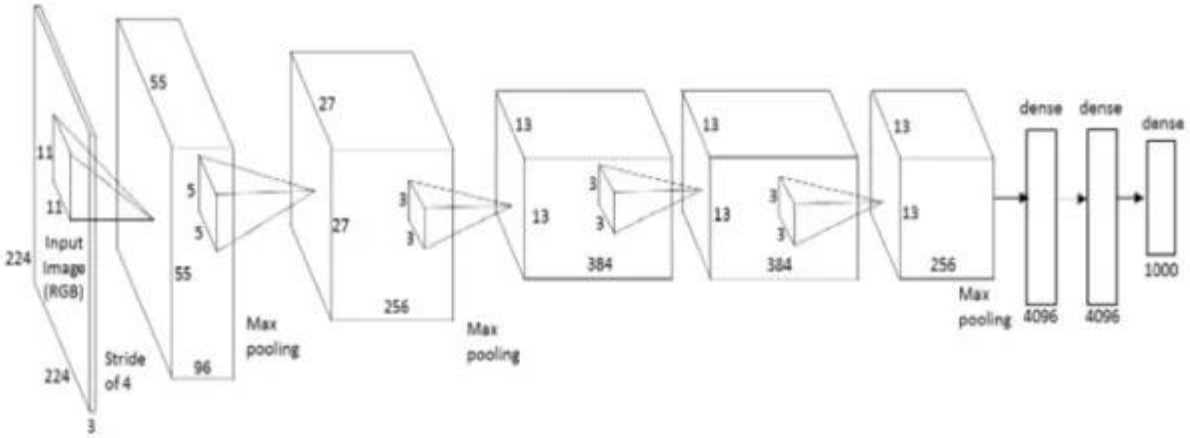
Lecture 7: DL Continued, Face Recognition, Ethical Concerns, Intro to Adversarial Attacks

Siddharth Garg

sg175@nyu.edu

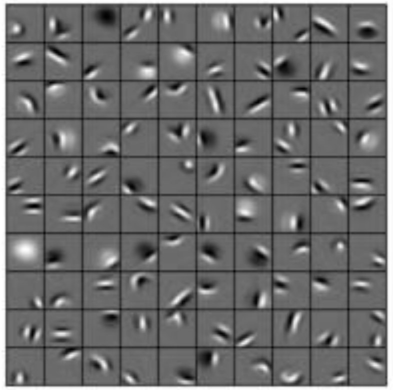
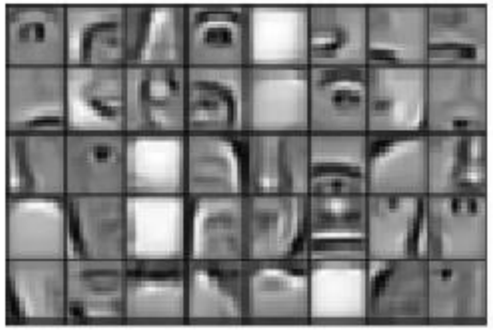
Alex Net

- Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton University of Toronto, 2012
- Key idea: Build a very deep neural network
- 60 million parameters, 650,000 neurons
- 5 conv layers + 3 FC layers
- Final is 1000-way softmax



Local Features

- Early layers in deep neural networks often find local features
- Small patterns in larger image
  - Examples: Small lines, curves, edges
- Build more complex classification from the local features


Localization via a Sliding Window

- Simple idea: Find local feature by sliding window
- Large image: $X \ N_1 \times N_2$ (e.g. 512 x 512)
- Small filter: $W \ K_1 \times K_2$ (e.g. 8 x 8)
- At each offset (i, j) compute:

$$Z[i,j] = \sum_{k_1=0}^{K_1-1} \sum_{k_2=0}^{K_2-1} W[k_1, k_2] X[i + k_1, j + k_2$$

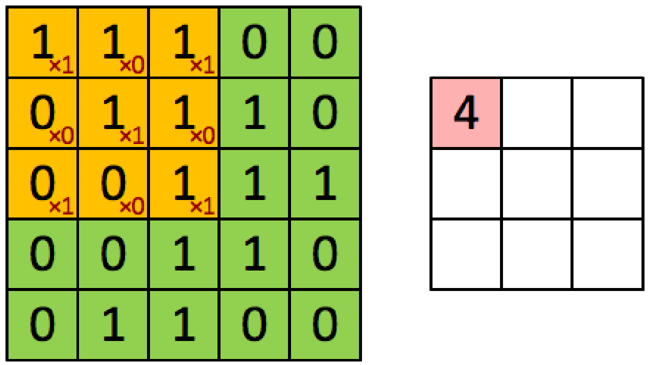
  - Correlation of W with image box starting at (i, j)
  - $Z[i,j]$ is large if feature is present around (i, j)

Convolution 2D Example

- Kernel

$$W = \widetilde{W} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

- Compute convolution in valid region

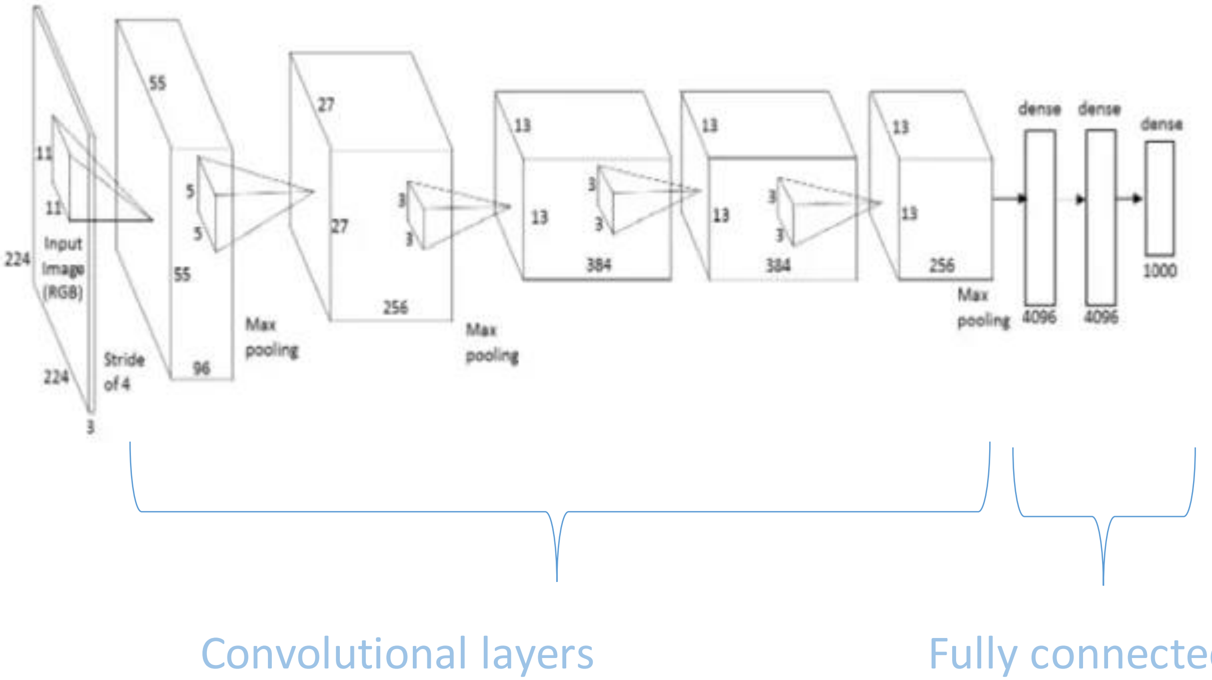


Image

Convolved Feature

https://stats.stackexchange.com/questions/199702/1d-convolution-in-neural-networks

Classic CNN Structure



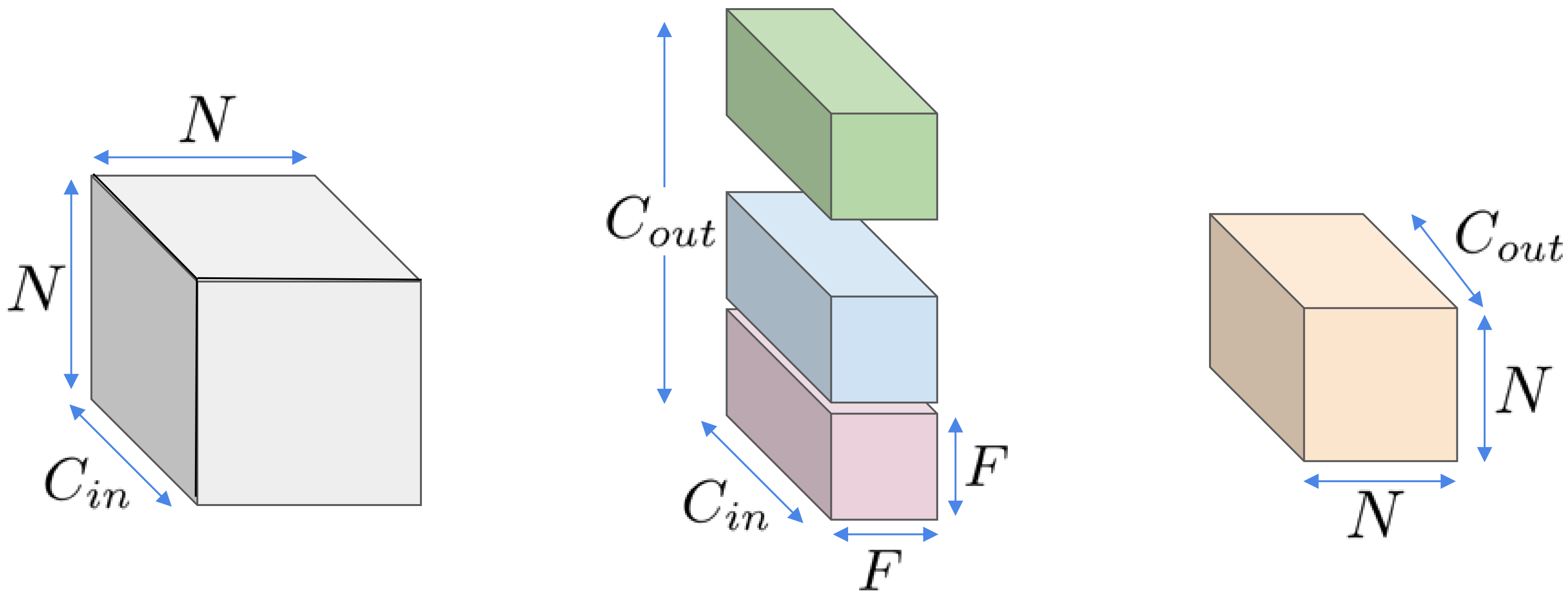
2D convolution with
Activation and
pooling / sub-sampling

Matrix multiplication &
activation

- Alex Net example
- Each convolutional layer has:
  - 2D convolution
  - Activation (eg. ReLU)
  - Pooling or sub-sampling

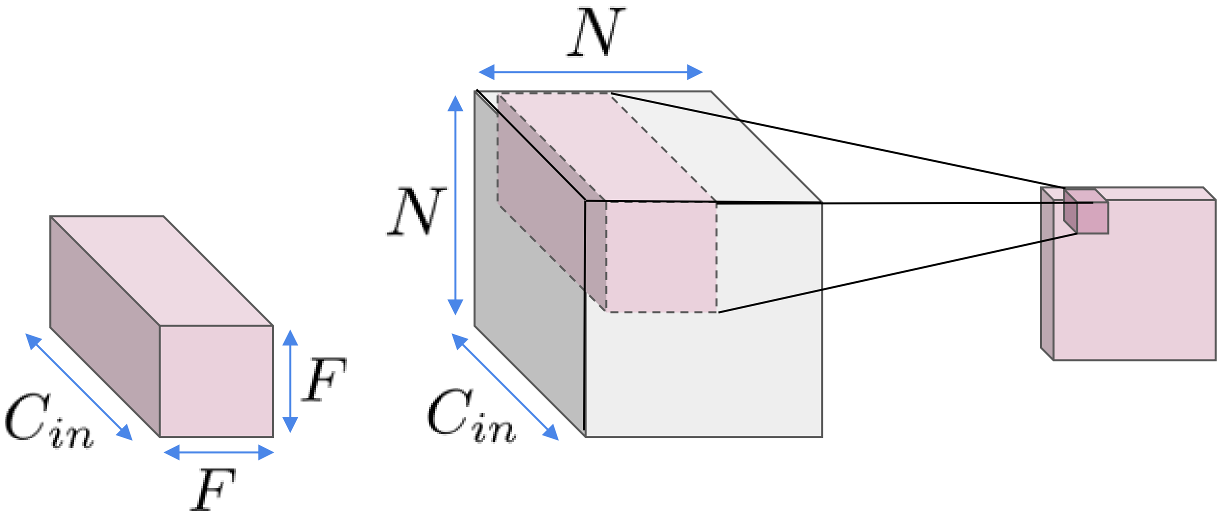
Convolutional Neural Networks

- but real-world DNNs also perform “convolution operations”
    - Inputs: 3-D tensor of activations
    - Weights: 4-D tensors representing filters
    - Outputs: 3-D tensors



input weights output

Convolutional Neural Networks



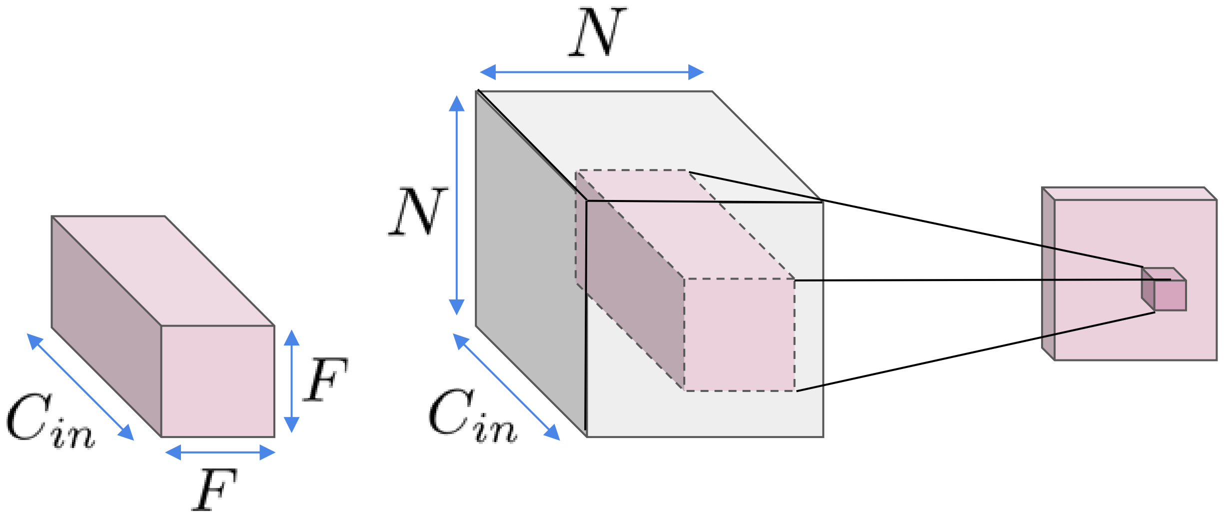
weights w

input a

output g

$$\boldsymbol{g}(1,1) = \sum_{r=0}^{C_{in}-1} \sum_{q=0}^{F-1} \sum_{p=0}^{F-1} \boldsymbol{w}(p,q,r).\boldsymbol{a}(1+p, 1+q, r)$$

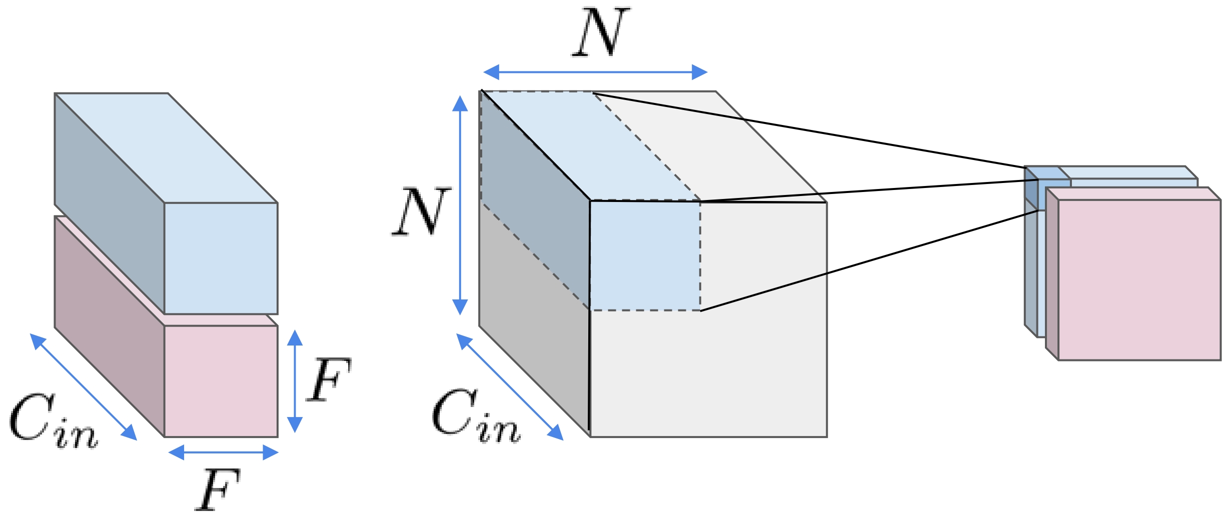
Convolutional Neural Networks



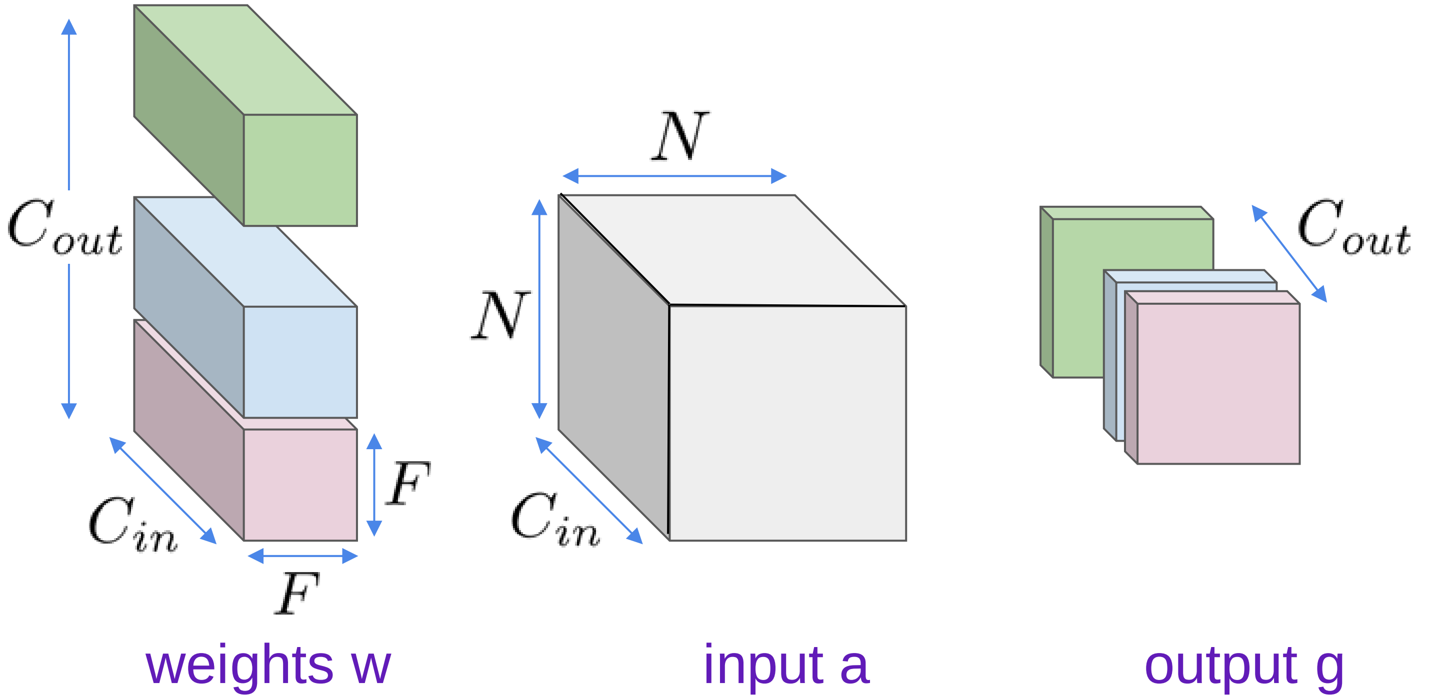
weights w input a output g

$$\boldsymbol{g}(x,y) = \sum_{r=0}^{C_{in}-1} \sum_{q=0}^{F-1} \sum_{p=0}^{F-1} \boldsymbol{w}(p,q,r).\boldsymbol{a}(x+p, y+q, r)$$

Convolutional Neural Networks



Convolutional Neural Networks



$$\boldsymbol{g}(x,y,z) = \sum_{r=0}^{C_{in}-1} \sum_{q=0}^{F-1} \sum_{p=0}^{F-1} \boldsymbol{w}(p,q,r,z).\boldsymbol{a}(x+p,y+q,r)$$

Activation and Sub-Sampling

- Convolution typically followed by activation and pooling
- Activation, typically ReLU
  - Zeros out portions of image
- Sub-sampling
  - Downsample output after activation
  - Different methods (striding, sub-sampling or max-pooling)
  - Output combines local features from adjacent regions
  - Creates more complex features over wider areas
- Details for sub-sampling not covered in this class
  - See web for more info

Convolution vs Fully Connected

- Convolution exploits translational invariance
  - Same features is scanned over whole image
- Greatly reduces number of parameters
- Example Consider first layer in LeNet
  - 32 x 32 image filtered by 6 channels 5 x 5 each
  - Creates 6 x 28 x 28 outputs (edges removed in convolution)
  - Fully connected would require 32 x 32 x 6 x 28 x 28 = 4.9 million parameters!
  - Convolutional layer requires only 6 x 5 x 5 = 125 parameters (plus bias terms)
- Reserve fully connected layers for last few layers.

Pre-Trained Networks

- State-of-the-art networks take enormous resources to train
  - Millions of parameters
  - Often days of training, clusters of GPUs
  - Extremely expensive
- Pre-trained networks in Keras
  - Load network architecture and weights
  - Models available for many state-of-the-art networks

Model	Size	Top-1 Accuracy	Top-5 Accuracy	Parameters	Depth
Xception	88 MB	0.790	0.945	22,910,480	126
VGG16	528 MB	0.715	0.901	138,357,544	23
VGG19	549 MB	0.727	0.910	143,667,240	26
ResNet50	99 MB	0.759	0.929	25,636,712	168
InceptionV3	92 MB	0.788	0.944	23,851,784	159
InceptionResNetV2	215 MB	0.804	0.953	55,873,736	572
MobileNet	17 MB	0.665	0.871	4,253,864	88

https://keras.io/applications/

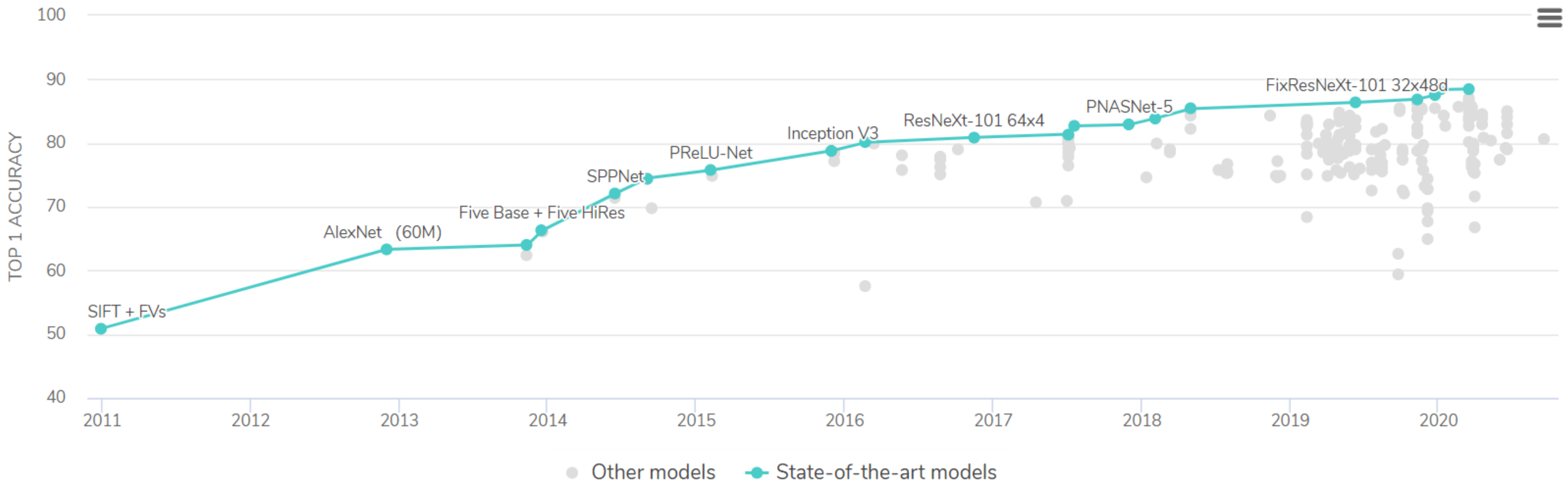
- Can be used for:
  - Making predictions
  - Building new, powerful networks (see lab)

State of the Art Today for Image Classification

https://kobiso.github.io/Computer-Vision-Leaderboard/imagenet.html

https://paperswithcode.com/sota/image-classification-on-imagenet

Image Classification on ImageNet



Deep Neural Networks for Face Recognition
A Brief Introduction

[Based on slides by T. Berg and Yang, Ranzato, Wolf, Taigman]

Motivation: General Goal

- Goal 1:
  - Given a picture of a person's face
  - Given a bag of possible names

    What's the name of the person in the picture?

- Goal 2:
  - Given two pictures of a person's face

    Are these of the same person?

Types of Face Recognition

- ‘Constrained’ – Mainly for traditional purposes
- ‘Unconstrained’ – General purpose

Constrained

NIST’s FR Vendor Test (FRVT) 2006

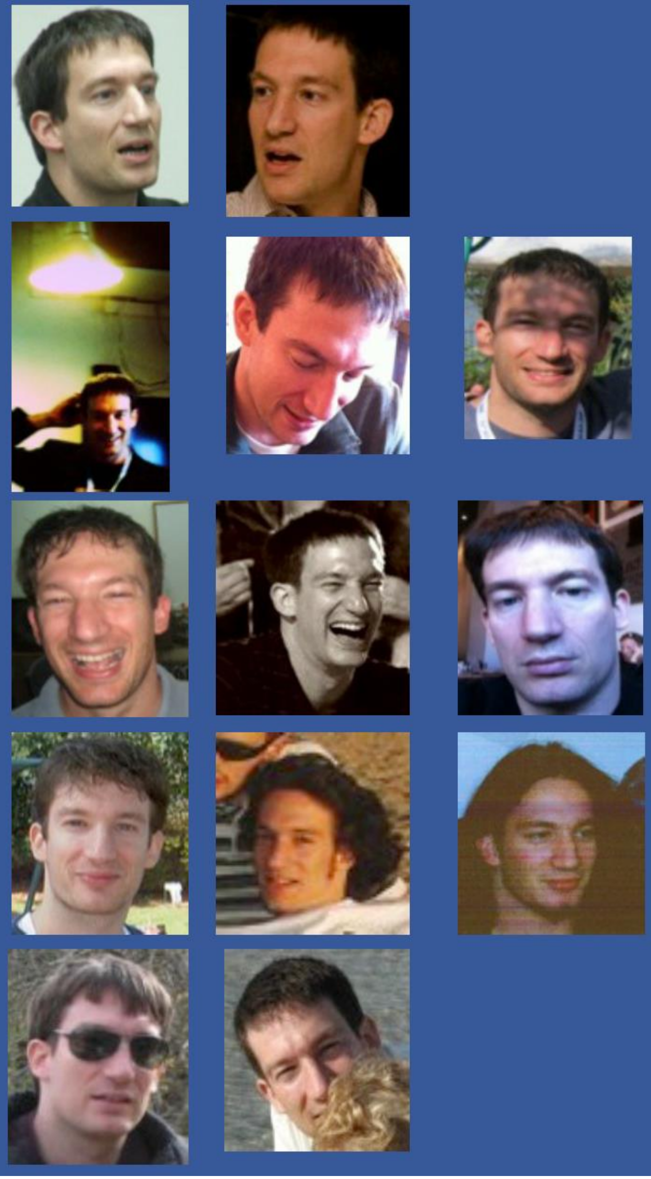
Unconstrained

In the wild

Challenges in Unconstrained Face Recognition

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

Probes for example



Gallery

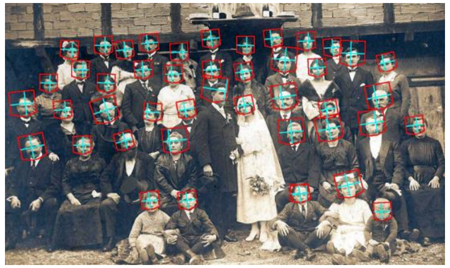
Unconstrained Face Recognition Era: The Labeled Faces in the Wild (LFW)

13,233 photos of 5,749 celebrities

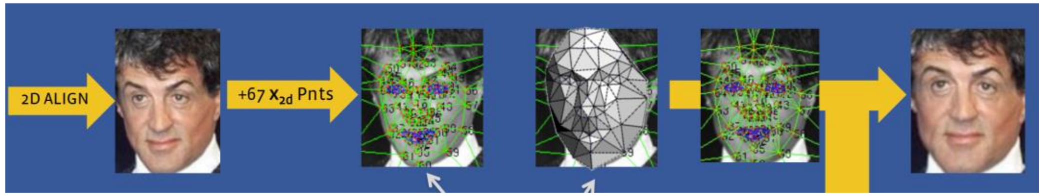
Labeled faces in the wild: A database for studying face recognition in unconstrained environments, Huang, Jain, Learned-Miller, ECCVW, 2008

Overview of Methods

- Face Detection
  - Localize the face



- Face Alignment
  - Factor out 3D transformation



- Feature Extraction
  - Find compact representation
- **Classification**
  - Answer the question

DeepFace [Taigman et al., CVPR 2014]

https://www.cv-foundation.org/openaccess/content_cvpr_2014/papers/Taigman_DeepFace_Closing_the_2014_CVPR_paper.pdf

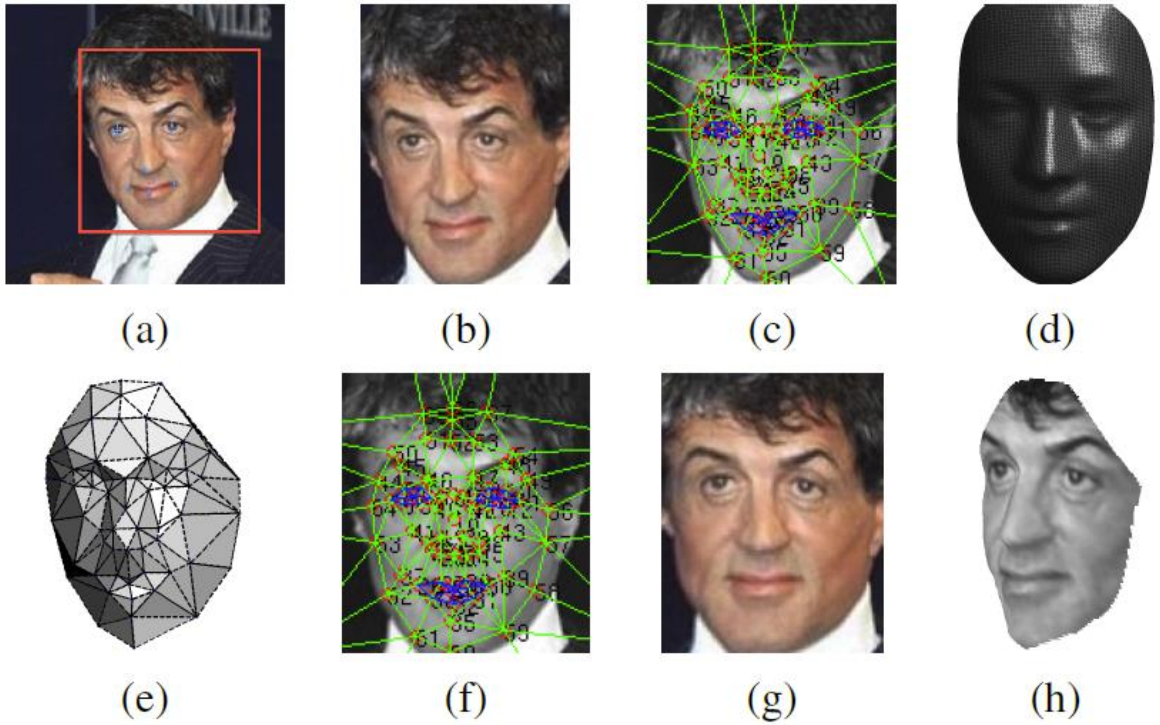
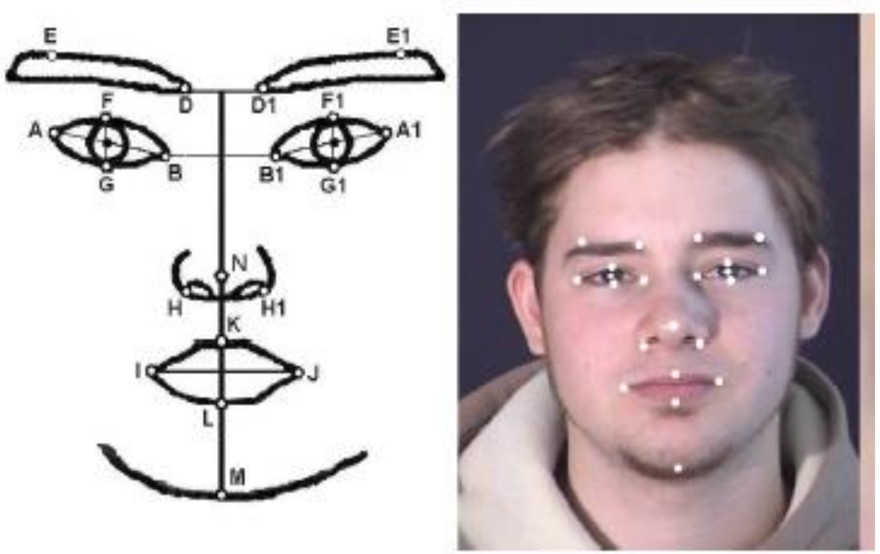


Figure 1. **Alignment pipeline.** (a) The detected face, with 6 initial fiducial points. (b) The induced 2D-aligned crop. (c) 67 fiducial points on the 2D-aligned crop with their corresponding Delaunay triangulation, we added triangles on the contour to avoid discontinuities. (d) The reference 3D shape transformed to the 2D-aligned crop image-plane. (e) Triangle visibility w.r.t. to the fitted 3D-2D camera; darker triangles are less visible. (f) The 67 fiducial points induced by the 3D model that are used to direct the piece-wise affine warpping. (g) The final frontalized crop. (h) A new view generated by the 3D model (not used in this paper).

Detect "fiducial points" in a face image that serve as "landmarks," for example, the corner of the eyes, lips, and so on.

- *Using Support Vector Regression based on "image descriptors"*



https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=5539996&casa_token=CtkEBevZvQIAAAAA:-oBCv5rOVBvH9Oif5Nv5bY2qaV8KQgDN-yvkukS-8zsk83DzsJ2UA5y9fQIiUW3oO6_78Lji_w&tag=1

DeepFace CNN

"Max Pooling" layer reduces sensitivity to local registration errors

No max pooling in subsequent layers because the relative position of global features matter

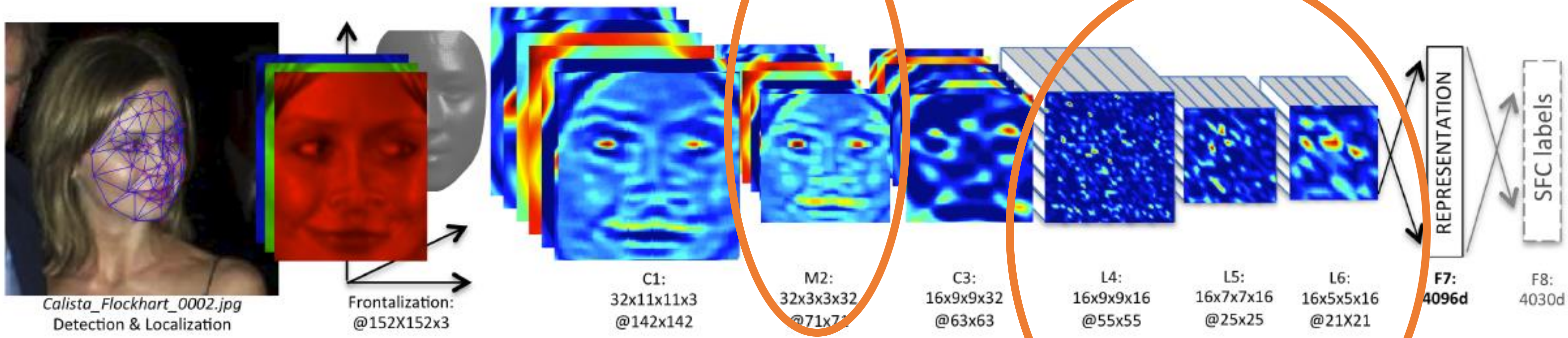
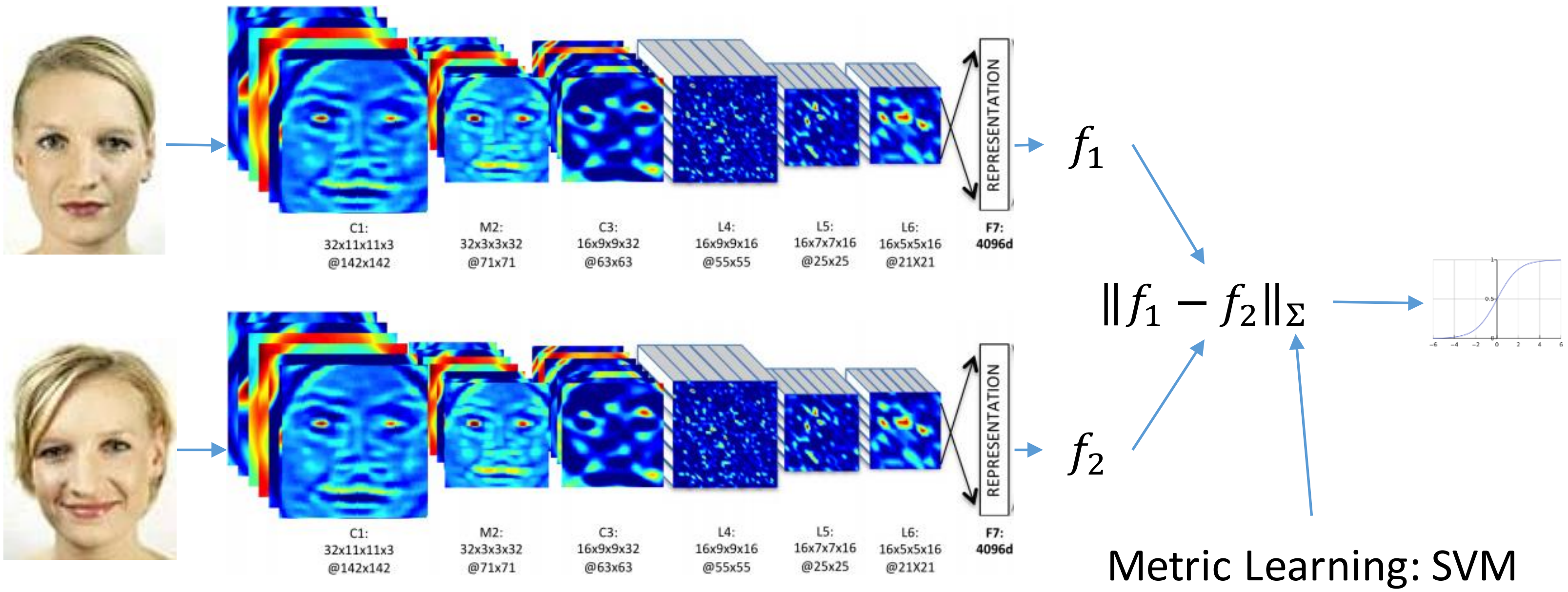


Figure 2. **Outline of the *DeepFace* architecture.** A front-end of a single convolution-pooling-convolution filtering on the rectified input, followed by three locally-connected layers and two fully-connected layers. Colors illustrate feature maps produced at each layer. The net includes more than 120 million parameters, where more than 95% come from the local and fully connected layers.

"Locally connected" convolutional layers, because of high spatial variation in image features

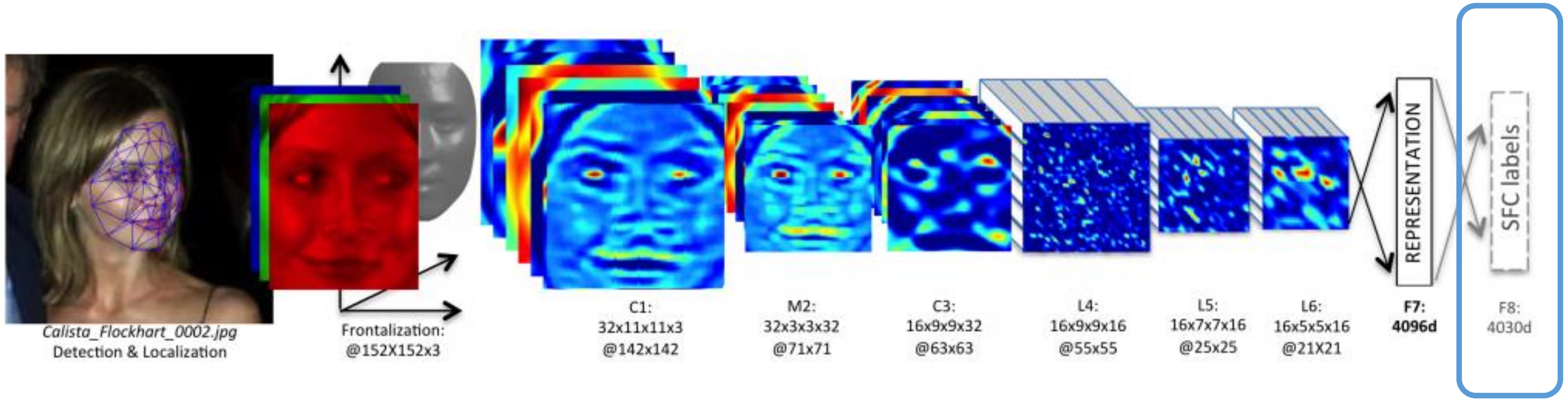
Classifier

- Same Person Task:



Classifier

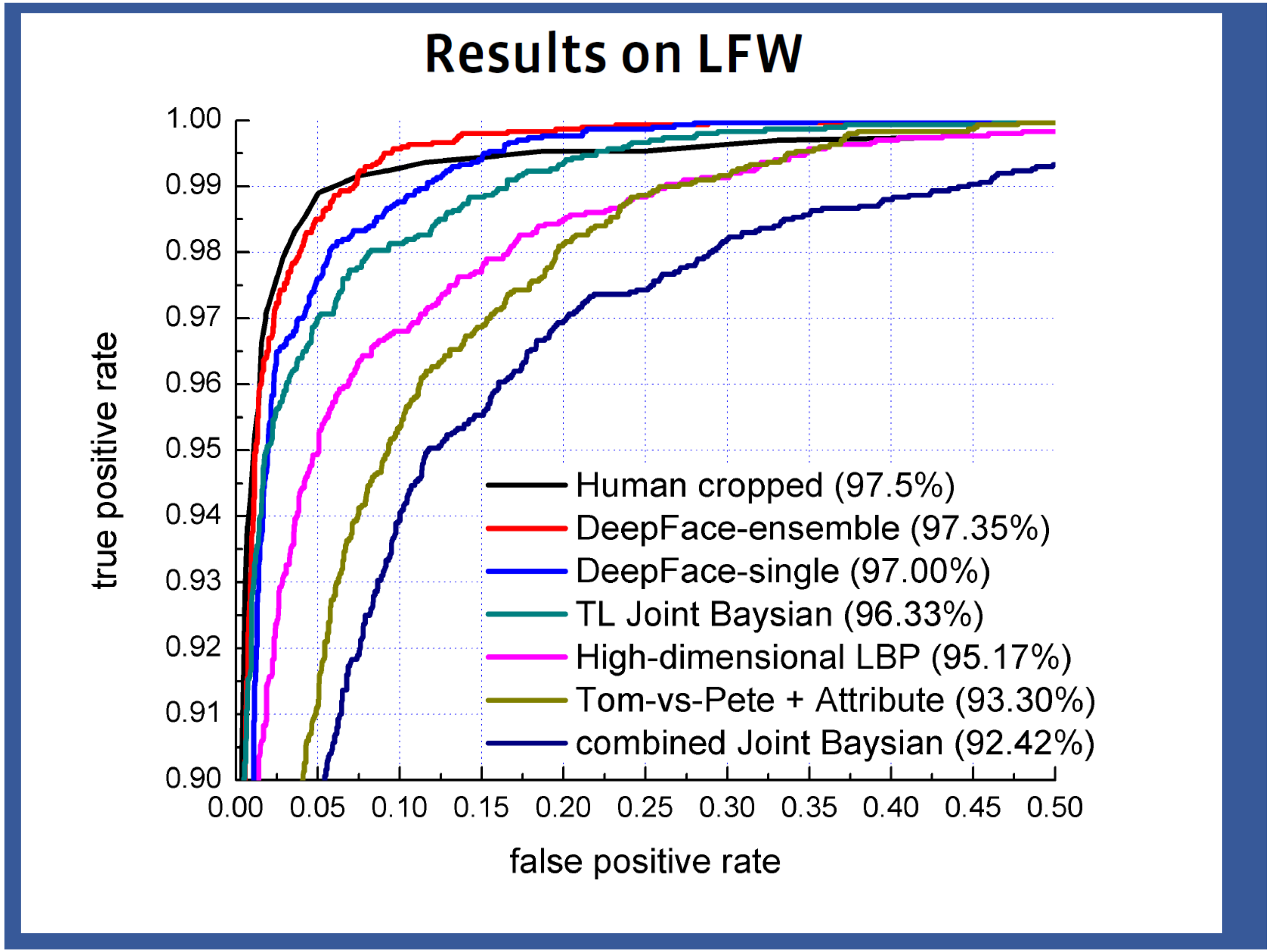
- Name of the Person Task:



Datasets

- The SFC Dataset
  – From Facebook
  – 800-1200 each, 4030 people, 4.4M in all
- The LFW Dataset
  – 13323 photos of 5749 celebrities

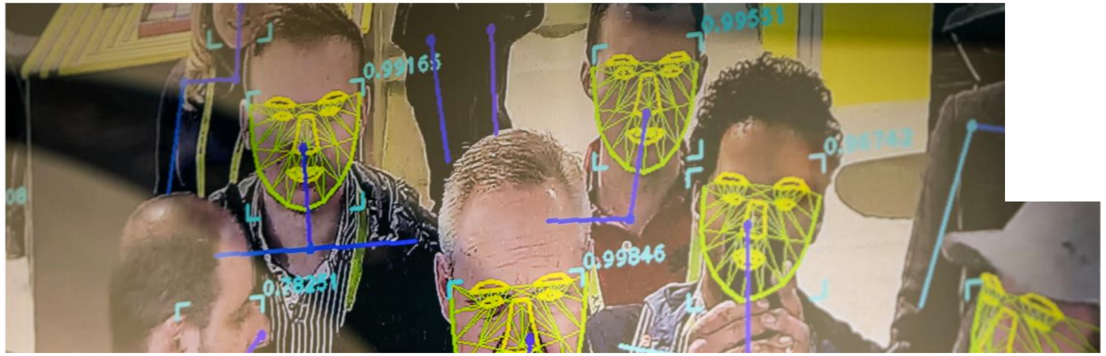
Comparison



Results on LFW
true positive rate
false positive rate
1.00
0.99
0.98
0.97
0.96
0.95
0.94
0.93
0.92
0.91
0.90
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
Human cropped (97.5%)
DeepFace-ensemble (97.35%)
DeepFace-single (97.00%)
TL Joint Baysian (96.33%)
High-dimensional LBP (95.17%)
Tom-vs-Pete + Attribute (93.30%)
combined Joint Baysian (92.42%)

Concerns Regarding Face Recognition

San Francisco Bans Facial Recognition Technology



CNN BUSINESS Markets Tech Media Success Perspectives Videos LIVE TV

UNHACKABLE

Portland passes broadest facial recognition ban in the US

By Rachel Metz, CNN Business

Updated 8:06 PM ET, Wed September 9, 2020

https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html

MIT Technology Review

Artificial intelligence Jun 26

A new US bill would ban the police use of facial recognition

These legislations reflect emerging concerns about compromise of individual privacy and enabling of a "surveillance state" giving broad powers to government and law enforcement

Concerns About Bias

The New York Times

PLAY THE CROSSWORD

Many Facial-Recognition Systems Are Biased, Says U.S. Study

Algorithms falsely identified African-American and Asian faces 10 to 100 times more than Caucasian faces, researchers for the National Institute of Standards and Technology found.

Introduction

- State-of-the-art deep neural networks have achieved near human-level performance in image classification tasks.
- However, deep neural networks have been shown to be susceptible to adversarial input attacks.
- A small, intentionally chosen perturbation added to a correctly classified image can mislead the classifier to output a totally different label, even though the perturbation is small enough that it appears imperceptible to humans.

Motivation Example

Figure 1: Clean and adversarial images with different prediction labels.

Attack's Types

Adversarial input attacks can be broadly classified into two types, one is non-targeted attack and the other is targeted attack.

- Non-targeted attack:
  Aiming to fool the neural network and output a label different than the original one.
- Targeted attack:
  Intentionally misleading the network to output a specific label designed by the attacker.
- E.g. a face recognition system for security entrance control:
  - Non-targeted attacks could lead to denial of legal access.
  - Targeted attacks bring the jeopardy of illegal entrance.

Adversary's Objective

Given an image x with a classification label $y = \text{classifier}(x)$, where classifier is the function of the neural network. The attacker aims to:

- find an image x' whose classification label is y', such that $y' = \text{classifier}(x') \neq y$.
- ensure $\|x' - x\| \leq \delta$, where δ is an upper bound of the distortion from x to x'.

Fast Gradient Sign Methods

- Non-targeted and targeted FGS methods are expressed as in Equation 2 and Equation 3:

$$x' \leftarrow \text{clip}(x + \epsilon \text{sign}(\nabla \ell_{F,y^*}(x))) \tag{2}$$

$$x' \leftarrow \text{clip}(x - \epsilon \text{sign}(\nabla \ell_{F,y'}(x))) \tag{3}$$

Here ϵ is a small constraint scalar, ℓ refers to the loss function and clip(x) ensures each pixel value falls in the setting range.

Iterative fast gradient sign methods:

- FGS methods have been extended to iterative versions, naming IFGS, that perturb each pixel with a small amount for multiple times.
- IFGS methods for non-targeted and targeted attacks are shown in Equation 4 and Equation 5:

$$x'_0 = x, \quad x'_{N+1} \leftarrow \text{clip}_\epsilon(x'_N + \alpha \text{sign}(\nabla \ell_{F,y^*}(x))) \tag{4}$$

$$x'_0 = x, \quad x'_{N+1} \leftarrow \text{clip}_\epsilon(x'_N - \alpha \text{sign}(\nabla \ell_{F,y'}(x))) \tag{5}$$

- IFGS methods are capable of generating adversarial inputs with smaller distortion when compared to basic FGS methods.

JSMA

- Jacobian-based saliency map attack (JSMA) modifies pixel pairs with the highest influence on the output of the network with unit step and increase the prediction probability of the target label with multiple iterations.
- Searching satisfied pixel pairs in JSMA brings up the computation cost dramatically in every iteration as the image dimensions grow.

Optimization Based Approaches

- CW method is capable of making much smaller perturbation than previous attacks to fool the network.
- When the pixel values are quantized to form valid inputs, the newly generated images often fail to mislead the network with the specific target label.
- A greedy search on the lattice defined by the discrete neighbor integers is essential after optimization.

Attacks Comparison

Table 1: Attacks Comparison

Attacks	Pros	Cons
FGS	fastest speed low computation cost	large perturbation
IFGS	smaller perturbation than FGS	slower than FGS
JSMA	small perturbation natively generate valid images	high computation cost unfeasible for ImageNet
CW	minimum perturbation	slowest speed high computation cost

Attacks Comparison

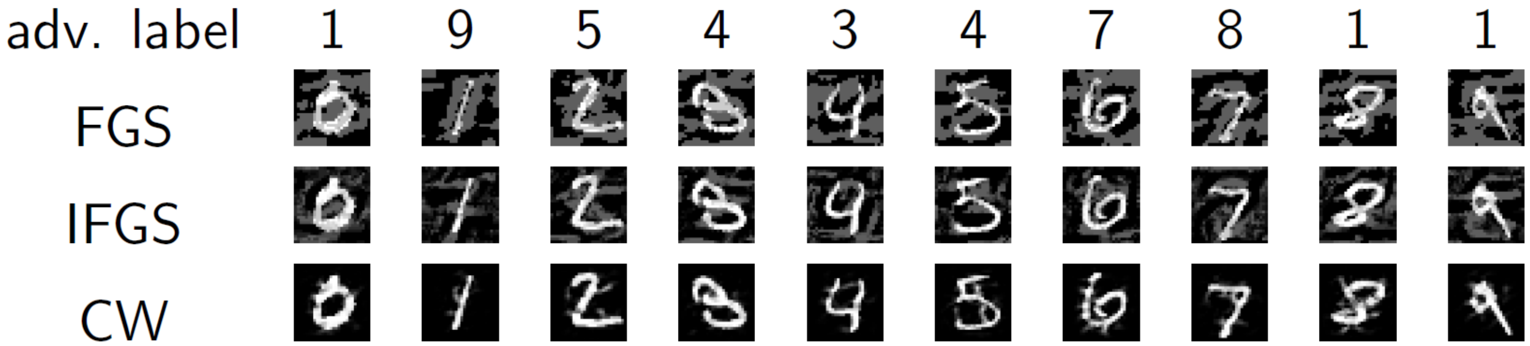


Figure 3: MNIST adversarial images generated by FGS, IFGS, CW, and their prediction labels.

Adversarial Perturbations Applied to FaceRec

- Attacks implemented on Face Recognition DNN implemented by Parkhi et al.
  - “Dodging” = Untargeted attack
  - “Impersonation” = Targeted attack

Figure 2: A dodging attack by perturbing an entire face. Left: an original image of actress Eva Longoria (by Richard Sandoval / CC BY-SA / cropped from https://goo.gl/7QUvRq). Middle: A perturbed image for dodging. Right: The applied perturbation, after multiplying the absolute value of pixels’ channels ×20.

Figure 3: An impersonation using frames. Left: Actress Reese Witherspoon (by Eva Rinaldi / CC BY-SA / cropped from https://goo.gl/a2sCdc). Image classified correctly with probability 1. Middle: Perturbing frames to impersonate (actor) Russel Crowe. Right: The target (by Eva Rinaldi / CC BY-SA / cropped from https://goo.gl/AO7QYu).

(a) (b) (c) (d)

Figure 4: Examples of successful impersonation and dodging attacks. Fig. (a) shows S_A (top) and S_B (bottom) dodging against DNN_B. Fig. (b)–(d) show impersonations. Impersonators carrying out the attack are shown in the top row and corresponding impersonation targets in the bottom row. Fig. (b) shows S_A impersonating Milla Jovovich (by Georges Biard / CC BY-SA / cropped from https://goo.gl/GlsWlC); (c) S_B impersonating S_C; and (d) S_C impersonating Carson Daly (by Anthony Quintano / CC BY / cropped from https://goo.gl/VfnDct).

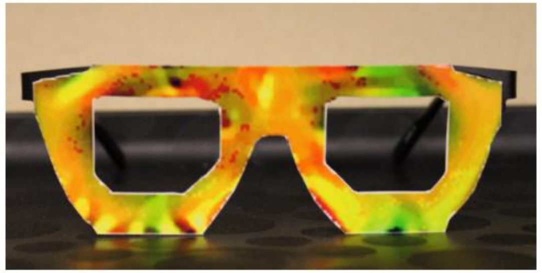


Figure 5: The eyeglass frames used by S_C for dodging recognition against DNN_B.