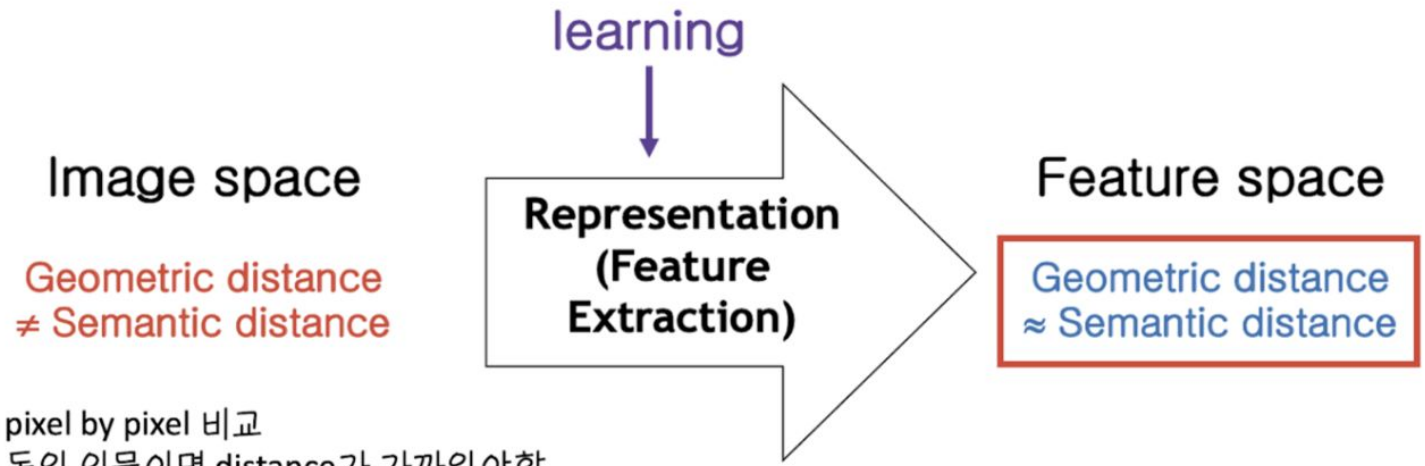
FaceNet: A Unified Embedding for Face Recognition and Clustering

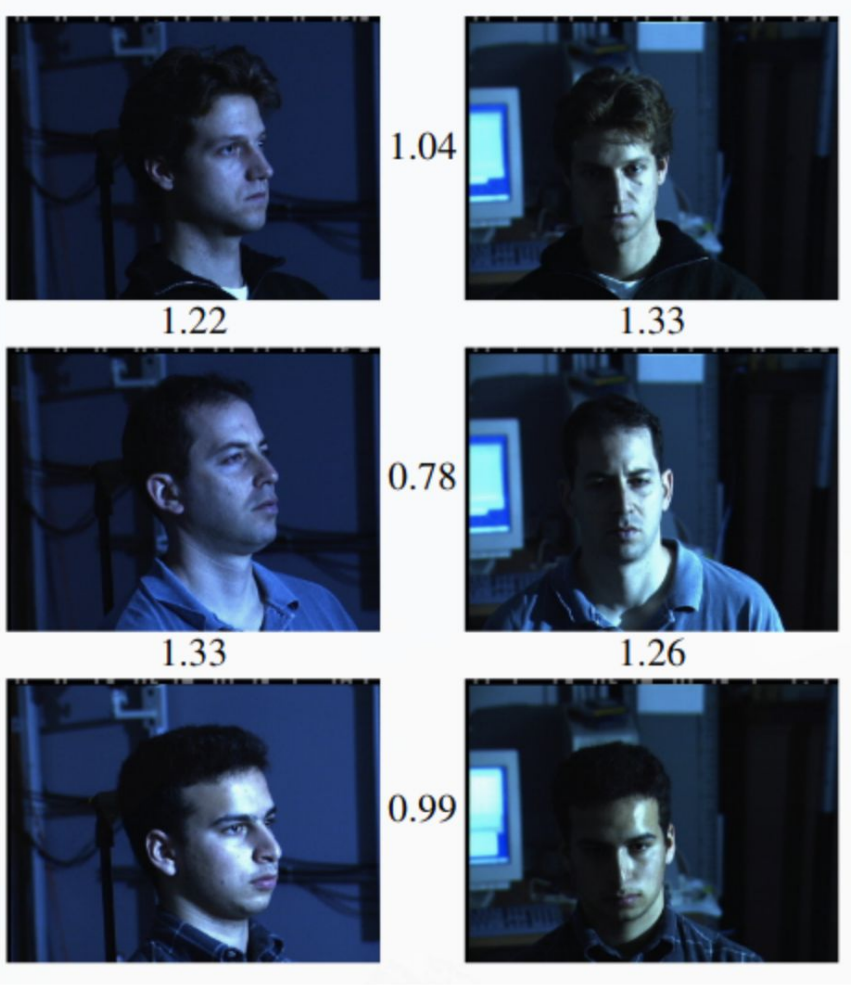
(Florian Schroff, Dmitry Kalenichenko, James Philbin, 2015)

Background - Metric Learning

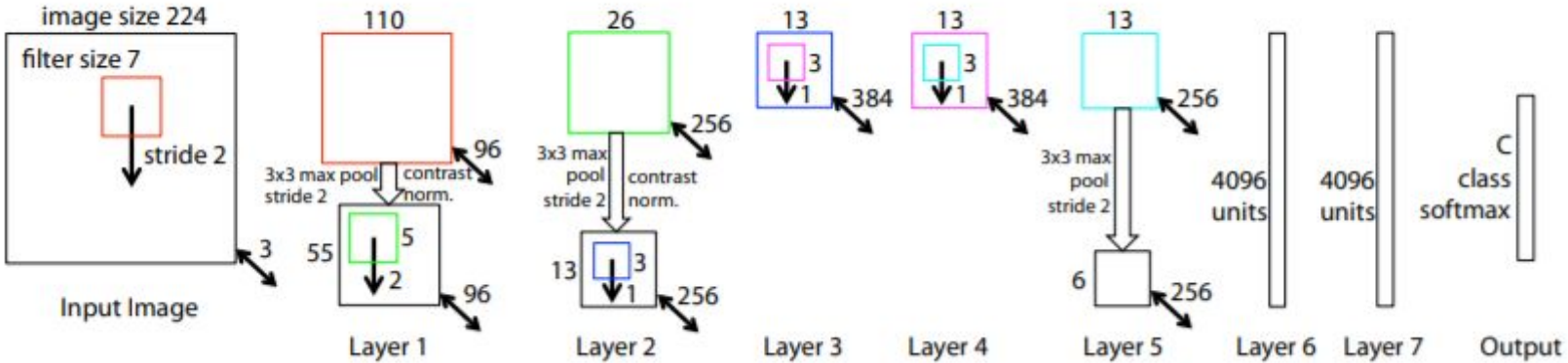


pixel by pixel 비교
동일 인물이면 distance가 가까워야함
→ but, 배경, 옷등에 의해 달라진다

Abstract & Introduction



Related Work - based model
- Visualizing and understandingconvolutional networks



Visualizing and Understanding Convolutional Networks, Matthew D. Zeiler and Rob Fergus
Dept. of Computer Science, New York University, USA, 2014

Related Work - based model - Inception model
- Going deeper with convolutions

Going deeper with convolutions, Szegedy, Dept of Computer Vision and Pattern Recognition, Cornell University, 2014

Method - Triplet Loss

Method - Triplet Loss

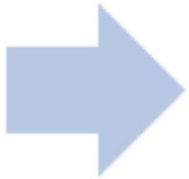
$$\|f(x_i^a) - f(x_i^p)\|_2^2 + \alpha < \|f(x_i^a) - f(x_i^n)\|_2^2$$

$$\text{Loss} = \sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]$$

Method - Triplet Selection

$$\text{argmax}_{x_i^p} \|f(x_i^a) - f(x_i^p)\|_2^2$$ Hard positive

$$\text{argmin}_{x_i^n} \|f(x_i^a) - f(x_i^n)\|_2^2$$ Hard negative



계산량 ⬆
오버피팅 문제

Method - Deep Convolutional Networks - NN1

layer	size-in	size-out	kernel	param	FLPS
conv1	220×220×3	110×110×64	7×7×3, 2	9K	115M
pool1	110×110×64	55×55×64	3×3×64, 2	0	
rnorm1	55×55×64	55×55×64		0	
conv2a	55×55×64	55×55×64	1×1×64, 1	4K	13M
conv2	55×55×64	55×55×192	3×3×64, 1	111K	335M
rnorm2	55×55×192	55×55×192		0	
pool2	55×55×192	28×28×192	3×3×192, 2	0	
conv3a	28×28×192	28×28×192	1×1×192, 1	37K	29M
conv3	28×28×192	28×28×384	3×3×192, 1	664K	521M
pool3	28×28×384	14×14×384	3×3×384, 2	0	
conv4a	14×14×384	14×14×384	1×1×384, 1	148K	29M
conv4	14×14×384	14×14×256	3×3×384, 1	885K	173M
conv5a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv5	14×14×256	14×14×256	3×3×256, 1	590K	116M
conv6a	14×14×256	14×14×256	1×1×256, 1	66K	13M
conv6	14×14×256	14×14×256	3×3×256, 1	590K	116M
pool4	14×14×256	7×7×256	3×3×256, 2	0	
concat	7×7×256	7×7×256		0	
fc1	7×7×256	1×32×128	maxout p=2	103M	103M
fc2	1×32×128	1×32×128	maxout p=2	34M	34M
fc7128	1×32×128	1×1×128		524K	0.5M
L2	1×1×128	1×1×128		0	
total				140M	1.6B

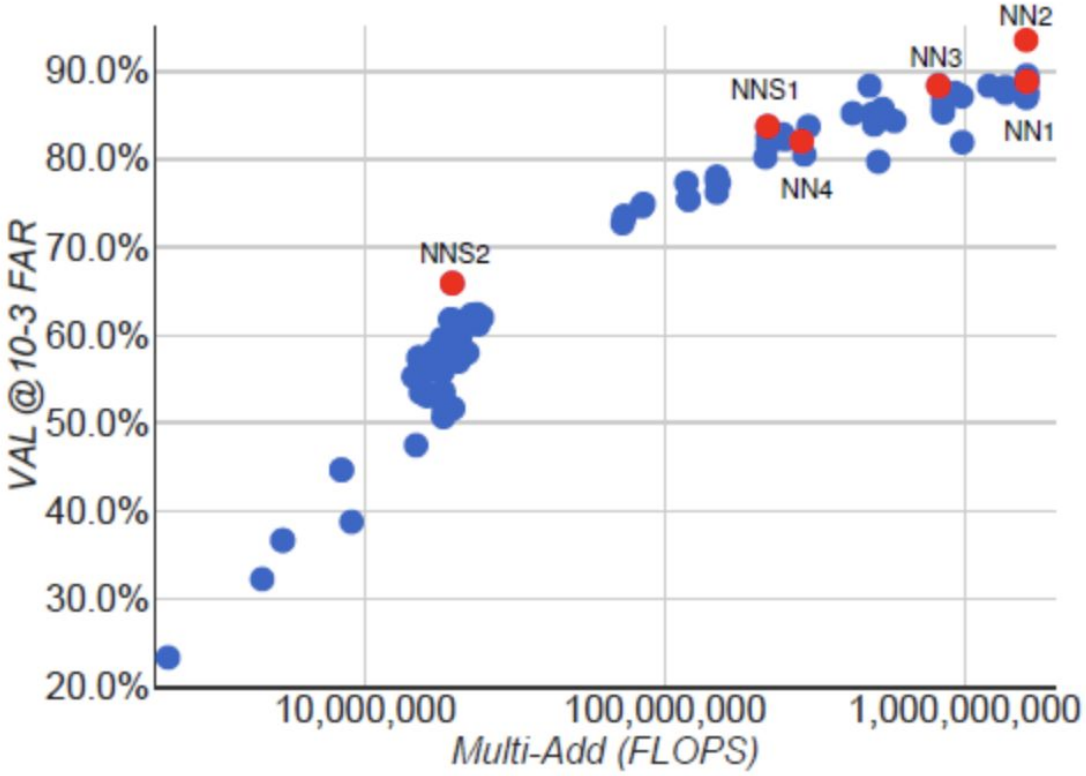
Method - Deep Convolutional Networks - NN2

type	output size	depth	#1×1	#3×3 reduce	#3×3	#5×5 reduce	#5×5	pool proj (p)	params	FLOPS
conv1 (7×7×3, 2)	112×112×64	1							9K	119M
max pool + norm	56×56×64	0						m 3×3, 2		
inception (2)	56×56×192	2		64	192				115K	360M
norm + max pool	28×28×192	0						m 3×3, 2		
inception (3a)	28×28×256	2	64	96	128	16	32	m, 32p	164K	128M
inception (3b)	28×28×320	2	64	96	128	32	64	L_2, 64p	228K	179M
inception (3c)	14×14×640	2	0	128	256,2	32	64,2	m 3×3,2	398K	108M
inception (4a)	14×14×640	2	256	96	192	32	64	L_2, 128p	545K	107M
inception (4b)	14×14×640	2	224	112	224	32	64	L_2, 128p	595K	117M
inception (4c)	14×14×640	2	192	128	256	32	64	L_2, 128p	654K	128M
inception (4d)	14×14×640	2	160	144	288	32	64	L_2, 128p	722K	142M
inception (4e)	7×7×1024	2	0	160	256,2	64	128,2	m 3×3,2	717K	56M
inception (5a)	7×7×1024	2	384	192	384	48	128	L_2, 128p	1.6M	78M
inception (5b)	7×7×1024	2	384	192	384	48	128	m, 128p	1.6M	78M
avg pool	1×1×1024	0								
fully conn	1×1×128	1							131K	0.1M
L2 normalization	1×1×128	0								
total									7.5M	1.6B

Method - Deep Convolutional Networks

architecture	VAL
NN1 (Zeiler&Fergus 220×220)	$87.9\% \pm 1.9$
NN2 (Inception 224×224)	$89.4\% \pm 1.6$
NN3 (Inception 160×160)	$88.3\% \pm 1.7$
NN4 (Inception 96×96)	$82.0\% \pm 2.3$
NNS1 (mini Inception 165×165)	$82.4\% \pm 2.4$
NNS2 (tiny Inception 140×116)	$51.9\% \pm 2.9$

Method - Deep Convolutional Networks



Datasets and Evaluation

$$\mathrm{TA}(d) = \{(i,j) \in \mathcal{P}_{\mathrm{same}} | D(x_i, x_j) \leq d\}$$

$$\mathrm{FA}(d) = \{(i,j) \in \mathcal{P}_{\mathrm{diff}} | D(x_i, x_j) \leq d\}$$

$$\mathrm{VAL}(d) = \frac{|\mathrm{TA}(d)|}{|\mathcal{P}_{\mathrm{same}}|}, \mathrm{FAR}(d) = \frac{|\mathrm{FA(d)}|}{|\mathcal{P}_{\mathrm{diff}}|}$$

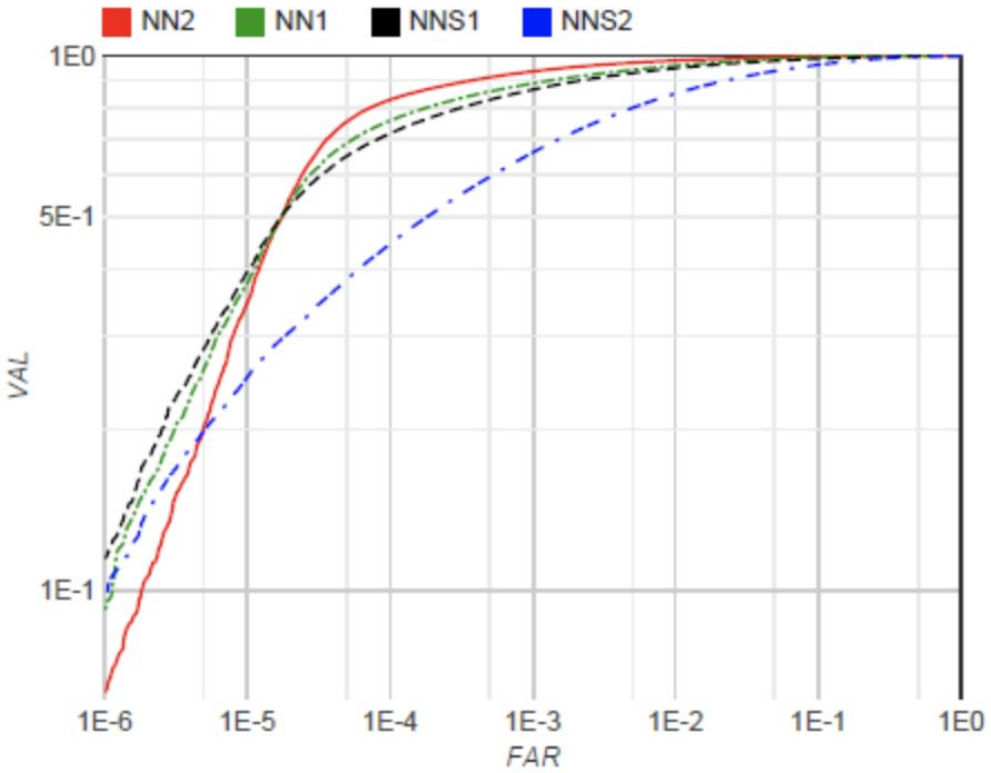
Datasets and Evaluation

- **Hold-out Data Set**
- **Personal Photos**
- **Academic Datasets**

Experiments - Computation Accuracy Trade-off

정확도 , FLOPS : Trade-off 관계

Experiments - Effect of CNN Model



Experiments - Sensitivity to Image Quality

jpeg q	val-rate
10	67.3%
20	81.4%
30	83.9%
50	85.5%
70	86.1%
90	86.5%

#pixels	val-rate
1,600	37.8%
6,400	79.5%
14,400	84.5%
25,600	85.7%
65,536	86.4%

Experiments - Embedding Dimensionality

#dims	VAL
64	$86.8\% \pm 1.7$
128	$87.9\% \pm 1.9$
256	$87.7\% \pm 1.9$
512	$85.6\% \pm 2.0$

Experiments - Amount of Training Data

#training images	VAL
2,600,000	76.3%
26,000,000	85.1%
52,000,000	85.1%
260,000,000	86.2%

Experiments - Performance on LFW

Experiments - Face Clustering