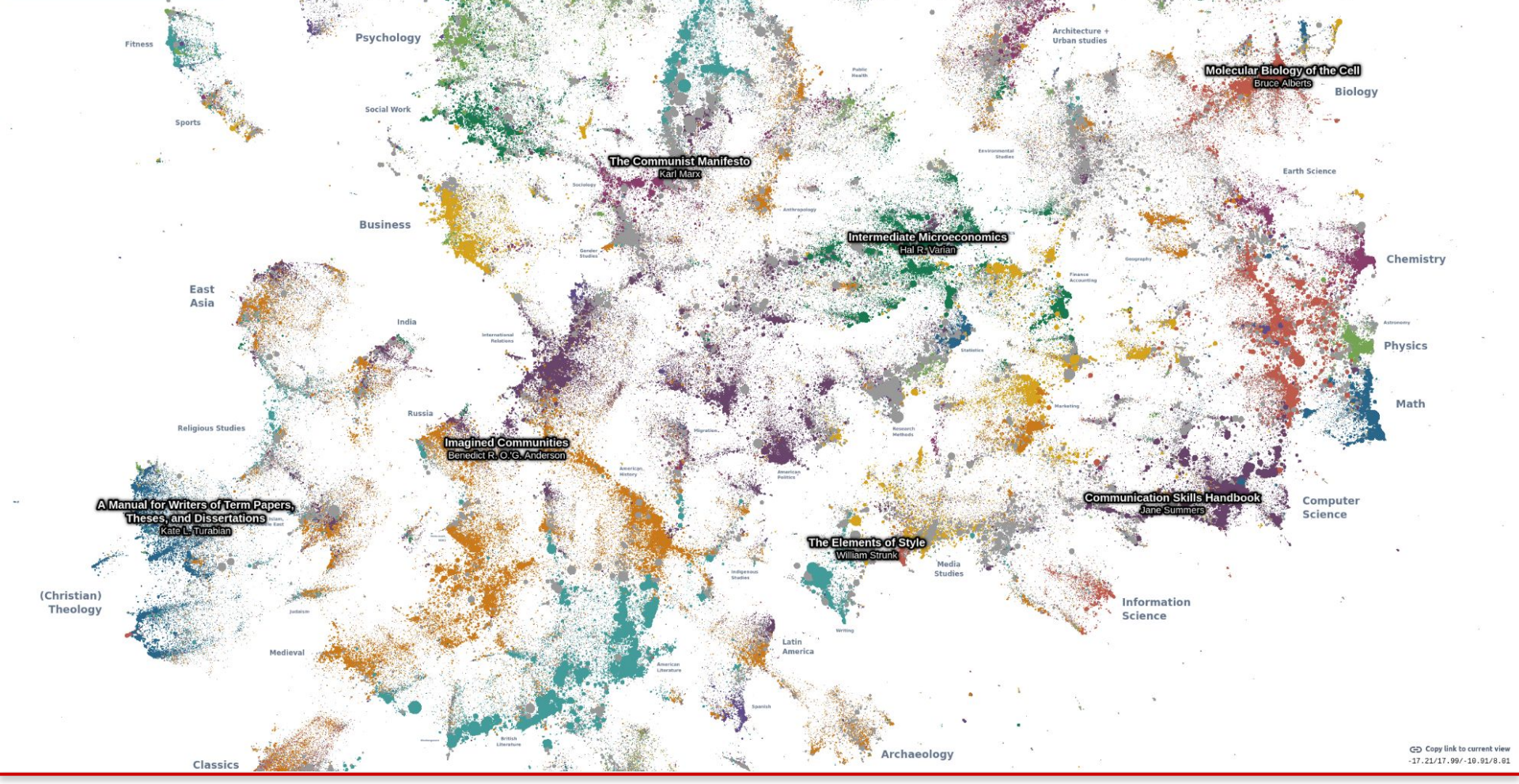
Image
Image Encoder
Image representation space
Linear projection
Contrastive embedding space
Cosine similarity
0.87
Text
A photo of a dog
Text Encoder
Text representation space
Linear projection

Embedding Space



kids playing in the snow

kids playing in the snow


Psychology
Fitness
Sports
Social Work
Architecture + Urban studies
Molecular Biology of the Cell
Bruce Alberts
Biology
The Communist Manifesto
Karl Marx
Business
Earth Science
Intermediate Microeconomics
Hal R. Varian
Chemistry
East Asia
India
Physics
Math
Religious Studies
Russia
Imagined Communities
Benedict R. O'G. Anderson
A Manual for Writers of Term Papers, Theses, and Dissertations
Kate L. Turabian
Communication Skills Handbook
Jane Summers
Computer Science
The Elements of Style
William Strunk
Media Studies
(Christian) Theology
Information Science
Medieval
Latin America
Classics
Archaeology
Copy link to current view
-17.21/17.99/-10.91/8.01

«Maken»

Foto: Jørgen Schyberg

Searching for similarities

1. Find an image

Searching for similarities

1. Find an image
2. Get its numerical representation

{0.9, 0.12, 0.45, …}

Searching for similarities

1. Find an image
2. Get its numerical representation
3. Compare to the numerical representation of all the other images

{0.9, 0.12, 0.45, …}

{0.9, 0.12, 0.45, …}
{0.1, 0.23, 0.27, …}
{0.2, 0.56, 0.87, …}
{0.9, 0.78, 0.62, …}
{0.6, 0.29, 0.32, …}
{0.9, 0.26, 0.79, …}
{0.9, 0.69, 0.16, …}
{0.8, 0.82, 0.87, …}
{0.5, 0.35, 0.10, …}
{0.3, 0.72, 0.97, …}
…

Searching for similarities

1. Find an image
2. Get its numerical representation
3. Compare to the numerical representation of all the other images
4. Rank by the comparison metric

{0.9, 0.12, 0.45, …}

{0.9, 0.12, 0.45, …} → **0.9**
{0.1, 0.23, 0.27, …} → 0.1
{0.2, 0.56, 0.87, …} → 0.2
{0.9, 0.78, 0.62, …} → **0.8**
{0.6, 0.29, 0.32, …} → 0.4
{0.9, 0.26, 0.79, …} → **0.7**
{0.9, 0.69, 0.16, …} → 0.1
{0.8, 0.82, 0.87, …} → 0.1
{0.5, 0.35, 0.10, …} → 0.2
{0.3, 0.72, 0.97, …} → 0.4
…

Searching for similarities

1. Find an image
2. Get its numerical representation
3. Compare to the numerical representation of all the other images
4. Rank by the comparison metric

{0.9, 0.12, 0.45, …}

{0.9, 0.12, 0.45, …} → **0.9**
{0.1, 0.23, 0.27, …} → 0.1
{0.2, 0.56, 0.87, …} → 0.2
{0.9, 0.78, 0.62, …} → **0.8**
{0.6, 0.29, 0.32, …} → 0.4
{0.9, 0.26, 0.79, …} → **0.7**
{0.9, 0.69, 0.16, …} → 0.1
{0.8, 0.82, 0.87, …} → 0.1
{0.5, 0.35, 0.10, …} → 0.2
{0.3, 0.72, 0.97, …} → 0.4
…

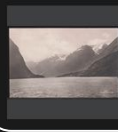
Searching for similarities

1. Find an image
2. Get its numerical representation
3. Compare to the numerical representation of all the other images
4. Rank by the comparison metric

Lake Louise, Alberta, Canada

Leden, Christian | 1909 | Mer informasjon

Lignende bilder:


Postkort fra Olden, Stryn kommune, Sogn og Fjordane bilde 001
1900
Likhet: ●●●●○


Wilse, Anders Beer
Wilse, Anders Beer
1921
Likhet: ●●●◐○

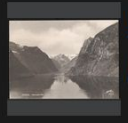

Luktvatn
Mittet & Co. AS
1925
Likhet: ●●●◐○

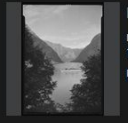

Romsdal.
Mittet & Co. AS
1900
Likhet: ●●●◐○


Romsdalsfjeldene
Mittet & Co. AS
1900
Likhet: ●●●◐○


Norge, Veblungsnes, Romsdalshorn og Vengetindene
Mittet & Co. AS
Likhet: ●●●◐○


Luktvatn
Mittet & Co. AS
1948
Likhet: ●●●◐○


Postkort fra Aurland kommune, Sogn og Fjordane bilde 011
1900
Likhet: ●●●◐○

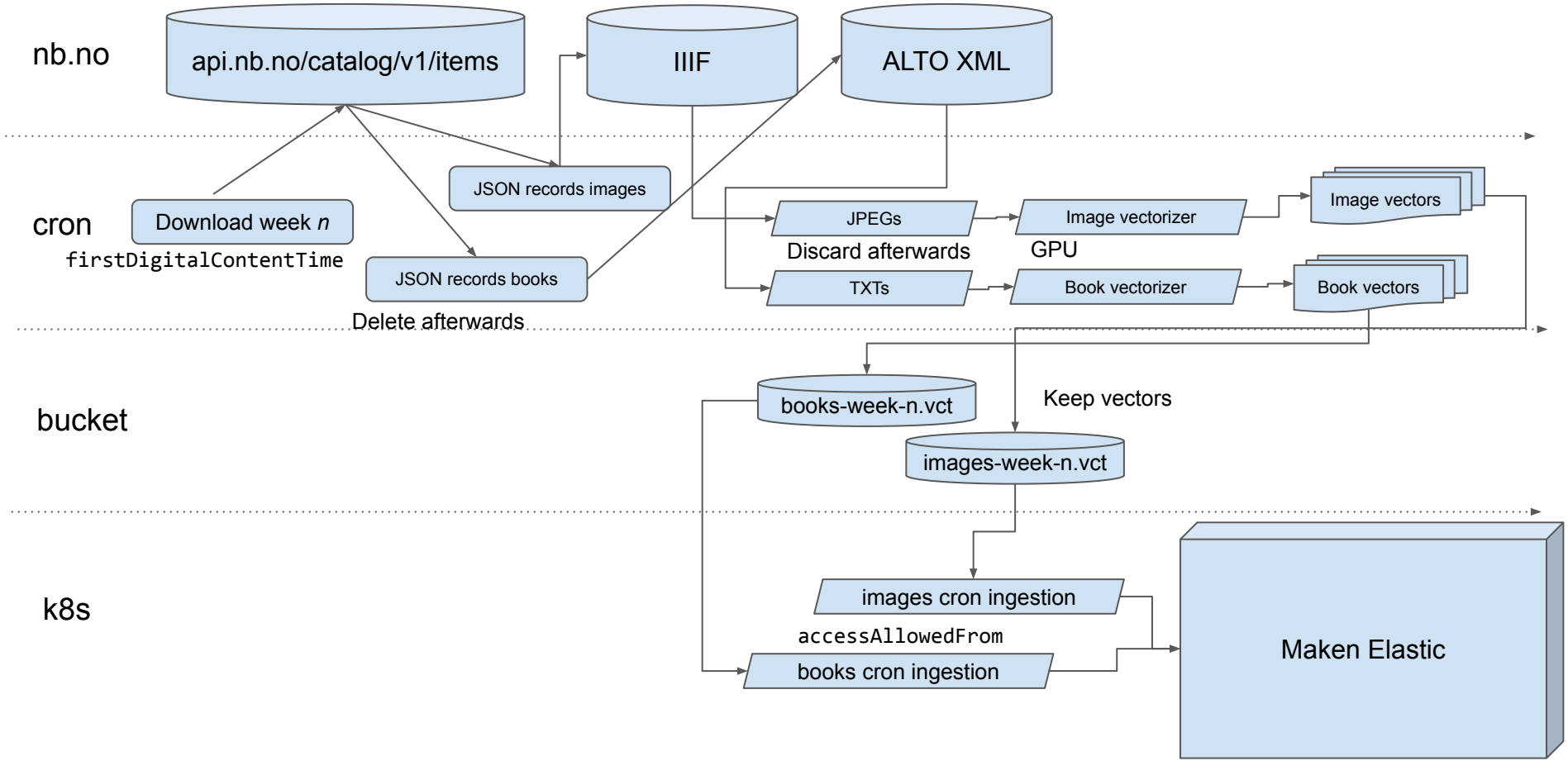

Passasjerskip i Eidfjorden?
Normanns kunstforlag
1950
Likhet: ●●●◐○


Postkort fra Loenvatnet, Stryn kommune, Sogn og Fjordane bilde 015
1900


Norge. Sommerkveld, Merok, Geiranger
Mittet & Co. AS
1938

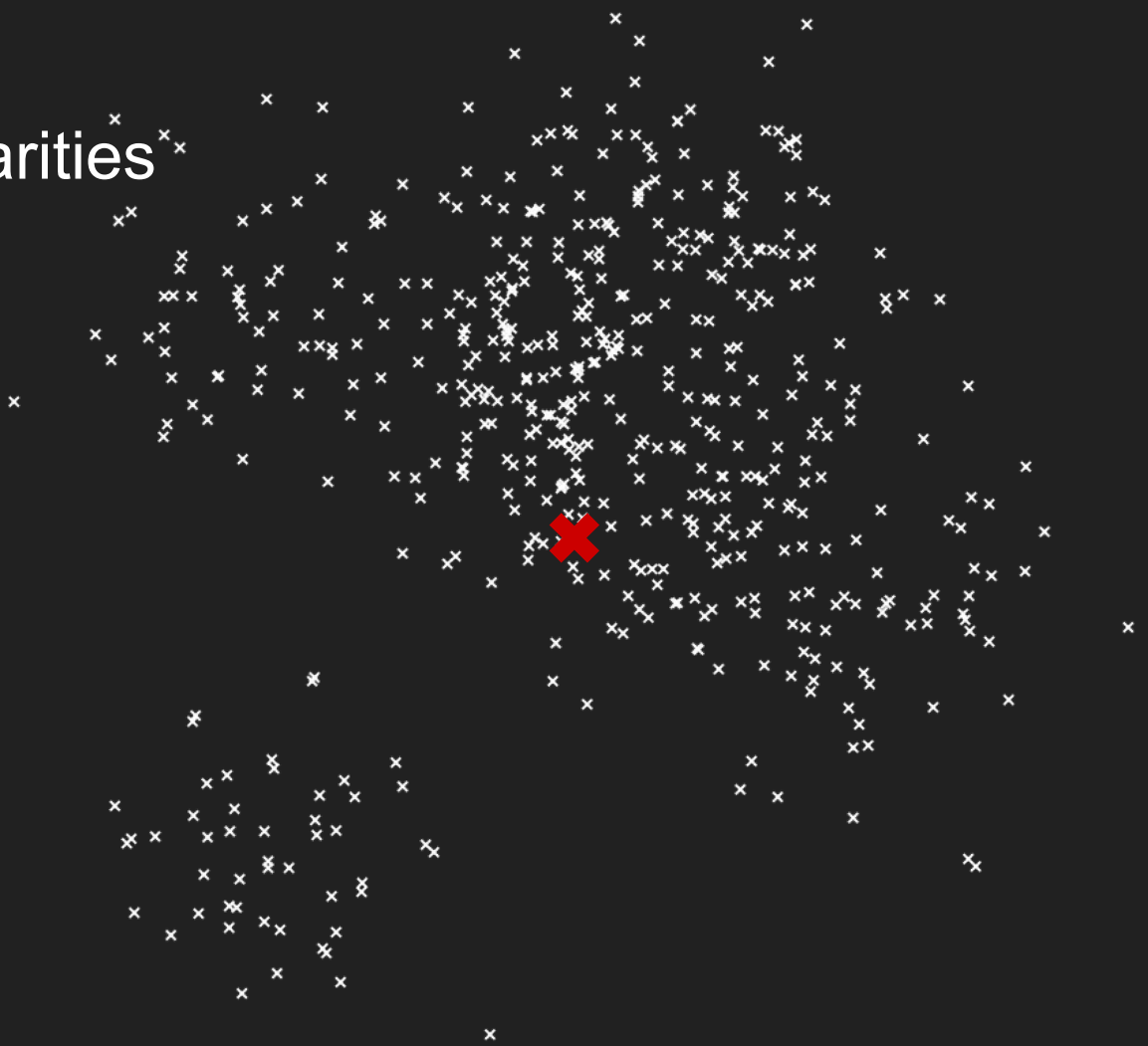

Skjolden
Mittet & Co. AS
1939
Likhet: ●●●◐○

https://github.com/NbAiLab/maken-es-data



Searching for similarities

- Precise ✅
- Slow ❌



Source: https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html

Searching for similarities (approximately)

Approximate Nearest Neighbors

Source: https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html

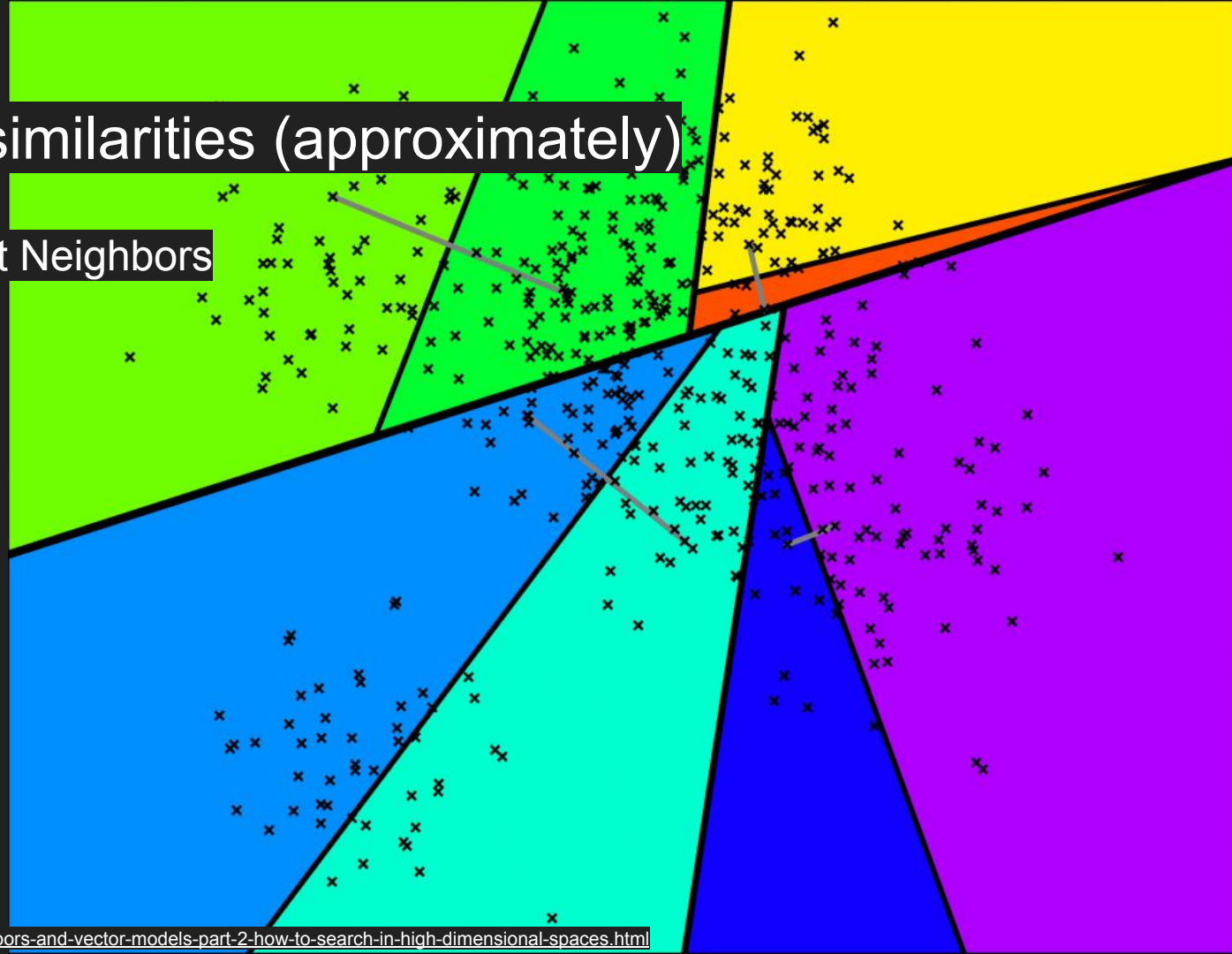
Searching for similarities (approximately)

Approximate Nearest Neighbors

Source: https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html

Searching for similarities (approximately)

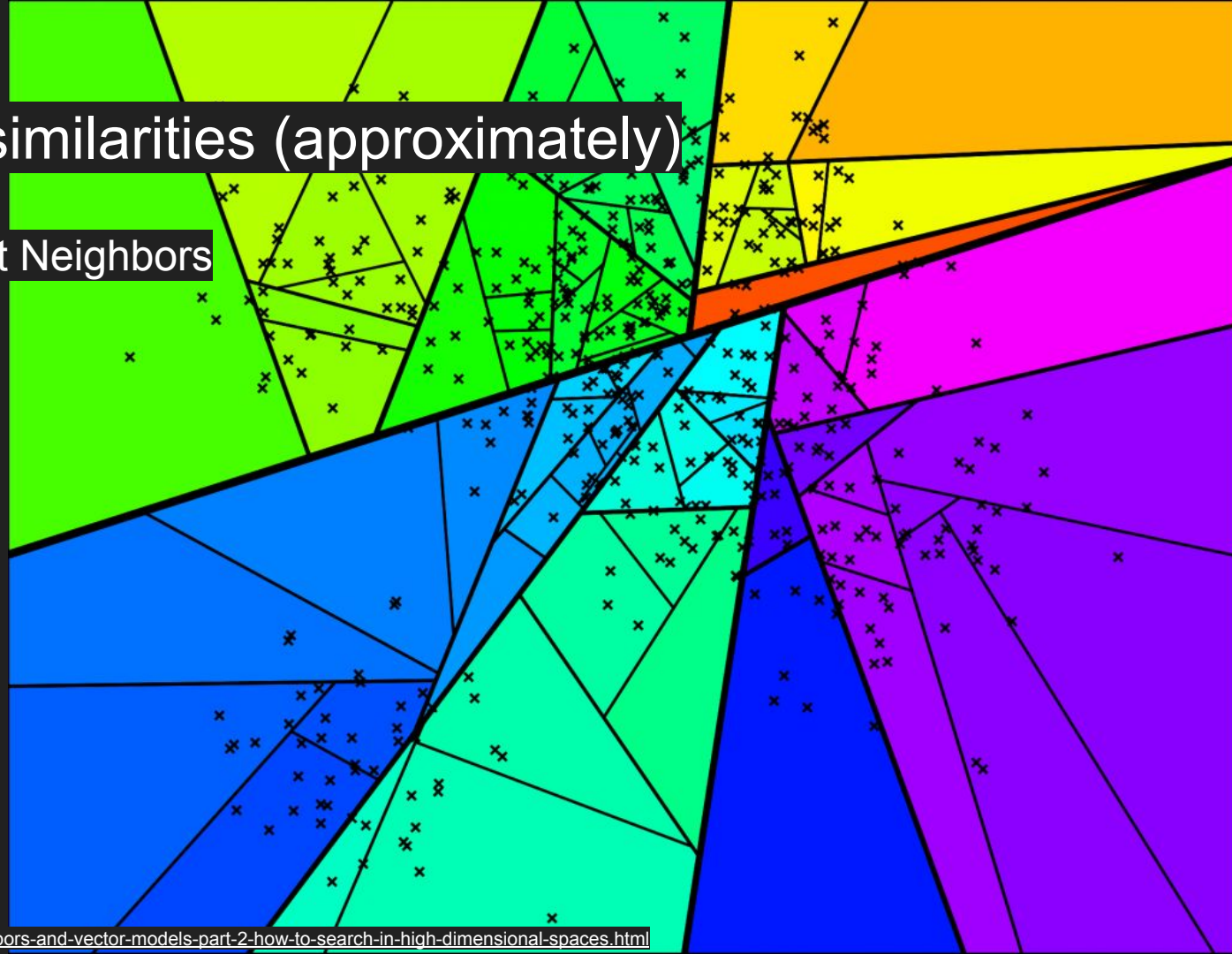
Approximate Nearest Neighbors



Source: https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html

Searching for similarities (approximately)

Approximate Nearest Neighbors

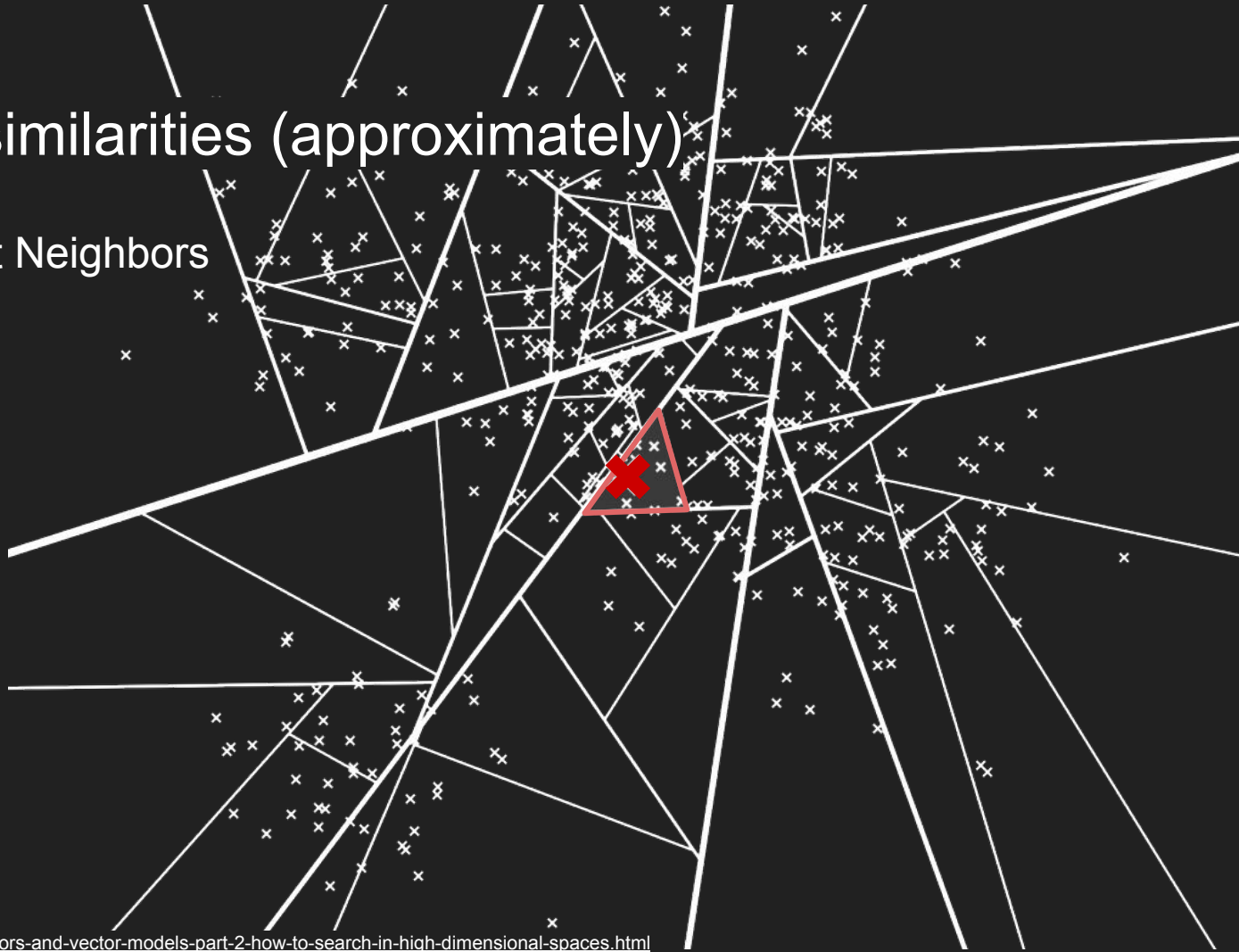


Source: https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html

Searching for similarities (approximately)

Approximate Nearest Neighbors

- Precise 🆗
- Fast ✅



Source: https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html

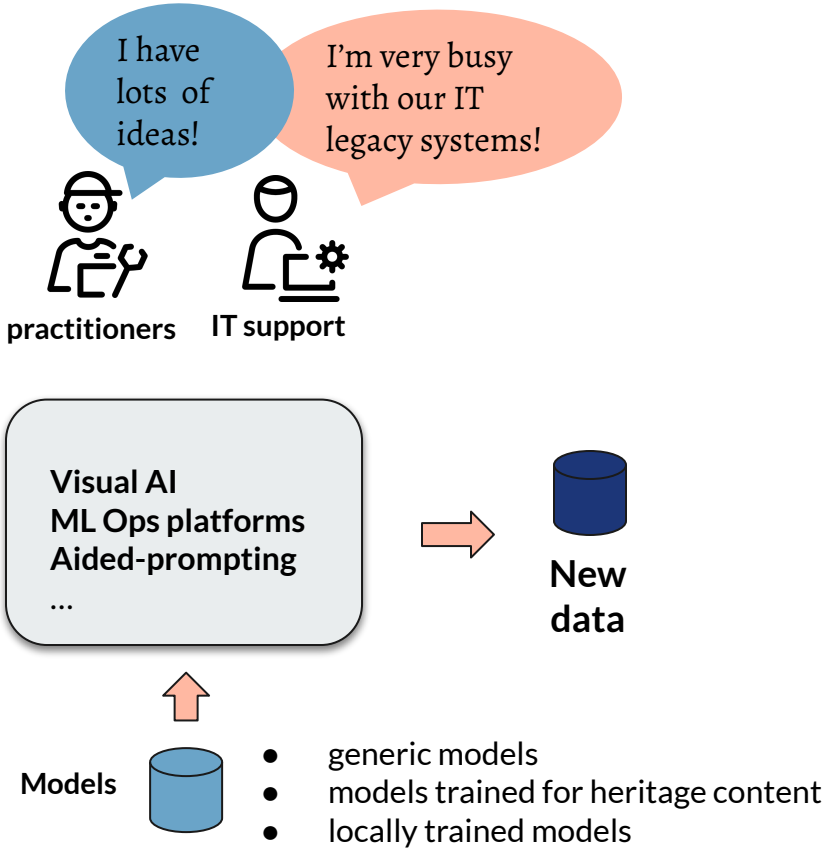
Managing AI workflows for the BnF everyday life?

Commodification of AI

Target: in-house + DH (BnF Datalab users)

Why: low staffing, limited IT resources

What for: building data processing pipelines on top of AI models ("Lego way")



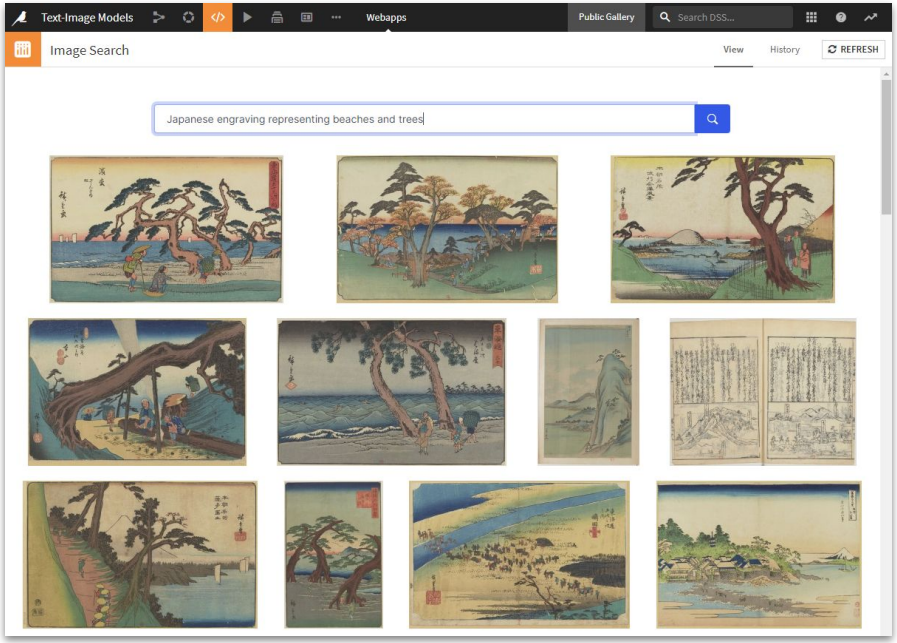
Dataiku DSS experiments

Use case #1: Gallica/classification

Semantic unsupervised image classification with CLIP model (OpenAI), using prompts:

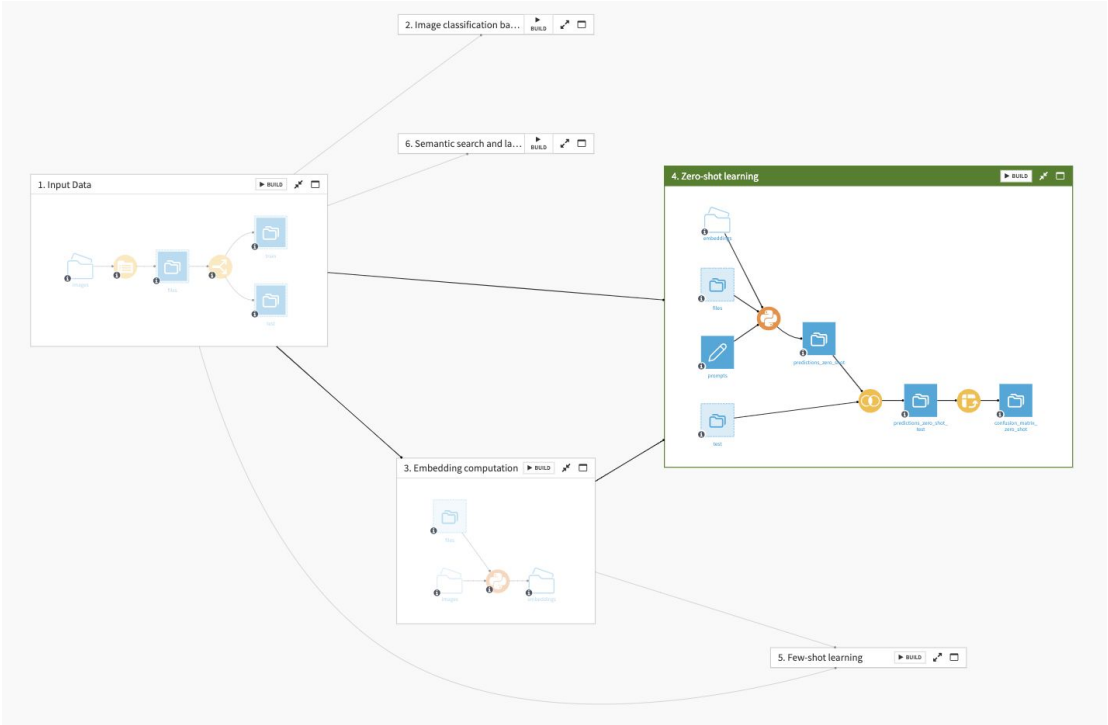
- *"japanese painting"*
- *"japanese ideograms"*
- *"book bindings"*
- *"blank pages"*

on the Gallica Japanese engravings collection



Dataiku experiments

- Zero annotated data (we leverage CLIP “knowledge”)
- Visual design of workflows (Dataiku)
- Accuracy ≃ **95%**



DEMO

https://gallery.dataiku.com/projects/EX_CLIP/
(go to the </> Webapps menu)

Dataiku experiments

Use case #2: Gallica/classification

Newspaper illustrations types classification:

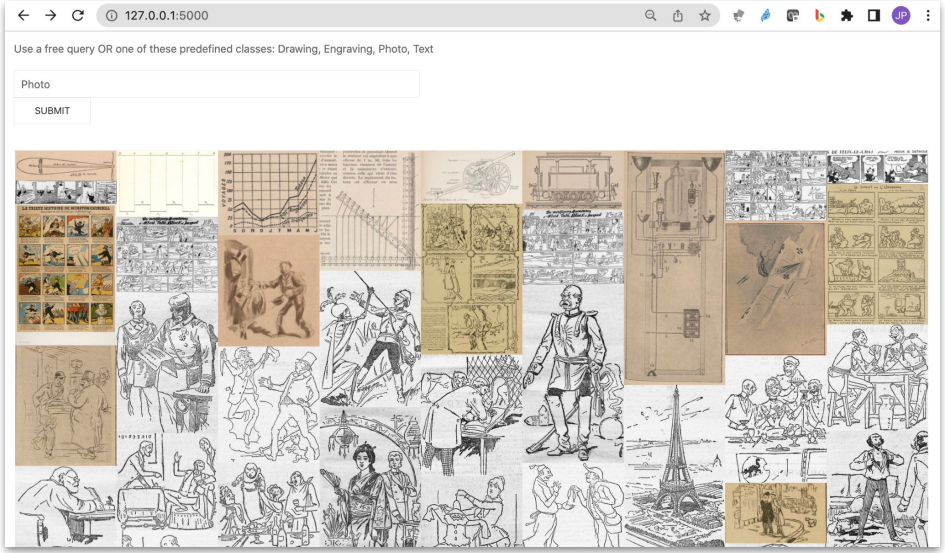
- GT: 9,500 tagged images from French newspapers (1910-1920)
- 4/10 classes (“prompts”)
- **Accuracy ≃ 90%**

Confusion matrix

GT\pred	Chart	Comic	Drawi	Engra	Games	Manus	Map	Photo	Score
Chart	251	0	108	15	8	12	3	2	16
Comic	0	281	11	0	0	0	0	0	0
Drawi	3	23	1738	15	0	2	0	31	0
Engra	0	32	51	537	0	1	0	73	0
Games	0	0	0	41	198	15	0	20	2
Manus	0	0	4	1	0	80	0	2	1
Map	6	0	7	0	0	0	292	2	2
Photo	11	11	73	54	1	0	3	2749	1
Score	0	0	2	0	0	0	0	0	136

```
Accuracy (micro average): 90.40
Accuracy (per classe):
Chart / Comics / Drawing / Engraving / Games / Manuscript / Map / Photo / Score
60.48 % / 96.23 % / 95.92 % / 77.38 % / 71.74 % / 90.91 % / 94.50 % / 94.70 % / 98.55 %
```

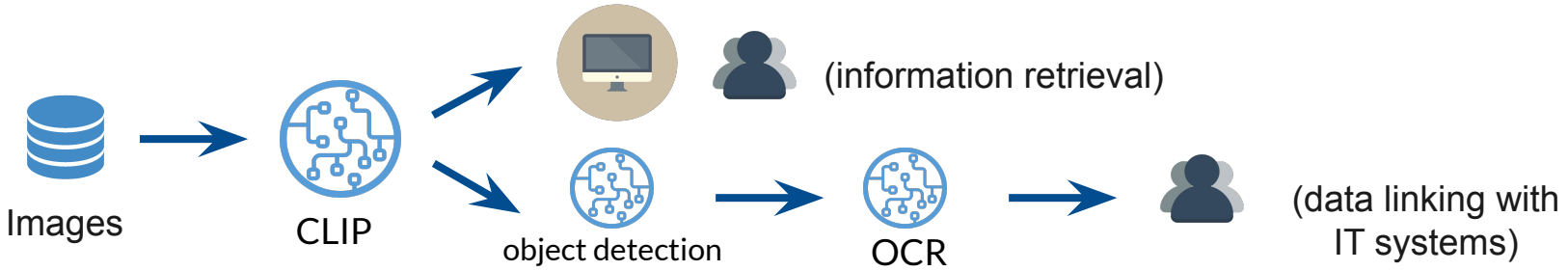
- `Ads,an monochrome illustrated ad printed in a heritage newspaper`
- `Chart,a chart or a diagram printed in a heritage newspaper`
- `Comics,a comics printed in a heritage newspaper`
- `Drawing,a monochrome drawing printed in a heritage newspaper`
- `Engraving,a color or grayscale old engraving printed in …`
- `Games,a crossword grid or a chess game or a word game or a checkers game printed in a heritage newspaper`
- `Manuscript,a handwritten text`
- `Map,a monochrome map printed in a heritage newspaper`
- `Photo,a black and white picture printed in a heritage newspaper`
- `Score,a printed musical score printed in a heritage newspaper`



Other experiments

Use case #3: Conservation dpt

Semantic indexation of a photo bank from the BnF books restoration workshops (50k files)

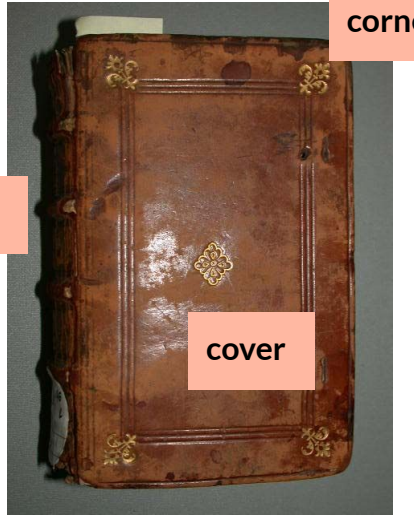


Other experiments

Use case #3: Conservation dpt

Evaluation on 500 images :
accuracy ≈ 34%
(multi-label scenario…)

but effective for an IR task on
a niche context (book restoration)

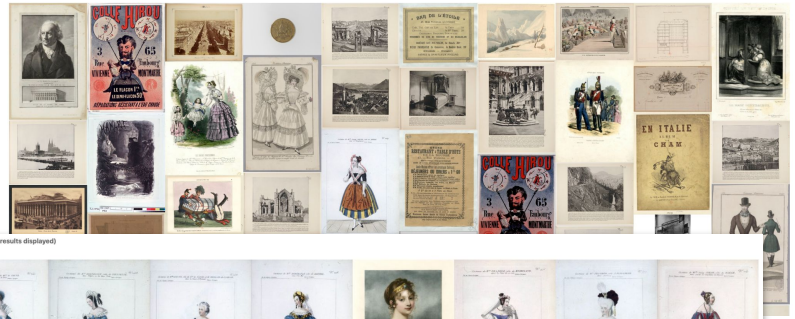


Confusion matrix

GT\pred	Coins	Dos	Etiqu	Fleur	Illus	Nerfs	Papie	Plat
Coins	9	17	0	1	0	1	1	1
Dos	2	36	0	6	0	3	0	1
Etiqu	5	17	11	6	0	2	0	14
Fleur	2	18	0	31	0	7	6	13
Illus	9	2	1	12	19	1	3	47
Nerfs	0	46	1	0	0	26	4	1
Papie	3	4	0	0	0	2	21	4
Plat	7	30	1	5	0	11	4	11

```
Accuracy (micro average): 33.81
Accuracy (per classe):
Coins / Dos / Etiquette_rondage / Fleuron_decor / Illustration / Nerfs / Papier_decor / Plat
30.00 % / 75.00 % / 20.00 % / 40.26 % / 20.21 % / 33.33 % / 61.76 % / 15.94 %
```

Other experiments with CLIP



Use case #4: IR for the digital humanities

"Histoire du quartier Richelieu" project
(INHA, ENC, BnF…)

- Multi-institutional data aggregation
- Visual analysis of small but heterogeneous corpora (engravings, maps, architectural drawings, ads, …)

https://quartier-richelieu.fr/

"a woman wearing a blue dress"

WISE Image Search Engine (WISE) [10mins]

- Developed by Horace Lee, Prasanna Sridhar, Abhishek Dutta (Research Software Engineers at University of Oxford)
- An image search engine built around CLIP
  - Easy to use
  - Fast search speeds
  - Use with your own image collections
- Multimodal search (search with natural language, images, or a combination of both)

- Demo on 50+ million images from Wikimedia Commons

WISE Image Search Engine (WISE) [10mins]

- Code is available open source: https://gitlab.com/vgg/wise/wise
- Feel free to contact us if you would like to use WISE in your own organisation / research
  - {horacelee, prasanna, adutta} @ robots.ox.ac.uk

Thanks!

Questions?

Javier de la Rosa
versae@nb.no
AI-lab
National Library of Norway

Jean-Philippe Moreux
jean-philippe.moreux@bnf.fr
BnF Bibliothèque nationale de France

Horace Lee
horace.lee@eng.ox.ac.uk
University of Oxford