MACHINES THAT SEE: IMAGE ANNOTATION USING DEEP LEARNING

Prof. Lior Wolf
Tel Aviv University

What is deep learning?

- Large Artificial Neural Networks
- Making computer perception = human perception

Voice analysis

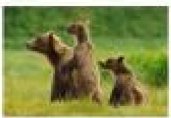
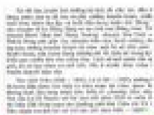
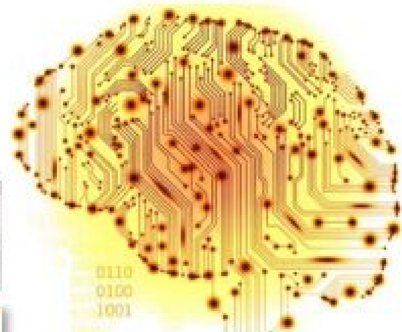


Image analysis



Text analysis

Robotics



Where did I hear about deep learning?

- Eveywhere in the news
- Fast evolving + strong competition

Stanford University Berkeley UNIVERSITY OF CALIFORNIA Baidu Google

IDSIA UNIVERSITY OF TORONTO NYU Microsoft facebook

Evolution of AI

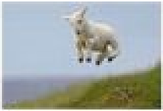


Image annotation

2015

Computer Science
@Tel Aviv University

Face recognition

NATURAL PERCEPTION

2014

facebook

CS
@TAU

Chess

MAN-MADE

1996

IBM

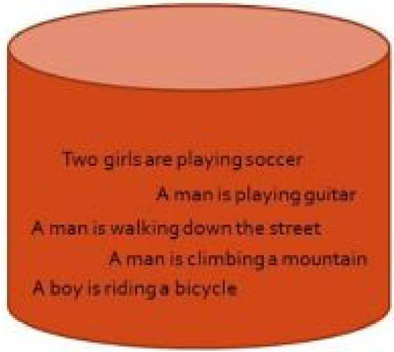

Evolution of AI
2015
Computer Science
@Tel Aviv University
Image annotation
LEARNING
2014
facebook
CS
@TAU
LEARNING
Face recognition
1996
SEARCH
IBM
Chess
LEARNING

FROM IMAGE TO TEXT AND BACK USING DEEP LEARNING

Associating Neural Word Embeddings with Deep Image Representations using Fisher Vectors

Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf
The Blavatnik School of Computer Science
Tel Aviv University

Task I: Image Annotation



Task II: Image Search

A man is playing guitar

Task II: Image Search

A man is playing guitar

Task III: Description synthesis

Input: new unseen image

Output: new description in English

Two girls playing soccer.

Technical Agenda

Image representation

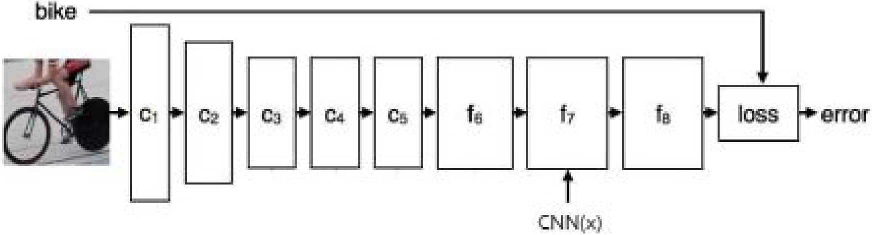
Sentence representation

Linking the two

Synthesis of new sentences

Image representation

We employ a pretrained Deep Convolutional Neural Network (CNN)



VGG by *Andrea Vedaldi and Andrew Zisserman*

Sentence representation

Natural Language Processing (NLP) as computer vision guys

1. What are the local descriptors?
   Google's Word2vec

2. How to combine (pool) the local descriptors?
   A new type of Fisher Vectors

Word2Vec

Word2Vec transforms word in English to representation with a semantic properties.

$$word2vec(\text{"A"}) = (131, 128, 111, 10, \ldots, 14, 11) \in \mathrm{R}^D$$
$$word2vec(\text{"playing"}) = (11, 61, 2, 13, \ldots, 11, 10) \in \mathrm{R}^D$$

.

.

.

$$word2vec(\text{"guitar"}) = (21, 122, 14, 1, \ldots, 110, 1) \in \mathrm{R}^D$$

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean.
Efficient Estimation of Word Representations in Vector Space.
In Proceedings of Workshop at ICLR, 2013.

Word2Vec

```
In [24]: model.most_similar('python')
Out[24]:
[('scripting', 0.91207885742 1875),
('bash', 0.9030072093009949),
('perl', 0.897027850151062),
('tcl', 0.8833462595939636),
('ruby', 0.872918367385 8643),
('c++', 0.8634607195854187),
('jython', 0.8467384576797485),
('groovy', 0.846560001373291),
('lua', 0.8416544795036316)]
```

Sentence representation

A man is playing guitar

Local Feature Extraction

$word2vec(\text{"A"}) = (131, 128, 111, 10, \ldots, 14, 11) \in \mathrm{R}^D$

$word2vec(\text{"playing"}) = (11, 61, 2, 13, \ldots, 11, 10) \in \mathrm{R}^D$

.

.

.

$word2vec(\text{"guitar"}) = (21, 122, 14, 1, \ldots, 110, 1) \in \mathrm{R}^D$

Sentence representation

$$word2vec(\text{"A"}) = (131, 128, 111, 10, \ldots, 14, 11) \in \mathrm{R}^D$$
$$word2vec(\text{"playing"}) = (11, 61, 2, 13, \ldots, 11, 10) \in \mathrm{R}^D$$

.

.

.

$$word2vec(\text{"guitar"}) = (21, 122, 14, 1, \ldots, 110, 1) \in \mathrm{R}^D$$

Pooling

$HGLMM\ FV \in R^m$

Generic representation: dataset independent

Text vector embeddings are not modeled fully by Gaussians

Univariate Gaussian: $g(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Univariate Laplacian: $l(x, m, s) = \frac{1}{2s} e^{-\frac{|x-m|}{s}}$

Univariate Hybrid Gaussian Laplacian: $l(x, m, s)^b \cdot g(x, \mu, \sigma)^{1-b}$

Multivariate Hybrid Gaussian Laplacian: $\prod_{d=1}^{D} l(x_d, m_d, s_d)^{b_d} \cdot g(x_d, \mu_d, \sigma_d)^{1-b_d}$

Multivariate Hybrid Gaussian Laplacian Mixture Models:

$$\sum_{k=1}^{K} \prod_{d=1}^{D} l\left(x_{k,d}, m_{k,d}, s_{k,d}\right)^{b_{k,d}} \cdot g\left(x_{k,d}, \mu_{k,d}, \sigma_{k,d}\right)^{1-b_{k,d}}$$

Hybrid Gaussian Laplacian Mixture Model EM

Let $X_{trn} = \{x_1, x_2, \ldots, x_n\} \in R^D$ be the train set for the EM.

- **Estimation Step:**

$$p(z_i = k | x = x_i; \lambda^t) = T_{k,i}^t = \frac{\tau_k^t \cdot l(x, m_k^t, s_k^t)^{b^t} \cdot g(x, \mu_k^t, \sigma_k^t)^{1-b^t}}{\sum_{r=1}^K \tau_r^t \cdot l(x, m_r^t, s_r^t)^{b^t} \cdot g(x, \mu_r^t, \sigma_r^t)^{1-b^t}}$$

- **Maximization Step:**

$$\tau_k^{(t+1)} = \frac{\sum_{i=1}^N T_{k,i}^{(t)}}{\sum_{r=1}^K \sum_{i=1}^N T_{r,i}^{(t)}} \qquad \mu_{k,d}^{(t+1)} = \frac{\sum_{i=1}^N T_{k,i}^{(t)} \cdot x_{i,d}}{\sum_{i=1}^N T_{k,i}^{(t)}} \qquad (\sigma_{k,d}^{(t+1)})^2 = \frac{\sum_{i=1}^N T_{k,i}^{(t)} \left(x_{i,d} - \mu_{k,d}^{(t+1)}\right)^2}{\sum_{i=1}^N T_{k,i}^{(t)}}$$

$$\sum_{m_{k,d}^{(t+1)} \le x_{i,d}} T_{k,i}^{(t)} = \sum_{m_{k,d}^{(t+1)} > x_{i,d}} T_{k,i}^{(t)} \qquad s_{k,d}^{(t+1)} = \frac{\sum_{i=1}^N T_{k,i}^{(t)} \left| x_{i,d} - m_{k,d}^{(t+1)} \right|}{\sum_{i=1}^N T_{k,i}^{(t)}} \qquad b_{k,d}^{(t+1)} = \begin{cases} 1 & \text{if } L_{b_{k,d}}^{(t+1)} > G_{b_{k,d}}^{(t+1)} \\ 0 & \text{otherwise} \end{cases}$$

Hybrid Gaussian Laplacian Mixture Model Fisher Vector

We prove that there is a hard selection for each coordinate and each component, therefore:

For $b_{k,d} = 0$:

$$\frac{\partial \mathcal{L}(X|\lambda)}{\partial \mu_{k,d}} = \sum_{i=1}^{N} T_{k,i} \cdot \frac{x_{i,d} - \mu_{k,d}}{\sigma_{k,d}^2}$$

$$\frac{\partial \mathcal{L}(X|\lambda)}{\partial \sigma_{k,d}} = \sum_{i=1}^{N} T_{k,i} \left(\frac{(x_{i,d} - \mu_{k,d})^2}{\sigma_{k,d}^3} - \frac{1}{\sigma_{j,d}} \right)$$

For $b_{k,d} = 1$:

$$\frac{\partial \mathcal{L}(X|\lambda)}{\partial m_{k,d}} = \sum_{i=1}^{N} \frac{T_{k,i}}{s_{k,d}} \cdot \begin{cases} 1 & \text{if } x_{i,d} > m_{k,d} \\ -1 & \text{otherwise} \end{cases}$$

$$\frac{\partial \mathcal{L}(X|\lambda)}{\partial s_{k,d}} = \sum_{i=1}^{N} T_{k,i} \left(\frac{|x_{i,d} - m_{k,d}|}{s_{k,d}^2} - \frac{1}{s_{k,d}} \right)$$

Fisher Information Matrix (FIM)

More algebra (14 page supplementary)

Bringing Images and Sentences to the same domain

We present each image by a CNN representation

$$CNN(Image) \in R^{D_{Image}}$$

We present each sentence by the HGLMM Fisher Vector representation

$$FV(Sentence) \in R^{D_{Sentence}}$$

Using **Canonical Correlation Analysis**, we learn two projections W_{Image} and $W_{sentence}$, such that:

$$W_{Image} \cdot CNN(Image) \in R^{common}$$

$$W_{Sentence} \cdot FV(Sentence) \in R^{common}$$

Image Annotation

Image Annotation

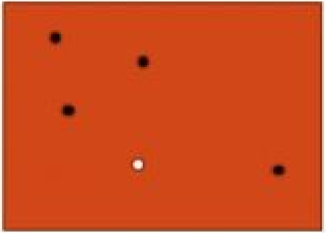


Image Annotation

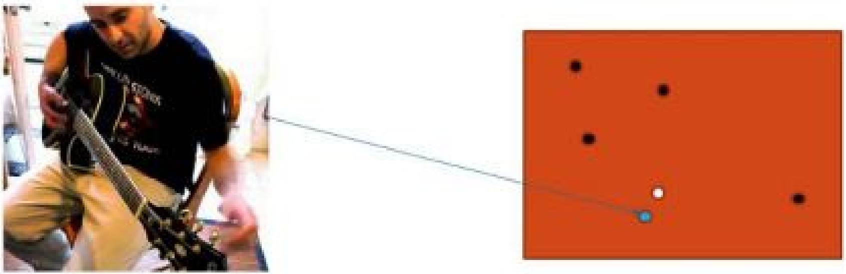


Image Search

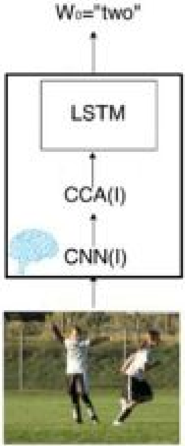
A man is playing guitar

Sentence Synthesis?

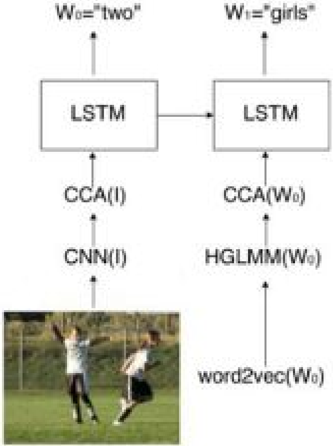
No need to model English explicitly

Intead, we use recurrent neural networks

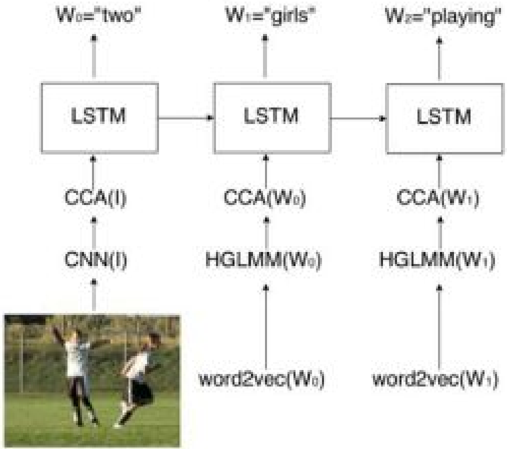
Sentence Synthesis



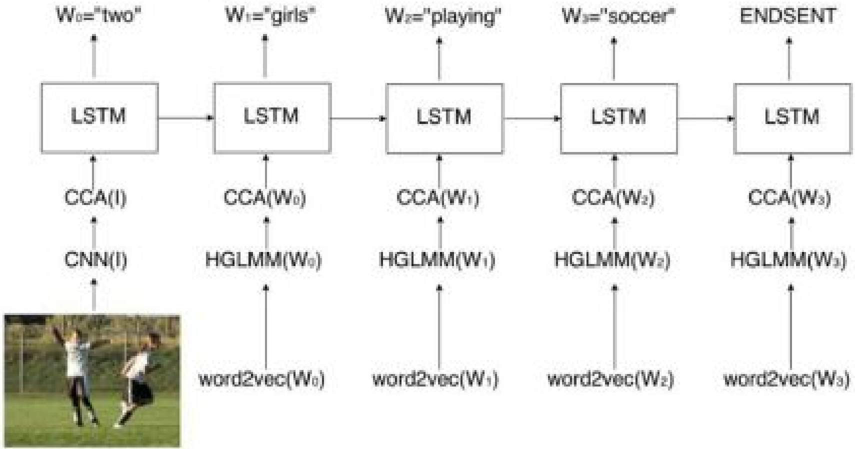
Sentence Synthesis



Sentence Synthesis



Sentence Synthesis



Results

Synthesis:
94% of the generated sentences are new
comparded to 20% in Google's system
imageannotator.cs.tau.ac.il

a dog with ball in its mouth

a basketball player in the uniform is running in the air

A man in red shirt is climbing up the rock face

A skier is jumping over snow covered hill

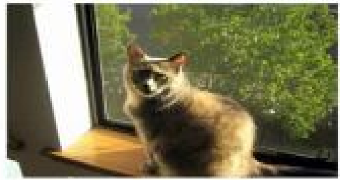
a boy is jumping into pool

two dogs are playing in the water

- The model was trained on only 8,000 images

Currently  < 

a man in red shirt is riding his bike



a dog is jumping up at the ground

a man is sitting on the ground

a man is holding up his hand on the ground

Image annotation

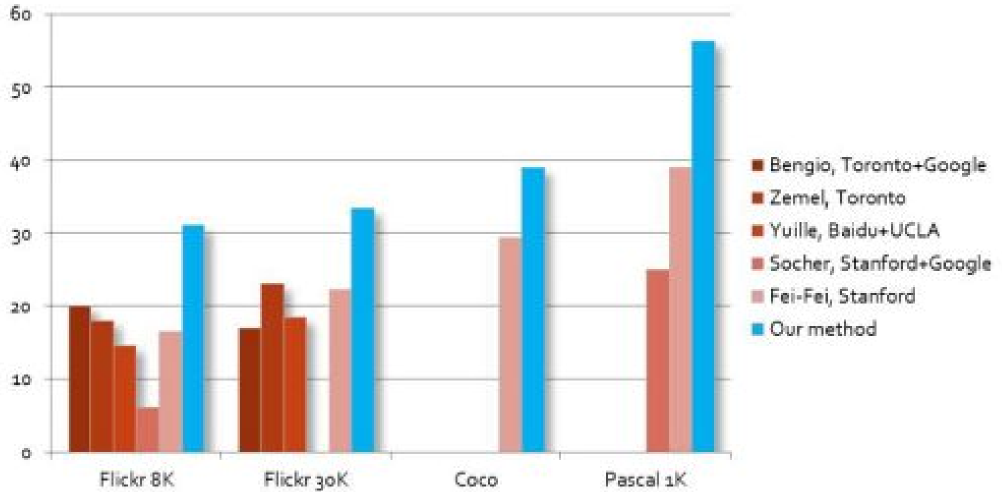
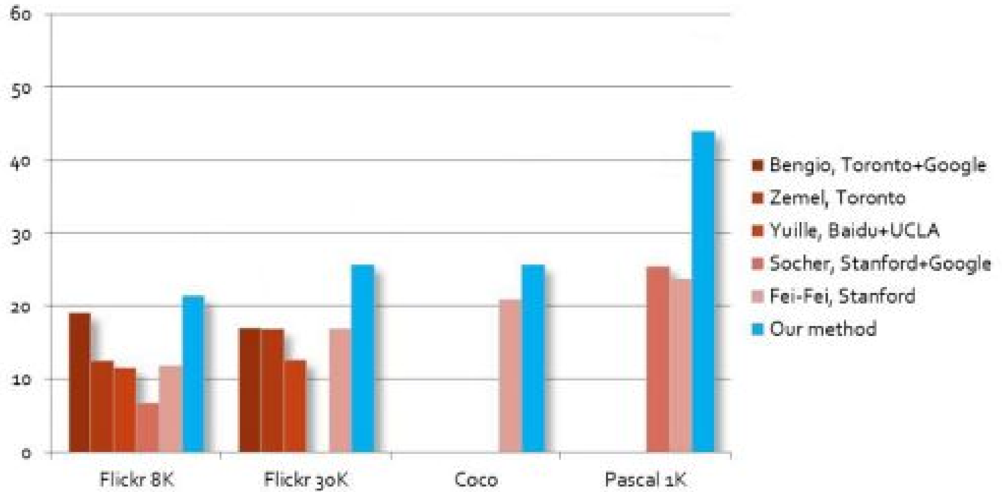


Image search



Future work

- Discuss images (Q&A)
- Train on much larger sets of images
- Describe videos

Input: unseen image + a question

What are the girls doing?

Output: the answer

They are playing soccer.

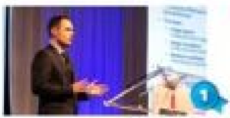
News from a few days ago

BROADCOM
FOUNDATION

University | Overview | University Research Competition | Winners

COMPETITION WINNERS

2015 2014 2013 2012



First Place

Benjamin Klein, Tel Aviv University, Israel, for his project "From Image to Text and Back Using Deep Learning."Klein is creating a unified semantic representation that allows computers to understand images and sentences using deep neural networks.In the short term, this work could be used for creating search engines on images and for systems that could assist the blind and visually impaired.

Image search

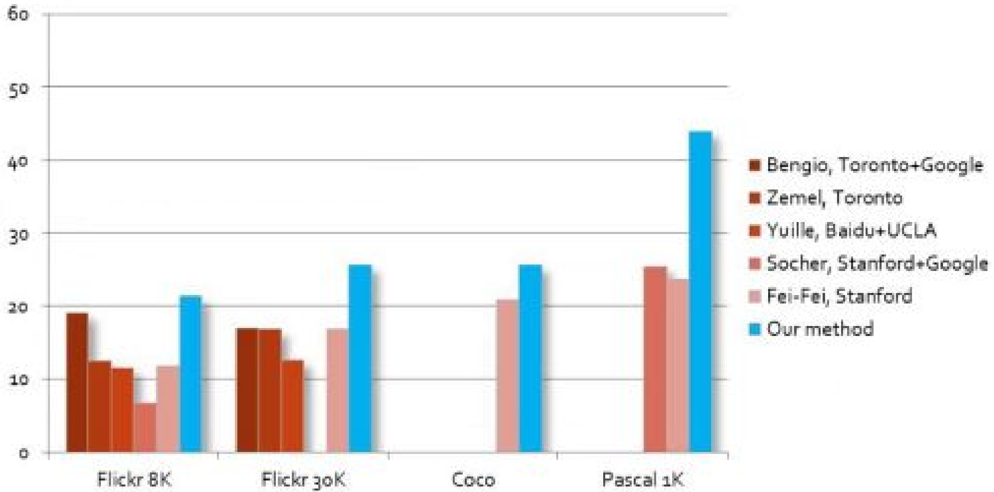


Image annotation

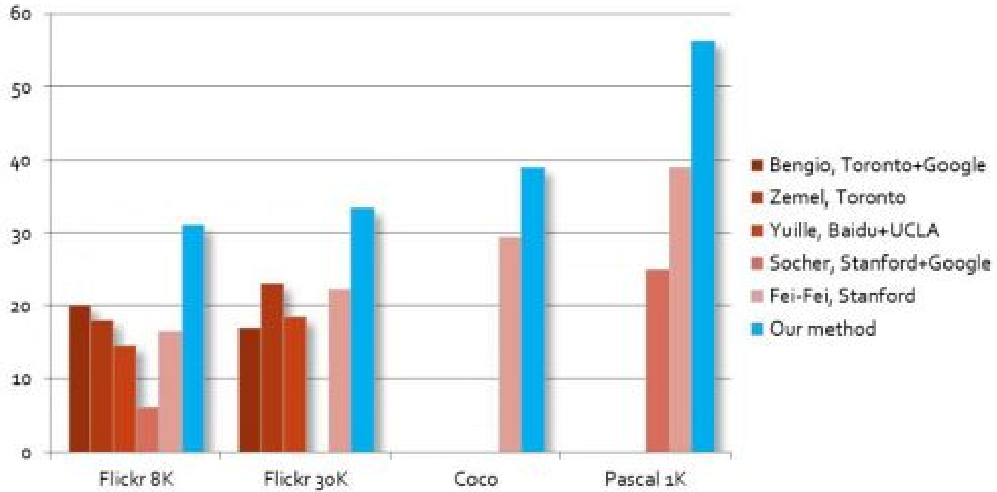
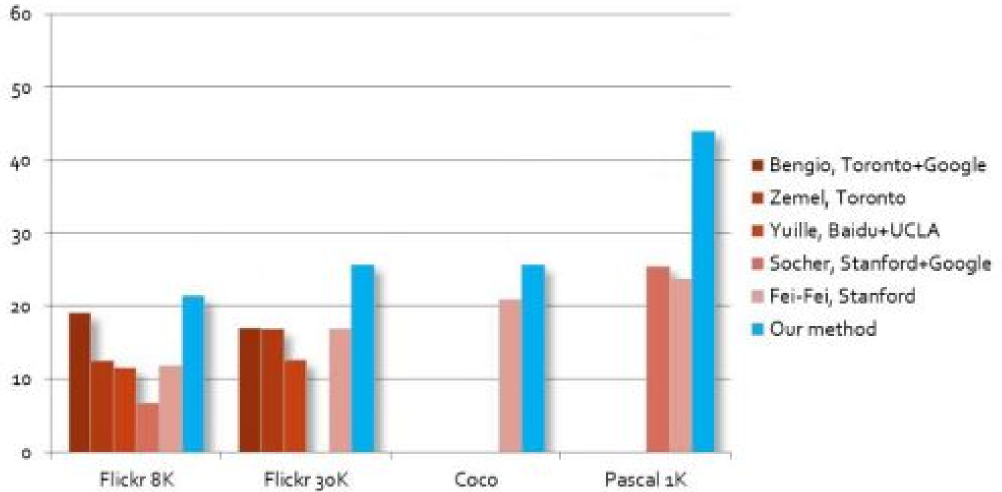


Image search



UNDERSTANDING NATURAL LANGUAGE

In Defense of Word Embedding for Generic Text Representation

Guy Lev, Benjamin Klein, and Lior Wolf

The Blavatnik School of Computer Science
Tel Aviv University, Tel Aviv, Israel

Generalized Gaussian Mixture Model

Can we enjoy both worlds?

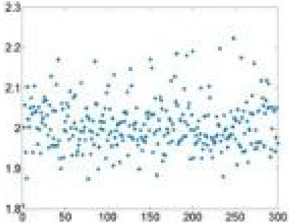
Univariate Gaussian: $g(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

Univariate Laplacian: $l(x,m,s) = \frac{1}{2s} e^{-\frac{|x-m|}{s}}$

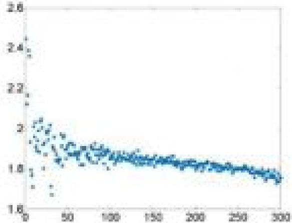
Univariate Hybrid Gaussian Laplacian: $l(x,m,s)^b \cdot g(x,\mu,\sigma)^{1-b}$

Univariate Generalized Gaussin: $ggd(x;m,s,p) = \frac{1}{2sp^{1/p}\Gamma(1+1/p)} exp\left(-\frac{|x-m|^p}{ps^p}\right)$

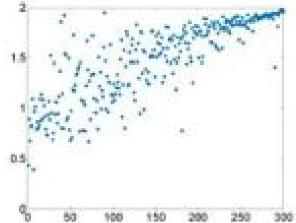
Distribution of p of word2vec



Original 300 dimensions



After PCA



After ICA

Answer Sentence Selection

Factual question: What is the brightest star visible from Earth?

Candidate sentences:

- In the year <number>, Voyager will pass within <number> light years, <number> trillion miles, of Proxima Centauri, the nearest star.
- Voyager will be headed toward Sirius, the brightest star in the heavens, after it leaves our solar system.
- Near Sirius in Year <number>.
- Then engineers will turn off Voyager <number>'s TV cameras and its infrared and visible light sensors.
- In the year <number>, Voyager <number> will make its closest approach to Sirius, the brightest star visible from Earth.

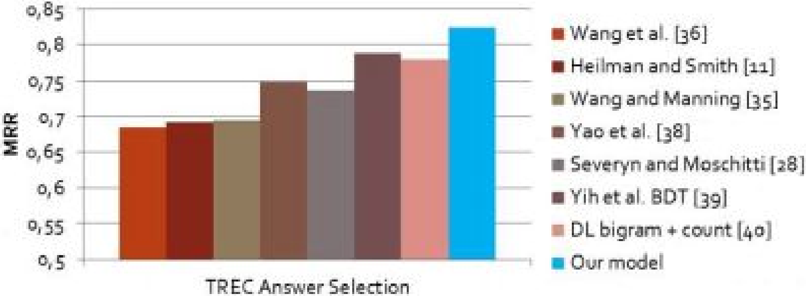
Answer Sentence Selection

Factual question: What is the brightest star visible from Earth?

Candidate sentences:

- In the year <number>, Voyager will pass within <number> light years, <number> trillion miles, of Proxima Centauri, the nearest star.
- Voyager will be headed toward Sirius, the brightest star in the heavens, after it leaves our solar system.
- Near Sirius in Year <number>.
- Then engineers will turn off Voyager <number>'s TV cameras and its infrared and visible light sensors.
- In the year <number>, Voyager <number> will make its closest approach to Sirius, the brightest star visible from Earth.

Answer Sentence Selection - Results



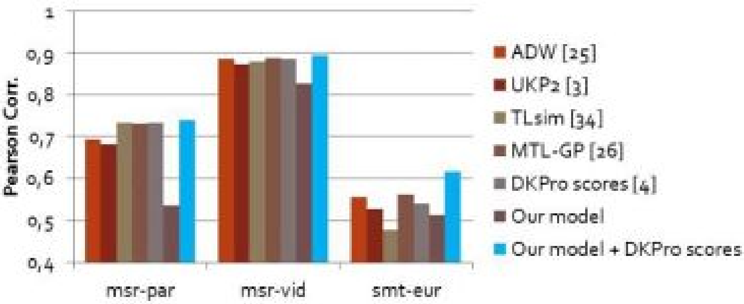
[36] Wang, Mengqiu, Noah A. Smith, and Teruko Mitamura: What is the Jeopardy model? a quasi-synchronous grammar for question answering. EMNLP 2007.
[11] Heilman, Michael, and Noah A. Smith: Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. ACL 2010.
[35] Wang, Mengqiu, and Christopher D. Manning: Probabilistic tree-edit models with structured latent variables for textual entailment and question answering. ACL 2010.
[38] Yao, Xuchen, et al.: Answer Extraction as Sequence Tagging with Tree Edit Distance. HLT-NAACL 2013.
[28] Severyn, Aliaksei, and Alessandro Moschitti: Automatic Feature Engineering for Answer Selection and Extraction. EMNLP 2013.
[39] Yih, Wen-tau, et al.: Question answering using enhanced lexical semantic models. 2013.
[40] Yu, Lei, et al.: Deep learning for answer sentence selection. NIPS 2014.

Semantic Sentence Similarity

- Input: a pair of sentences
- Output: a similarity score between 0-5

A woman is riding a horse. A woman is riding a donkey.	2.8
A man is riding a bicycle. A man is riding a bike.	5
Someone is drawing. Someone is dancing.	0.3

Semantic Sentence Similarity - Results



[25] Pilehvar, T.M., Jurgens, D., Navigli, R.: Align, disambiguate and walk: a unified approach for measuring semantic similarity. ACL 2013.
[3] Bär, D., Biemann, C., Gurevych, I., Zesch, T.: UKP: computing semantic textual similarity by combining multiple content similarity measures. ACL 2012.
[34] Šarić, Frane, et al.: Takelab: Systems for measuring semantic text similarity. ACL 2012
[26] Rios, Miguel, Lucia Specia: Uow: Multi-task learning gaussian process for semantic textual similarity. SemEval 2014
[4] Bär, Daniel, Torsten Zesch, and Iryna Gurevych: DKPro Similarity: An Open Source Framework for Text Similarity. ACL 2013.

Topic Classification

"LeBron James had 26 points, seven rebounds and seven assists and became the youngest player in NBA history to score 2,000 points, as the red-hot Cleveland Cavaliers pasted the Chicago Bulls, 96-74, at Gund Arena."

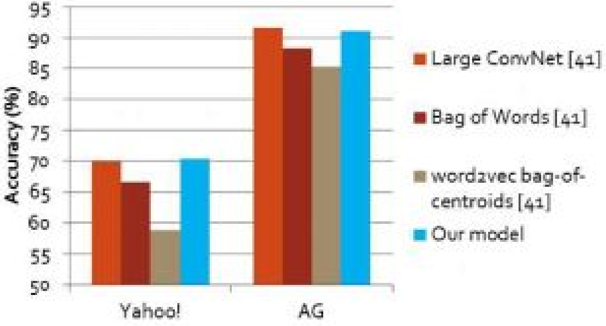
Business	Sports	Politics	…	Health

Topic Classification

"LeBron James had 26 points, seven rebounds and seven assists and became the youngest player in NBA history to score 2,000 points, as the red-hot Cleveland Cavaliers pasted the Chicago Bulls, 96-74, at Gund Arena."

Business	Sports	Politics	...	Health

Topic Classification - Results



[41] Zhang, X., LeCun, Y.: Text Understanding from Scratch. ArXiv e-prints

"it might also be the case that the hope for linear separability of word2vec is not valid at all"

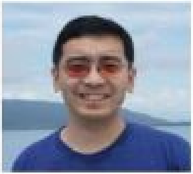
HUMAN LEVEL FACE RECOGNITION

- **DeepFace: Closing the Gap to Human-Level Performance in Face Verification**
- **Web-Scale Training for Face Identification**

Yaniv Taigman Ming Yang Marc'Aurelio Ranzato Lior Wolf

Facebook AI Group
Menlo Park, CA, USA
{yaniv, mingyang, ranzato}@fb.com

Tel Aviv University
Tel Aviv, Israel
wolf@cs.tau.ac.il


Why faces?

1. One class. Billions of unique instances.
2. The most frequent entity in the media by far: e.g. ~1.2 faces / Photo on avg
3. Many applications

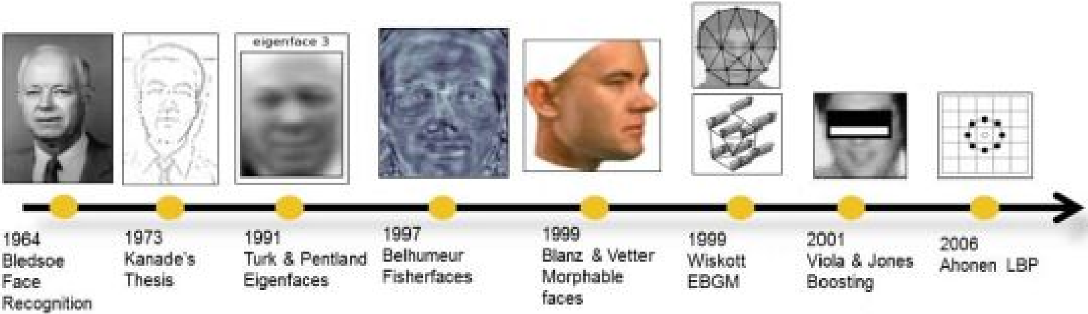
Tag Suggestions

Face Recognition main objective

Find a representation & similarity measure such that:

- Intra-subject similarity is high
- Inter-subject similarity is low

Milestones in Face Recognition



Slightly modified version of Anil Jain's timeline

Problem Solved?

NIST's best-performer's gets:

1. Its internal dataset with 1.6 million identities: 95.9%
2. On LFW (public) with 'only' 4,249 identities: 56.7%

→ Answer: No.

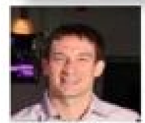
L. Best-Rowden, H. Han, C. Otto, B. Klare, and A. K. Jain.

Challenges in Unconstrained Face Recognition

1. Pose
2. Illumination
3. Expression
4. Aging
5. Occlusion

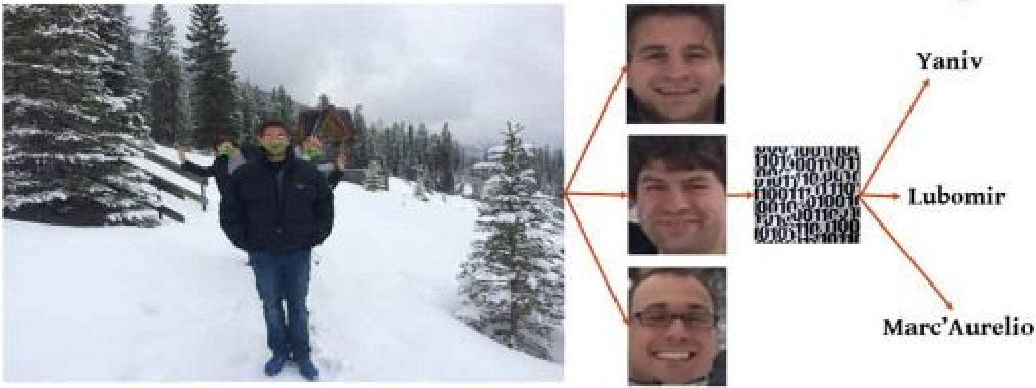
Probes for example

Gallery

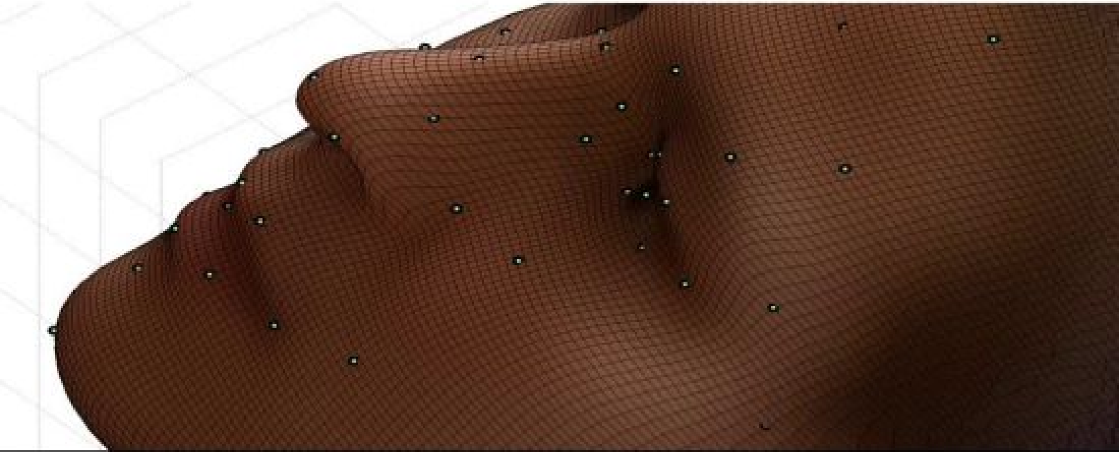


Face Recognition Pipeline

Detect → Align → Represent → Classify



Faces are 3D objects



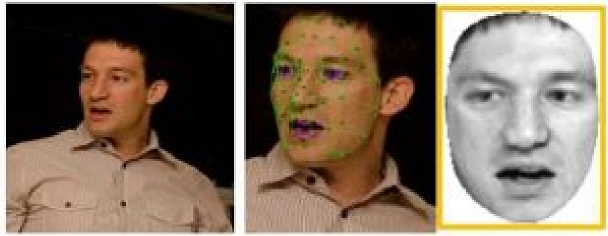
Face alignment ('Frontalization')

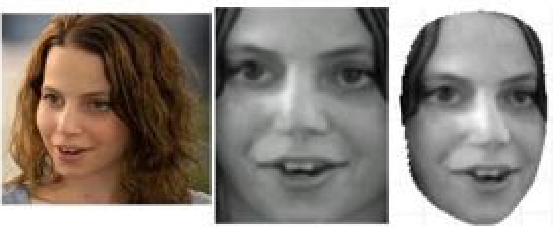
Detect

2D-Aligned

3D-Aligned

Examples

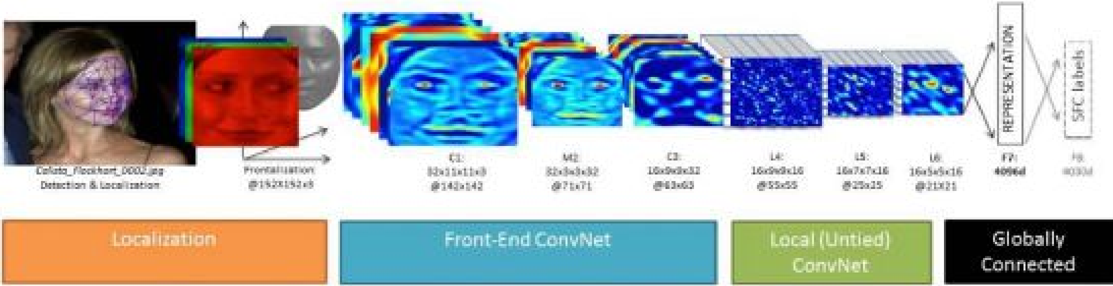




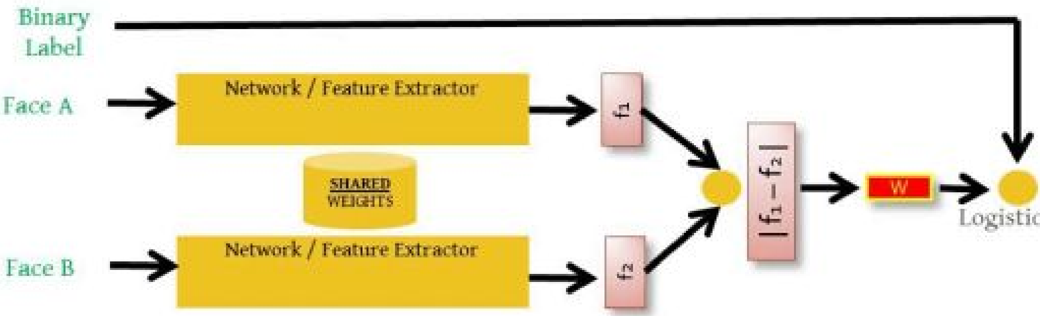
SFC Training Dataset

4.4 million photos blindly sampled, belonging to 4,030 identities

Deep Neural Networks on aligned inputs



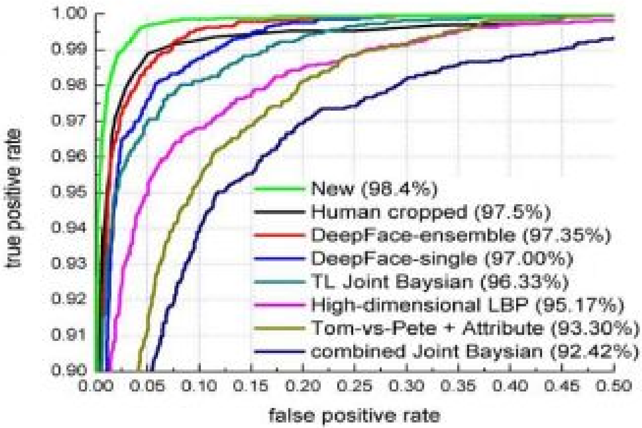
Deep Siamese Architecture [1]



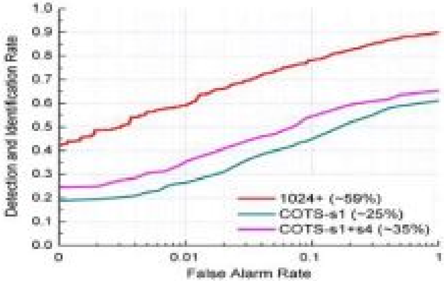
$$p = \frac{1}{1 + e^{-(W * |f_1 - f_2| + b)}}$$

$$E = -ylog(p) - (1-y)log(1-p)$$

Second round results



Comparison to NIST's State Of The Art



Second-round DeepFace

Same system that achieved 92% Rank-1 accuracy on a table of 1.6 million identities. (NIST's State-Of-The-Art, Constrained)

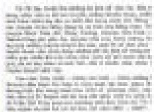
Method	DeepFace [20]	BLS [3]*	COTS-s1 [1]	COTS-s1+s4 [1]	**1024+**	**Fusion**
Verification	97.35	93.18	-	-	98.00	98.37
Rank-1	64.9	18.1	56.7	66.5	82.1	82.5
DIR @ 1%	44.5	7.89	$\boxed{25}$	35	$\boxed{59.2}$	61.9

@brink of human performance: 1.72% FN

I'VE SPOKEN ENOUGH
ANY QUESTIONS?

Voice analysis

Image analysis



Text analysis

Robotics

@brink of human performance: 1.72% FN