# Similarity search for weakly supervised Machine Learning

### Matthijs Douze

Facebook Al research, Paris

Toulouse symposium on deep learning, Oct 18, 2018



## About me





- Eng + PhD ENSEEIHT, 2004
  - Computer vision
- 10 years at INRIA
  - image/video matching
  - large-scale 3D reconstruction
- Facebook since 2015
  - similarity search
  - knn-graphs (clustering, low-shot learning)
  - unsupervised learning
  - video alignment
  - neural net memorization



















# Facebook Al Research



#### ECOSYSTEMS

**PRODUCTS** 

**TECHNOLOGIES** 





•Created in 2013 by Yann Le Cun •Fundamental, open and collaborative research •160+ people: 50/40/10 scientists/engineers/ students



Locations:

- Menlo Park (HQ of Facebook) [2014]
- NYC [2014]
- Paris [2015] + London [2018]
- Montreal [2017]
- Pittsburgh [2018]
- Seattle [2018]



### Values:

 Openness: publish and open-source • Freedom: researchers have complete control on their agenda Collaboration: with internal and external partners Excellence: focus on most impactful projects (publish or perish) Scale: operate at large scale (computation. data)





#### PERCEPTION



#### **UNDERSTANDING & LEARNING**

#### Artificial intelligence

From Wikipedia, the free encyclopedia

"AI" redirects here. For other uses, see AI and Artificial intellig

Artificial intelligence (All of Al research defines itse environment and takes ac Colloquially, the term "arti functions that humans as solving" (known as Mach facilities once thought to optical character recognit having become a routine successfully understandir systems (such as Chess networks, and interpreting Al research is divided into

	Who invented Convolutional Neura
	I would say Yann LeCun et al but I mi be wrong.
¢,	Hmmm. Tough one. I would say Compu
¢,	I would say their bodies' heat and elect activity but I might be wrong.
ÇÊ,	I would say James Chadwick but I mig be wrong.





PREDICTION



### **ALEXNET (2012)**

### MSRA\_2015 (2015)



### **MASK R-CNN (2017)**

### **MASK R-CNN (2017)**

Mask-RCNN He, Gkioxari, Dollar & Girshick [2018]



### BEFORE | 2017



## Live Pose Detection

### AFTER I 2018



#### PERCEPTION



#### **UNDERSTANDING & LEARNING**

#### Artificial intelligence

From Wikipedia, the free encyclopedia

"AI" redirects here. For other uses, see AI and Artificial intellige

#### Artificial intelligence (AI)

of AI research defines itse environment and takes ac Colloquially, the term "arti functions that humans as solving" (known as Mach facilities once thought to optical character recognit having become a routine successfully understandir systems (such as Chess networks, and interpreting AI research is divided inte





## A man riding a wave on a surfboard in the water.



A giraffe standing on the grass next to a tree.



## An airplane is parked on the tarmac at an airport.



## A man riding a motorcycle on a beach.





## Q: How many of Warsaw's inhabitants

spoke Polish in 1933?



Warsaw		

This article is about the Polish capital. For other uses, see illiura 'Warszowa' redirects here. For other uses, see Warszowa (deardy

"City of Warsaw" redirects here. For the Decord Work! War lighter squadron, see No. 314 Polait Fighter 3

Read Edit View Notory

Q

ands on the Watula River in east-central Poland, roughly 280 kilometres (160 m) from the Battic Sea and 300 res (190 m) from the Carpathian Mountains. Its population is estimated at 1.750 million residents within ropolitan area of 3.105 million residents, which makes Warsaw the M in Union, <sup>102044</sup> The oily limits cover \$16.3 square kilometres (190.8 sq.ml), while the metropolitary are 6.102.40 square kilometres (2.365.38 sq.m).<sup>20</sup>

in 2012 the Economial Intelligence Unit ranked Names as the 3 wried as one of the most liveable offes in Central Europe. Today Warsaw is considered an "Aphe-" global oly. ajor international tourist destination and a significant outural, political and economic hub.1999 Warsan a wide variety of industries, is characterised by PMCD manufa rulacturing and lood processing. The city is a significant centre of research and development, BPO, ITO, as we as of the Polish media industry. The Warsaw Stock Exchange is one of the largest and most important in Europe.<sup>[16]</sup> Frontes, the European Union agency for external border security, has its headquarter Warsaw. It has been said that Warsaw, together with Frankfurt, London, Plans and Barcelone is one of the cities w he highest number of skyscrapers in the European Union.<sup>747</sup> Warsaw has also been called "Eastern Europe's chil

Reading Wikipedia to answer open-domain questions Chen, Fisch, Weston & Bordes [2017]

### **READING WIKIPEDIA**





#### PERCEPTION





#### Artificial intelligence

From Wikipedia, the free encyclopedia

"AI" redirects here. For other uses, see AI and Artificial intellige

Artificial intelligence (Al of Al research defines itse environment and takes ad Colloquially, the term "arti functions that humans as solving" (known as Mach facilities once thought to optical character recognit having become a routine successfully understandir systems (such as Chess networks, and interpreting Al research is divided into

	I would say Yann LeCun et al but I mi be wrong.
Ç,	Hmmm. Tough one. I would say Compu
Č,	I would say their bodies' heat and elect activity but I might be wrong.
	I would say James Chadwick but I mig be wrong.

#### PLANNING

#### PREDICTION





### **GETTING AN INTUITION OF PHYSICS**



### PREDICTING FUTURE FRAMES





### **PREDICTING FUTURE FRAMES**







Ground truth

Approx. posterior



DesIGN: Design Inspiration from Generative Networks. Sbai, Couprie, Elhoseiny, Bordes & LeCun [2018]



#### PERCEPTION





#### Artificial intelligence

From Wikipedia, the free encyclopedia

"AI" redirects here. For other uses, see AI and Artificial intellige

Artificial intelligence (All of Al research defines itse environment and takes ad Colloquially, the term "arti functions that humans as solving" (known as Machi facilities once thought to optical character recogni having become a routine successfully understandir systems (such as Chess networks, and interpreting Al research is divided into

	I would say Yann LeCun et al but I mi be wrong.
Ç,	Hmmm. Tough one. I would say Compu
Č,	I would say their bodies' heat and elect activity but I might be wrong.
	I would say James Chadwick but I mig be wrong.

## PLANNING



### PREDICTION



### BEFORE TRAINING



### AFTER TRAINING





.....

## **Common Sense** Humanoid robot Efficient Learning Long-term planning

and a second second

"Good enough" translation Narrow Chatbots Go (game) Handwriting Recognition Chess

## UNSUPERVISED LEARNING

### "Autonomous" driving

### Visual-impaired assistance

### Speech Recognition

### Picture segmentation

## **SUPERVISED** LEARNING



## **Common Sense** Humanoid robot Efficient Learning Long-term planning

and a second second

"Good enough" translation Narrow Chatbots Go (game) Handwriting Recognition Chess

## UNSUPERVISED LEARNING

### "Autonomous" driving

### Visual-impaired assistance

### Speech Recognition

### Picture segmentation

## **SUPERVISED** LEARNING



# Supervised/unsupervised learning

### Supervised learning: the magenet case saturn

- training set: a large and balanced set of positive examples, labelled unambiguously
- excellent testbed for the improvement of image classification algorithms
- leap forward in 2010-2015
  - see this morning's presentation
- now: more or less solved



[ConvNets and ImageNet Beyond Accuracy: Understanding Mistakes and Uncovering Biases, Pierre Stock, Moustapha

school-bus scorpion-101 screwdriver segway self-propelled-law sextant sheet-music skateboard skunk skyscraper smokestack snail snake sneaker snowmobile soccer-ball socks soda-can spaghetti speed-boat spider spoon stained-glass starfish-101 steering-wheel stirrups sunflower-101 superman sushi swan swiss-army-knife sword syringe <u>لم ، ام ام الم</u>





## Supervised learning Behind the scenes

- Getting many images is not a problem
- Manual annotation: an endeavor
  - millions of images
  - several opinions
  - quality control...
- Al powered by... Human intelligence
- How can we reduce the level of supervision?



## Less supervised learning

- Noisy supervision: learn from hashtags -scale up to 3B images
- Weak supervision: object detector with only image-level supervision
- Low-shot learning: train from few images per class
- Zero-shot learning: learn from metadata, eg. a description rather than examples

[Exploring the Limits of Weakly Supervised Pretraining, Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin

#### Target task: ImageNet-1k





# Embeddings



like Schroepfe

Earlier this year we announced Surround 360, a camera that shoots and renders stereoscopic 360 video. Today we're making both the hardware specs and the software freely available so filmmakers can have access to technology that will help them create great 3D-360 content more quickly. You can find everything at https://github.com/facebook/Surround360.

Our team built most of the camera with off-the-shelf hardware to make it easier for others to build too. We also created softwa... See More



	425		12 Comments	71 Shares	11K Views
┢ Like	Comment	A Share		Тор С	Comments -

## **Rost embedding**

Video

embedding

### User embedding









# **Comparing embeddings**

- Using the distance between embeddings
- should be a measure of semantic similarity



- the k-nearest neighbor classifier
- in the following: 2 works on reducing supervision using similarity search



0.6

## **Indexing embeddings** Build index for a collection:





Query:

### Criteria: compact, fast, accurate





# Low-shot learning

[Low-shot learning with large-scale diffusion, Douze, Szlam, Hariharan, Jégou, CVPR'18]

# Problem setup

Too little data to effectively train a CNN

- Typical approach: transfer learning
  - train an embedding on other classes
  - SVM or logistic regression classifier on top of the embeddings

# • Assume we have only 1,2, ..10 training images per class

# **Diffusion on a knn-graph**

- We have:
  - a large set of background images
  - semantic embeddings for the images
- knn-graph
  - nodes = images
  - edges = links to the k nearest images



## Diffusion on a knn-graph Matrix view

$$L_{0} = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix} \qquad \qquad L_{i+1} = \begin{bmatrix} \mathbf{W}_{\mathrm{LL}} & \mathbf{W}_{\mathrm{LB}} \\ \mathbf{W}_{\mathrm{BL}} & \mathbf{W}_{\mathrm{BB}} \end{bmatrix} \times L_{i}$$

- matrix symmetrized and row-normalized
- no normalizations necessary during diffusion
- L1-normalize the L vector at the end of iterations
- early stopping (validated parameter)
- Can be performed on all classes at once
  - sparse matrix dense martrix multiplications





# Visualize

triumphal arch



koala



### **Failure cases** – correct path first, and path produced by the method:







mosque

mosque

- Visualization of the strongest path
- starting from the target images
- follow strongest edges













(triumphal arch)



koala



koala









(koala)

(jack-o'-lantern)



mosque

mosque

(planetarium)

## Results

 classification performance on a s (n=1..20) training images

	out-of-domain diffusion			in-domain	logistic	combined		[16]	
n	none	F1M	F10M	F100M	Imagenet	regression	h +F10M	+ F100M	
1	57.1	60.0	61.4	62.3	68.0	57.3	62.0	62.6	60.6
2	62.5	65.5	66.8	67.8	73.2	66.0	68.7	69.2	68.9
5	68.4	70.6	71.9	73.1	77.8	76.4	76.9	77.4	77.3
10	72.7	74.2	75.3	76.2	80.1	80.9	81.3	81.5	80.6
20	76.0	77.0	77.5	78.6	81.4	83.7	83.9	84.1	82.5

### classification performance on a subset of ImageNet classes, with few

## Results

 runtime depends on number of edges

 graph completion is slow

useful in practice?





# Deep clustering

[Deep Clustering for Unsupervised Learning of Visual Features, Caron, Bojanowski, Joulin, Douze, ECCV'18]

## **Unsupervised** feature learning Experimental context

- Given many unlabeled images, find an embedding
- How can we check whether the embedding is good? fine-tune the convnet on another dataset that has fewer images (Pascal VOC)
- - see how it performs
- Related work:
  - invariance to data aguentation
  - pretext tasks: counting, patch layout prediction [Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles, Noroozi & Favaro, ECCV'16]



## Our approach Keep it simple!

- random initialization of the convnet
- Iterate
  - k-means clustering of descriptors
  - train convnet to predict the clusters



## relies on the convolutional structure of the feature extraction

## Parameters



(a) Clustering quality (b) Clu

convergence in ~300 epochs

- Sobel filter
- data augmentation

(b) Cluster reassignment

(c) Influence of k

## Visualization

## Direct visualization of the first convolutional filters



# Visualization

## strongest images for some filters (last convolutional layer)



Filter 97

Filter 116









Filter 194



Filter 182



#### Filter 119



















## Results

- Image classification performance when using features at a certain level of the network
- variants:
  - use images from Imagenet vs. Flickr
  - use power iteration clustering vs. k-means

- Method
- ImageN Random Random
- Pathak Donahu Pathak Owens Wang at Doersch Bojanov Zhang e
- Norooz Noroozi

DeepCl

DeepClu

FC6-8ALLFC6-8ALLFC6-8ALLfet labels $78.9$ $79.9$ $ 56.8$ $ 48.9$ h-rgb $33.2$ $57.0$ $22.2$ $44.5$ $15.2$ $30.9$ h-sobel $29.0$ $61.9$ $18.9$ $47.9$ $13.0$ $32.9$ et al. $[38]$ $34.6$ $56.5$ $ 44.5$ $ 29.9$ e et al. $[20]^*$ $52.3$ $60.1$ $ 46.9$ $ 35.9$	)     )
Tet labels $78.9$ $79.9$ $ 56.8$ $ 48.9$ n-rgb $33.2$ $57.0$ $22.2$ $44.5$ $15.2$ $30.9$ n-sobel $29.0$ $61.9$ $18.9$ $47.9$ $13.0$ $32.9$ et al. $[38]$ $34.6$ $56.5$ $ 44.5$ $ 29.9$ e et al. $[20]^*$ $52.3$ $60.1$ $ 46.9$ $ 35.9$	) [ )
n-rgb $33.2$ $57.0$ $22.2$ $44.5$ $15.2$ $30.$ n-sobel $29.0$ $61.9$ $18.9$ $47.9$ $13.0$ $32.$ et al. $[38]$ $34.6$ $56.5$ $ 44.5$ $ 29.$ e et al. $[20]^*$ $52.3$ $60.1$ $ 46.9$ $ 35.$	[ )
n-sobel29.061.918.947.913.032. $et al. [38]$ 34.656.5-44.5-29.6 $e et al. [20]^*$ 52.360.1-46.9-35.6	)
et al. $[38]$ $34.6$ $56.5$ - $44.5$ - $29.$ e et al. $[20]^*$ $52.3$ $60.1$ - $46.9$ - $35.7$	
e et al. $[20]^*$ 52.3 60.1 - 46.9 - 35.4	7
	2
$et \ al. \ [27] \qquad - \ 61.0 \qquad - \ 52.2 \qquad - \ -$	
$et \ al. \ [44]^* \qquad 52.3  61.3  -  -  -$	
nd Gupta $[29]^*$ 55.6 63.1 $32.8^{\dagger}$ 47.2 $26.0^{\dagger}$ 35.4	F <sub>4</sub>
$t et al. [25]^* 55.1 65.3 - 51.1 $	
wski and Joulin $[19]^*$ 56.7 65.3 $33.7^{\dagger}$ 49.4 $26.7^{\dagger}$ 37.	†
et al. $[28]^*$ 61.5 65.9 43.4 <sup>†</sup> 46.9 35.8 <sup>†</sup> 35.4 <sup>†</sup>	3
$et \ al. \ [43]^* \qquad \qquad 63.0  67.1 \qquad -  46.7 \qquad -  36.4$	)
and Favaro $[26]$ – 67.6 – 53.2 – 37.	)
$et \ al. \ [45] - 67.7 - 51.4 - 36.$	)
uster <b>72.0 73.7 51.4 55.4 43.2 45.</b>	1
uster YFCC100M 67.3 69.3 45.6 53.0 39.2 42.2	





Conclusion

- We have to go to less supervised learning
- Similarity search:
  - non-parametric method for media matching
  - efficient / scalable
- A fundamental tool for unsupervised learning
  - not only for vision: see [Word translation without parallel data, Conneau, Lample, Ranzato, Denoyer, Jégou, Arxiv'18]
- All works are open-sourced
  - check it out!

Source: sed ut unde omnis



Source: sed ut unde omnis

