Insight
Centre for Data Analytics

DEEP
LEARNING
WORKSHOP

Dublin City University
28-29 April 2017

DCU

Learning where to look: focus and attention in deep vision

Kevin McGuinness
kevin.mcguinness@dcu.ie

Research Fellow
Insight Centre for Data Analytics
Dublin City University

Overview

Visual attention models and their applications

Deep vision for **medical image analysis**

Deep **crowd analysis**

Interactive deep vision: image segmentation

Visual Attention Models and their Applications

The importance of visual attention

The importance of visual attention

The importance of visual attention

The importance of visual attention

Why don’t we see the changes?

We don’t really see the whole image

We only focus on small specific regions: the **salient** parts

Human beings reliably attend to the same regions of images when shown

What we perceive

Where we look

What we actually see

Can we predict where humans will look?

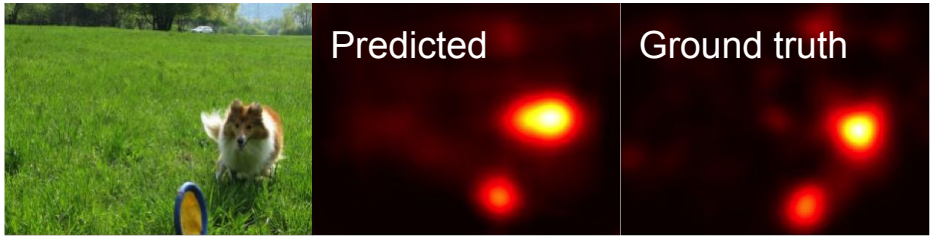
Yes! Computational models of visual saliency

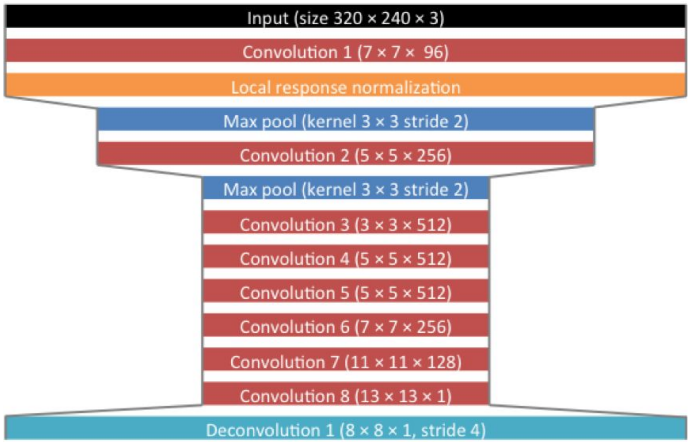
Why might this be useful?

SalNet: deep visual saliency model

Predict map of visual attention from image pixels (find the parts of the image that stand out)

- Feedforward 8 layer “fully convolutional” architecture
- Transfer learning in bottom 3 layers from pretrained VGG-M model on ImageNet
- Trained on SALICON dataset (simulated crowdsourced attention dataset using mouse and artificial foveation)
- Top-5 in MIT 300 saliency benchmark http://saliency.mit.edu/results_mit300.html





Pan, McGuinness, et al. **Shallow and Deep Convolutional Networks for Saliency Prediction**, CVPR 2016 http://arxiv.org/abs/1603.00845


Image
Ground truth
Prediction

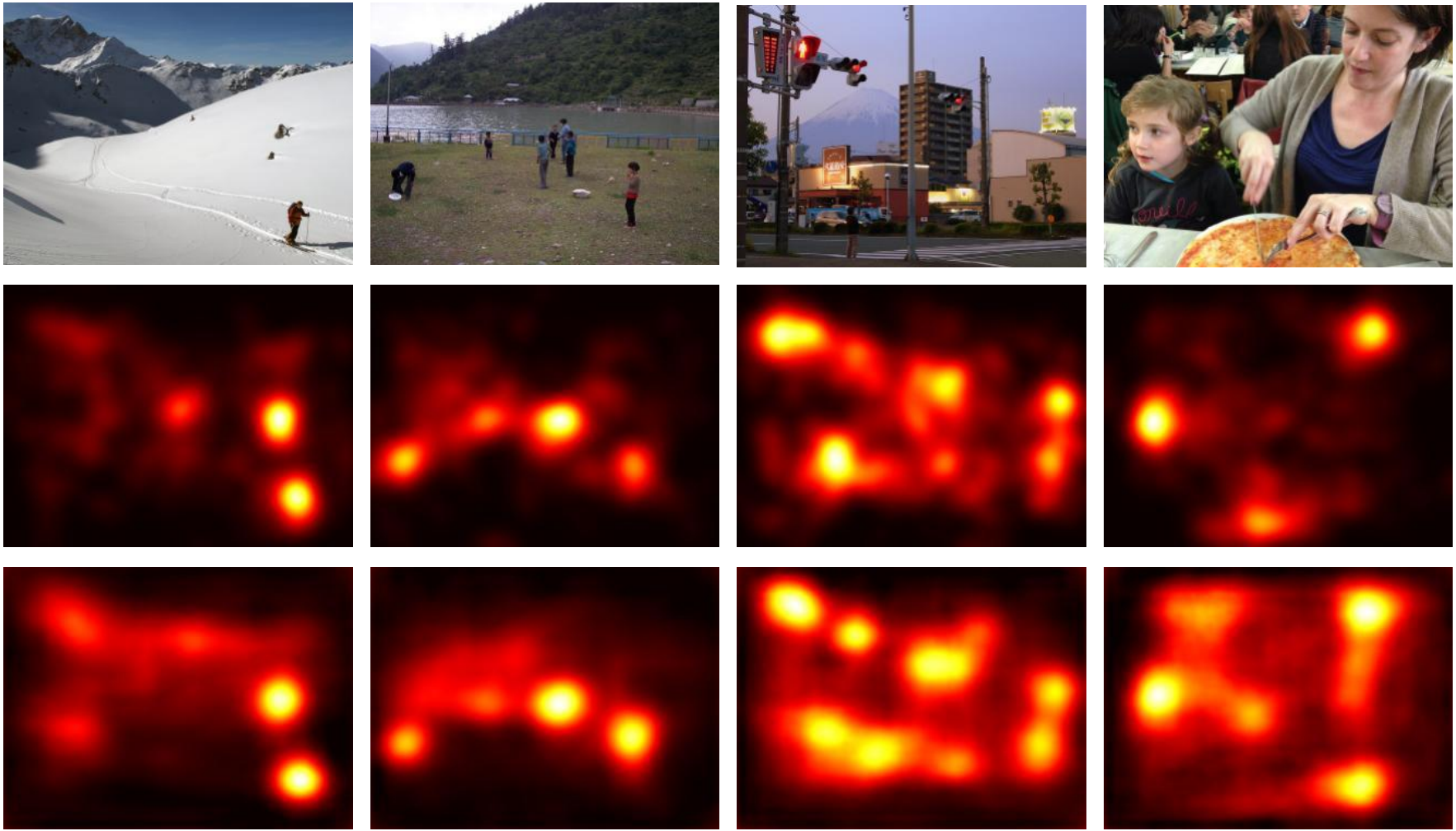


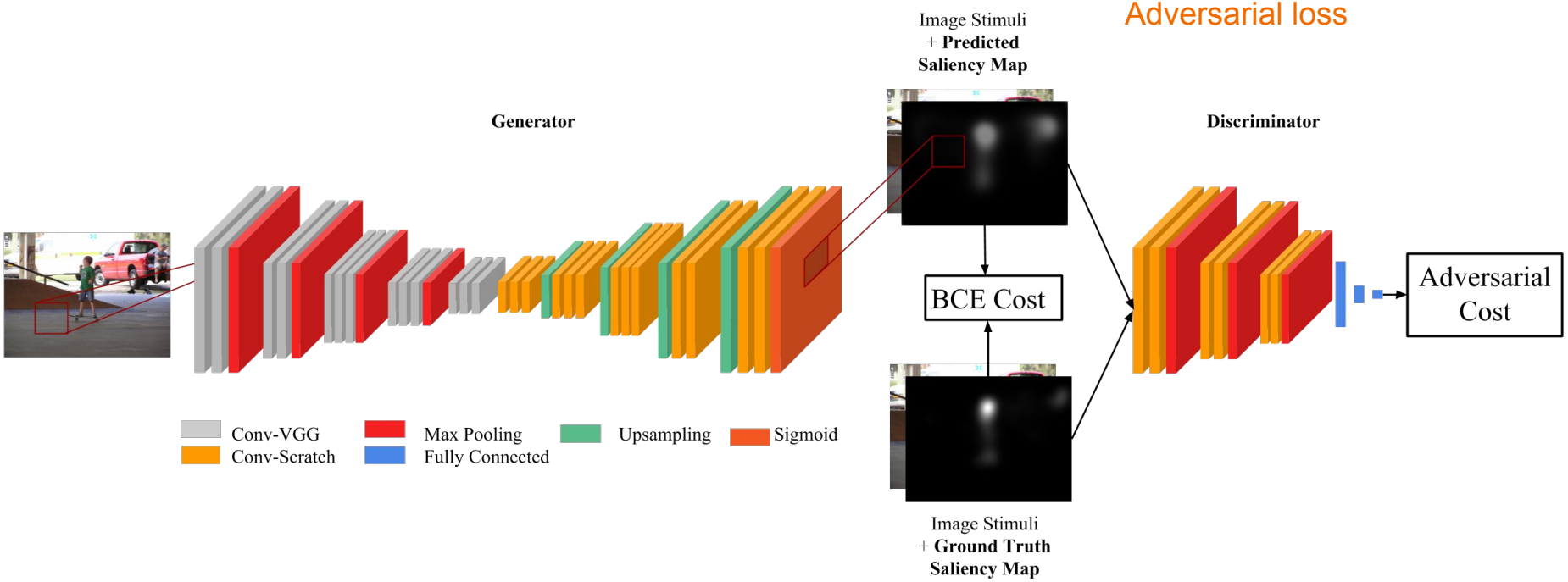
Image
Ground truth
Prediction

SalGAN

Data loss

$$\alpha \cdot \mathcal{L}_{BCE} - \log D(I, \hat{S}),$$

Adversarial loss



Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol and Xavier Giro-i-Nieto. "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks." arXiv. 2017.

SalNet and SalGAN benchmarks

SALICON (test)	AUC-J ↑	Sim ↑	EMD ↓	AUC-B ↑	sAUC ↑	CC ↑	NSS ↑	KL ↓
DSCLRCN [24](*)	-	-	-	0.884	0.776	0.831	3.157	-
SalGAN	-	-	-	**0.884**	**0.772**	**0.781**	**2.459**	-
ML-NET [5]	-	-	-	(0.866)	(0.768)	(0.743)	2.789	-
SalNet [25]	-	-	-	(0.858)	(0.724)	(0.609)	(1.859)	-
MIT300	AUC-J ↑	Sim ↑	EMD ↓	AUC-B ↑	sAUC ↑	CC ↑	NSS ↑	KL ↓
Humans	0.92	1.00	0.00	0.88	0.81	1.0	3.29	0.00
Deep Gaze II [21](*)	0.88	(0.46)	(3.98)	0.86	0.72	(0.52)	(1.29)	(0.96)
DSCLRCN [24](*)	0.87	0.68	2.17	(0.79)	0.72	0.80	2.35	0.95
DeepFix [17](*)	0.87	0.67	2.04	(0.80)	(0.71)	0.78	2.26	0.63
SALICON [9]	0.87	(0.60)	(2.62)	0.85	0.74	0.74	2.12	0.54
SalGAN	**0.86**	**0.63**	**2.29**	**0.81**	**0.72**	**0.73**	**2.04**	**1.07**
PDP [11]	(0.85)	(0.60)	(2.58)	(0.80)	0.73	(0.70)	2.05	0.92
ML-NET [5]	(0.85)	(0.59)	(2.63)	(0.75)	(0.70)	(0.67)	2.05	(1.10)
Deep Gaze I [19]	(0.84)	(0.39)	(4.97)	0.83	(0.66)	(0.48)	(1.22)	(1.23)
iSEEL [29](*)	(0.84)	(0.57)	(2.72)	0.81	(0.68)	(0.65)	(1.78)	0.65
SalNet [25]	(0.83)	(0.52)	(3.31)	0.82	(0.69)	(0.58)	(1.51)	0.81
BMS [31]	(0.83)	(0.51)	(3.35)	0.82	(0.65)	(0.55)	(1.41)	0.81

Applications of visual attention

Intelligent image cropping

Image retrieval

Improved image classification

Intelligent image cropping

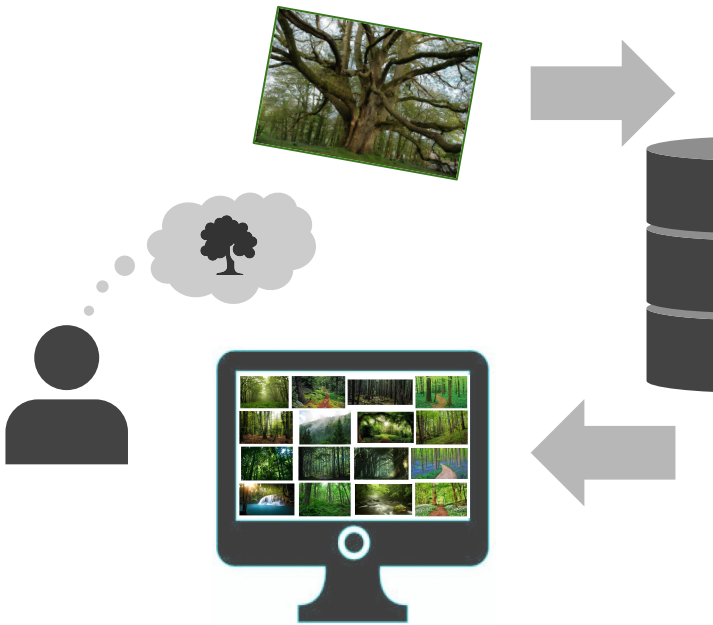
Image retrieval: query by example

Given:

- An example query image that illustrates the user's information need
- A very large dataset of images

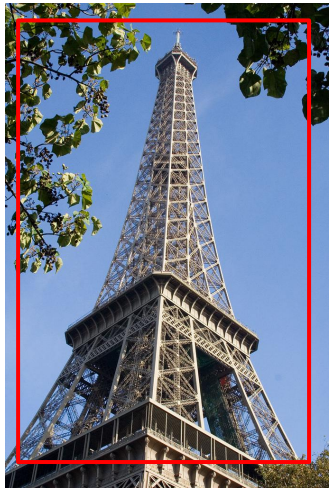
Task:

- Rank all images in the dataset according to how likely they are to fulfil the user's information need

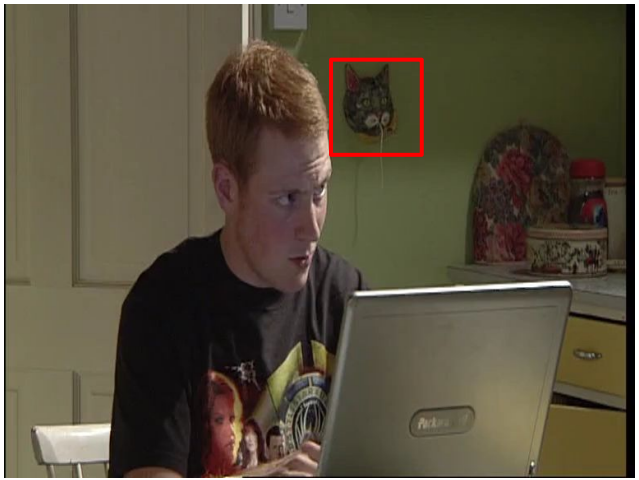


Retrieval benchmarks

Oxford Buildings
2007



Paris Buildings
2008



TRECVID INS
2014

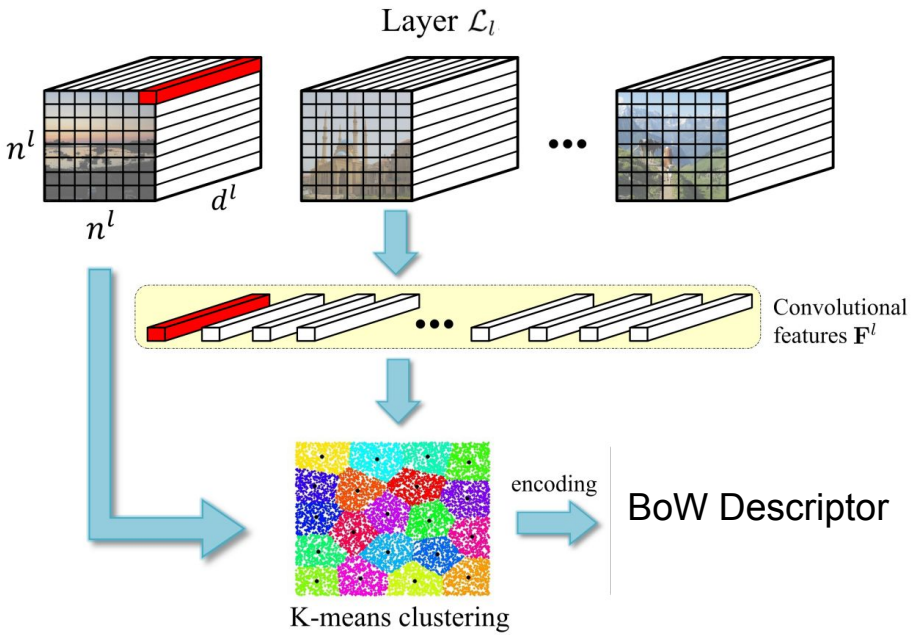
Bags of convolutional features instance search

Objective: rank images according to relevance to query image

Local CNN features and BoW

- Pretrained VGG-16 network
- Features from conv-5
- L2-norm, PCA, L2-norm
- K-means clustering -> BoW
- Cosine similarity
- Query augmentation, spatial reranking

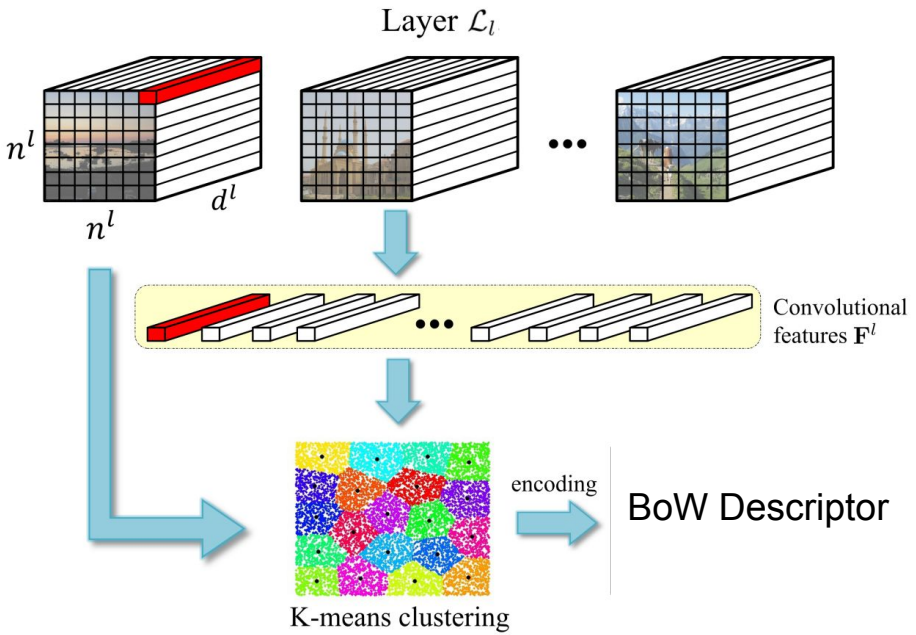
Scalable, fast, high-performance on Oxford 5K, Paris 6K and TRECVid INS



Mohedano et al. **Bags of Local Convolutional Features for Scalable Instance Search**, ICMR 2016 http://arxiv.org/abs/1604.04653

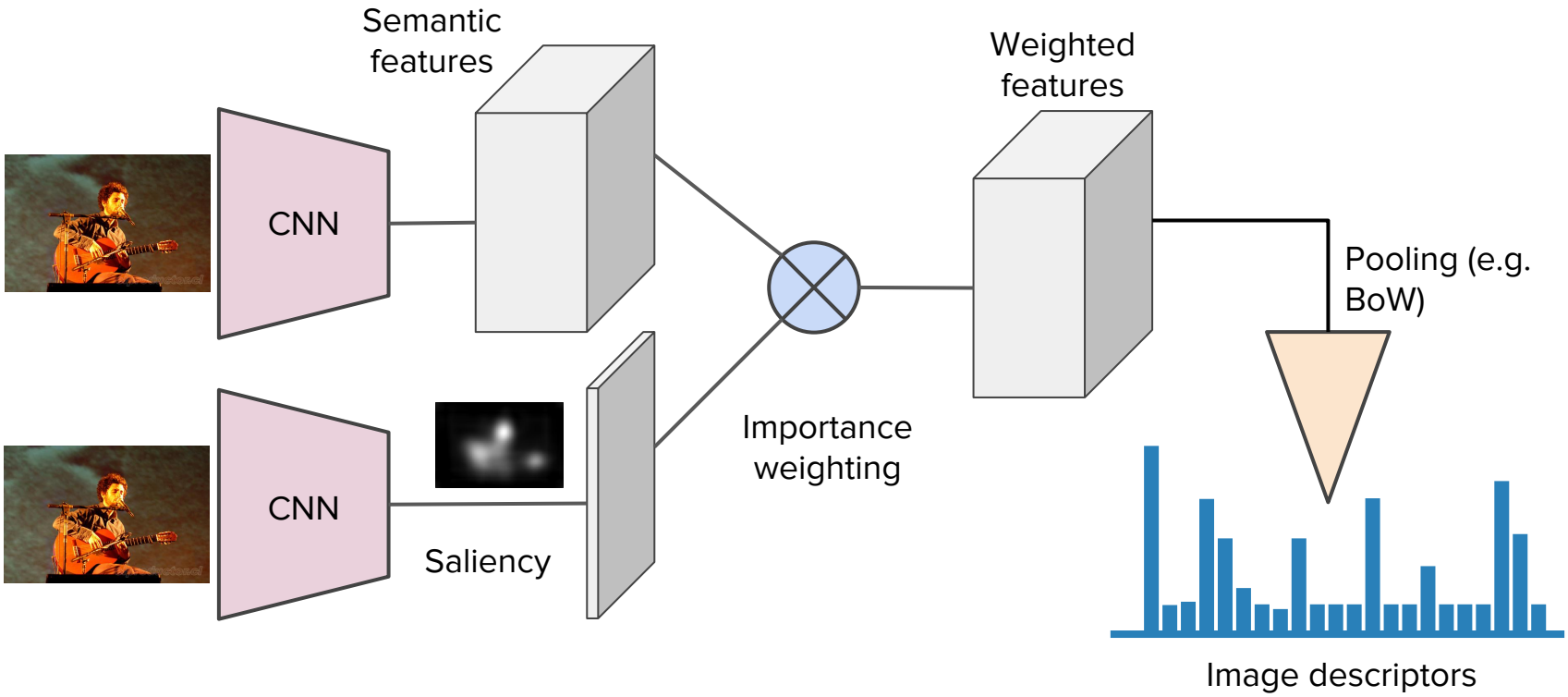
Bags of convolutional features instance search

		Oxford 5k	Paris 6k	INS 23k
BoW	GS	0.650	0.698	**0.323**
	LS	**0.739**	**0.819**	0.295
Sum pooling (as ours)	GS	0.606	0.712	0.156
	LS	0.583	0.742	0.097
Sum pooling (as in [7])	GS	0.672	0.774	0.139
	LS	0.683	0.763	0.120



Mohedano et al. **Bags of Local Convolutional Features for Scalable Instance Search**, ICMR 2016 http://arxiv.org/abs/1604.04653

Using saliency to improve retrieval

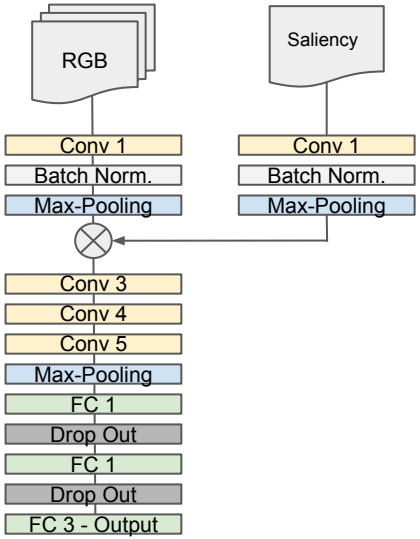


Saliency weighted retrieval

	Oxford		Paris		INSTRE	
	Global	**Local**	**Global**	**Local**	**Global**	**Local**
No weighting	0.614	0.680	0.621	0.720	0.304	0.472
Center prior	0.656	0.702	0.691	0.758	0.407	0.546
Saliency	0.680	0.717	0.716	0.770	0.514	0.617
QE saliency	-	**0.784**	-	**0.834**		**0.719**

Mean Average Precision

Using saliency to improve image classification



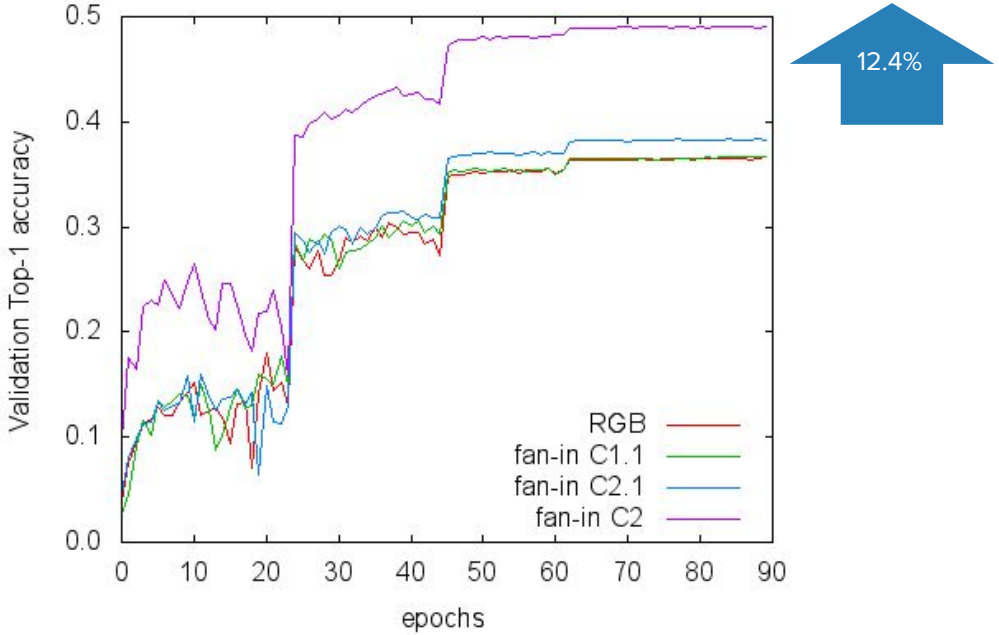
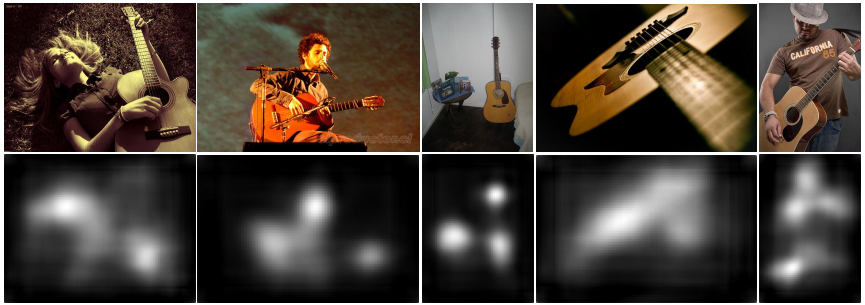


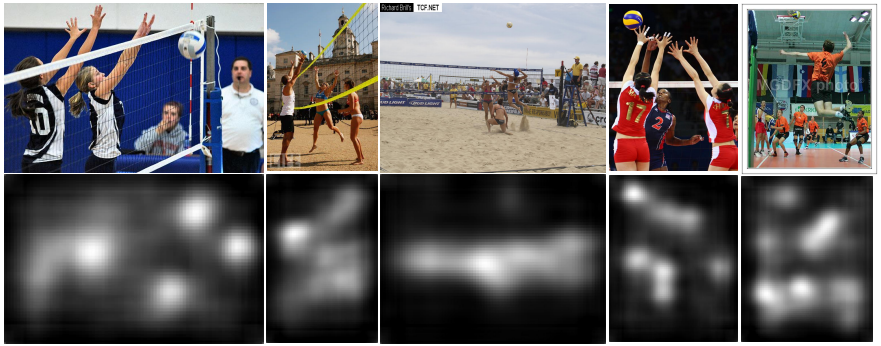
Figure credit: Eric Arazo

Why does it improve classification accuracy?

Acoustic guitar +25 %



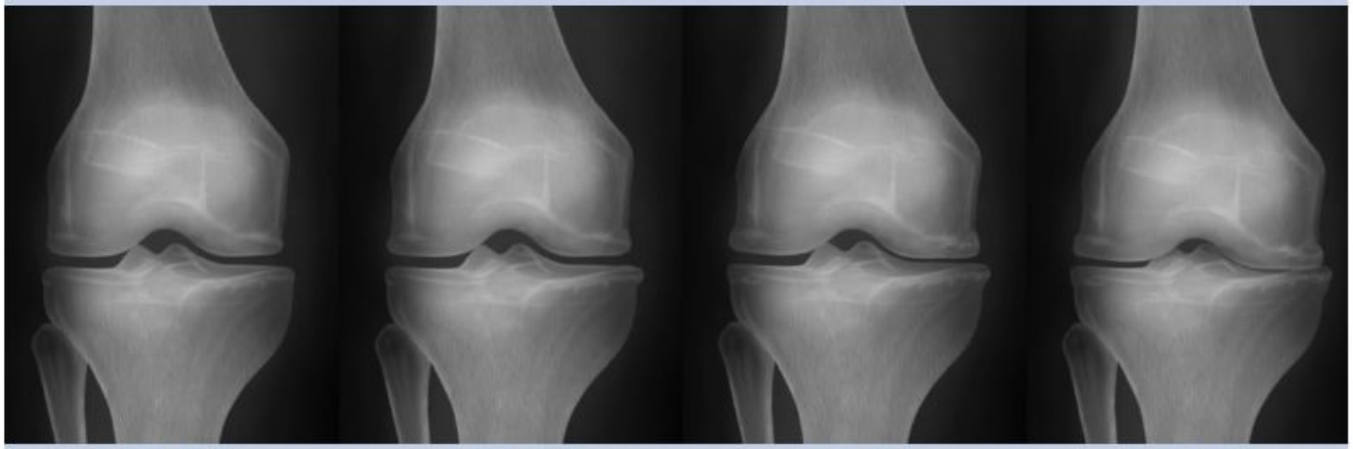
Volleyball +23 %



Deep Vision for Medical Image Analysis

Task: predict KL grade from X-Ray images

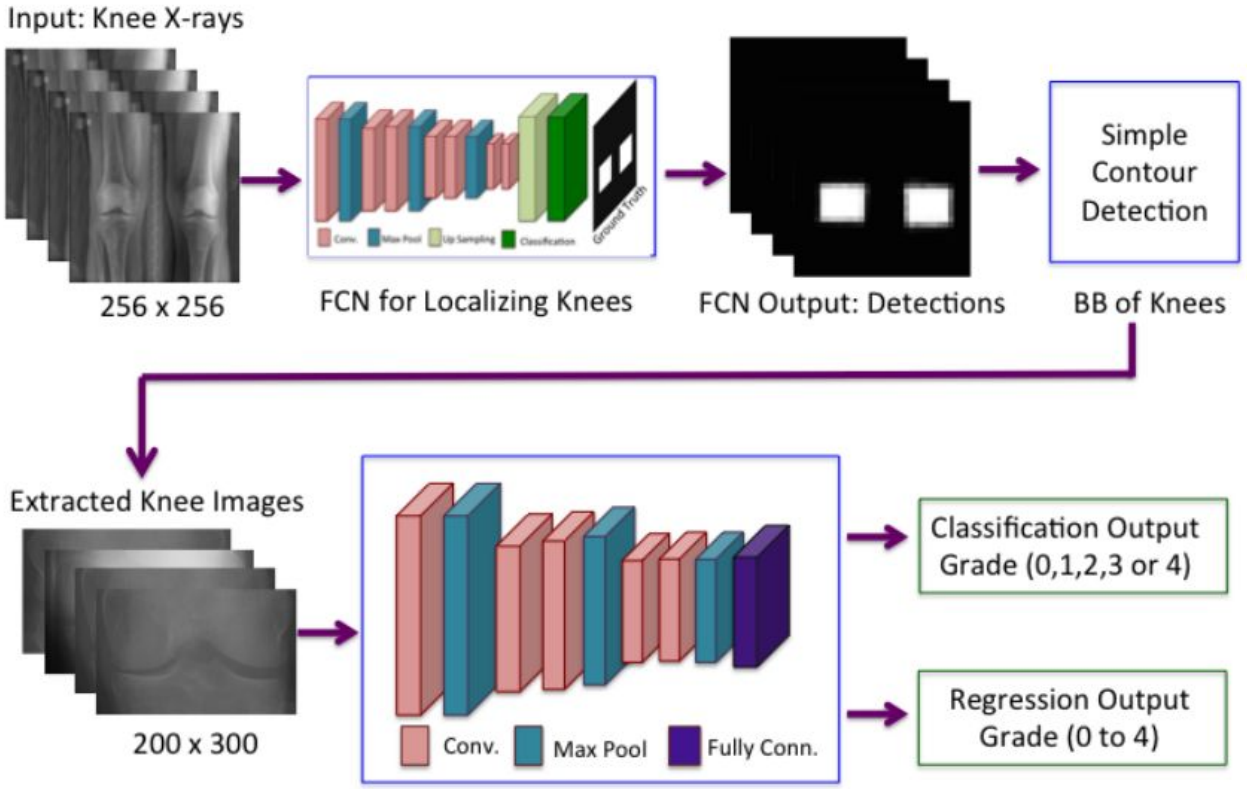
Kellgren-Lawrence (KL) grading scale



Grade 1 | Grade 2 | Grade 3 | Grade 4

CLASSIFICATION	Normal	Doubtful	Mild	Moderate	Severe
DESCRIPTION	No features of OA	Minute osteophyte: doubtful significance	Definite osteophyte: normal joint space	Moderate joint space reduction	Joint space greatly reduced: subchondral sclerosis

Pipeline: locate and classify



Detection performance

FCN detection performance

Test Data	J≥0.25	J≥0.5	J≥0.75	Mean	Std.Dev
OAI	**100%**	**99.9%**	89.2%	0.83	0.06
MOST	99.5%	98.4%	85.0%	0.81	0.09
Combined OAI-MOST	99.9%	**99.9%**	**91.4%**	0.83	0.06

Template matching: (J > 0.5) 8.3%

SVM on handcrafted features: (J > 0.5): 38.6%

Multi-objective learning helps!

Same network used to regress on KL grade and predict a discrete KL grade

Jointly train on both objectives

Grade	Joint training for Clsf & Reg				Training for only Clsf			
	Precision	Recall	F_1	AUC	Precision	Recall	F_1	AUC
0	0.68	0.80	0.74	0.87	0.63	0.82	0.71	0.83
1	0.32	0.15	0.20	0.71	0.25	0.04	0.06	0.66
2	0.53	0.63	0.58	0.82	0.47	0.57	0.51	0.78
3	0.78	0.74	0.76	0.96	0.76	0.71	0.73	0.94
4	0.81	0.75	0.78	0.99	0.78	0.77	0.77	0.99
Mean	0.61	0.63	0.61	-	0.56	0.60	0.56	-

Comparison with the state of the art

Method	Test Data	Accuracy	Mean-Squared Error
Wndchrm	OAI	29.3%	2.496
Wndchrm	MOST	34.8%	2.112
Fine-Tuned BVLC CaffeNet	OAI	57.6 %	0.836
Our CNN trained from Scratch	OAI & MOST	**60.3%**	0.898

How far are we from human-level accuracy?

Most errors are between grade 0 and 1 and grade 1 and 2.
Human experts have a hard time with these grades too.

Agreement among humans on OAI

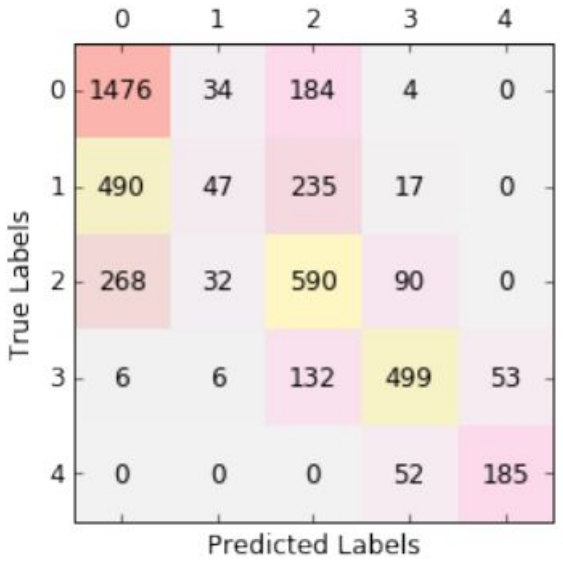
- weighted kappa of 0.70 [0.65-0.76]

Human machine agreement

- weighted kappa of 0.67 [0.65-0.68]

Predictions agree with the “gold standard” about as well as the “gold standard” agrees with itself.

Confusion matrix



Neonatal brain image segmentation

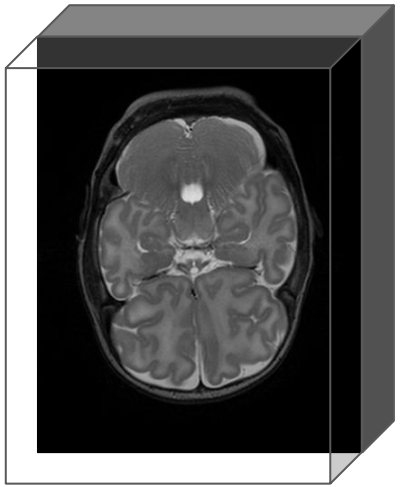
Volumetric **semantic segmentation:** label each pixel with class of brain matter.

Applications:

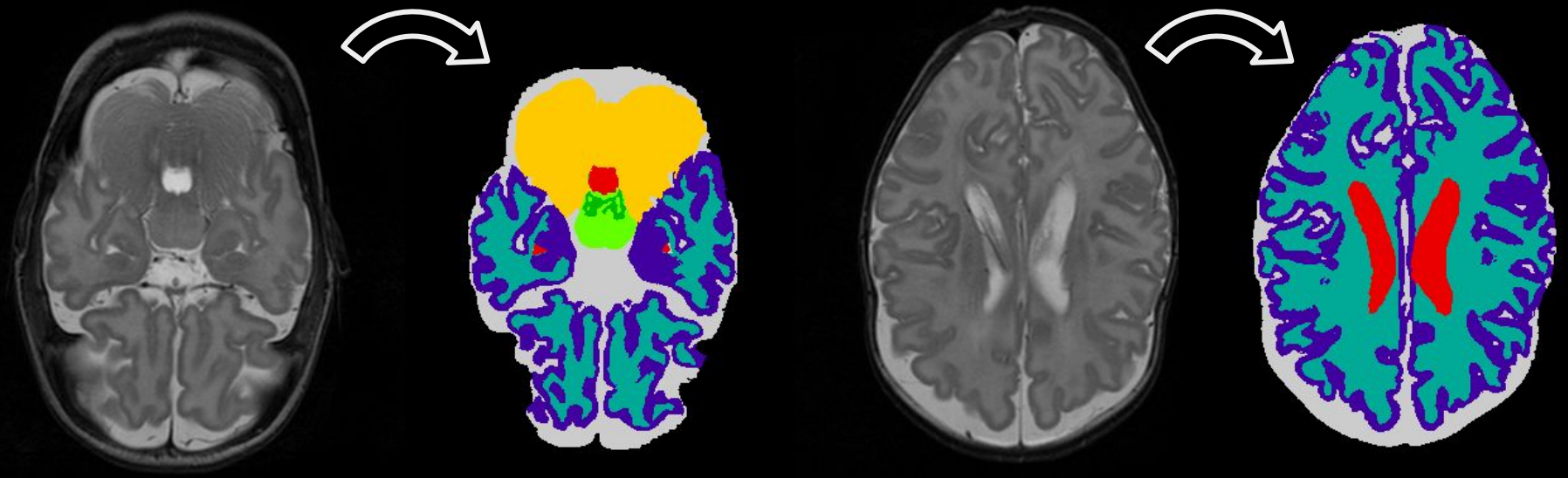
- Prerequisite for volumetric analysis
- Early identification of risk factors for impaired brain development

Challenge:

- Neonatal brains very different
- Sparse training data! Neobrains challenge has 2 training examples



The task



Cerebellum

Unmyelinated white matter

Cortical grey matter

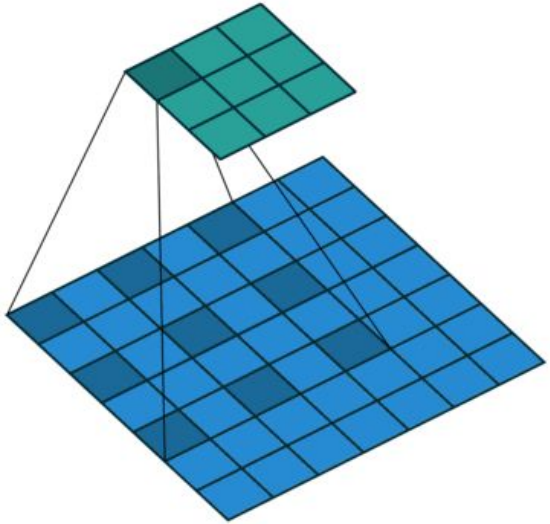
Ventricles

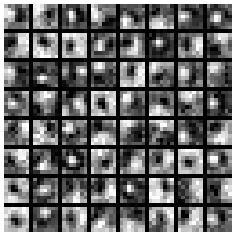
Cerebrospinal fluid

Brainstem

Model

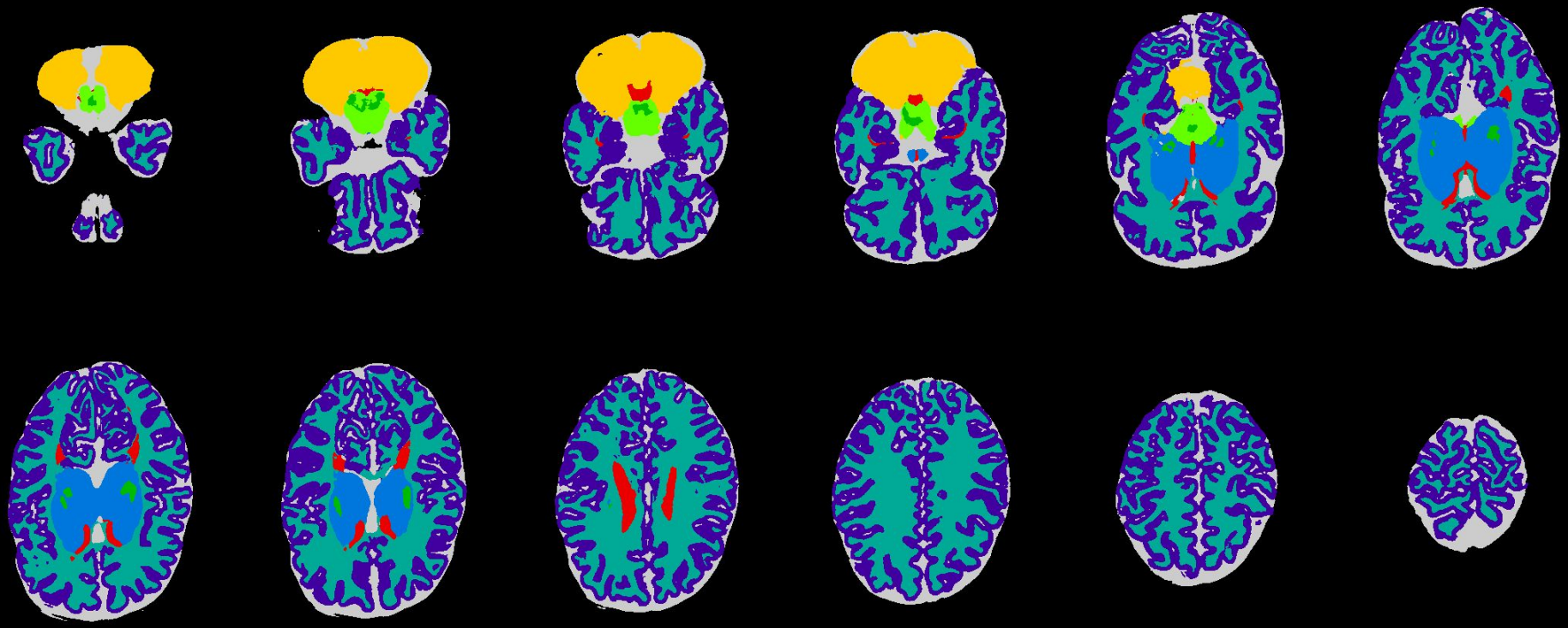
- 8 layer FCN
- 64/96 convolution filters per layer
- **Atrous (dilated) convolution** to increase receptive field without sacrificing prediction resolution
- 9D per pixel softmax over classes
- Binary cross entropy loss
- L^2 regularization
- **Aggressive data augmentation**: scale, crop, rotate, flip, gamma
- Train on 2 axial volumes (~50 slides per volume) for 500 epochs using Adam optimizer





Animation from: http://nicolovaligi.com/deep-learning-models-semantic-segmentation.html

Sample results



Neobrains challenge

New state of the art on Neobrains infant brain segmentation challenge for axial volume segmentation

Deep learning with only 2 training examples!

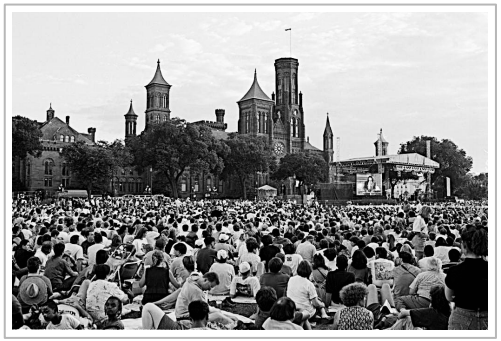
No ensembling yet. Best competing approach is a large ensemble.

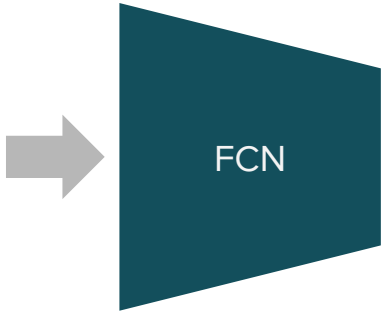
Second best is also a deep net.

Tissue	Ours	LRDE_LTCI	UPF_SIMBioSys
Cerebellum	0.92	0.94	0.94
Myelinated white matter	0.51	0.06	0.54
Basal ganglia and thalami	0.91	0.91	0.93
Ventricles	0.89	0.87	0.83
Unmyelinated white matter	0.93	0.93	0.91
Brainstem	0.82	0.85	0.85
Cortical grey matter	0.88	0.87	0.85
Cerebrospinal fluid	0.83	0.83	0.79
UWM+MWM	0.93	0.93	0.90
CSF+Ven	0.84	0.84	0.79
	0.85	**0.80**	**0.83**

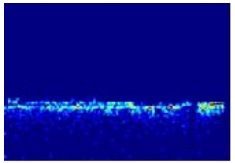
Deep Vision for Crowd Analysis

Fully convolutional crowd counting





Crowd density estimate



Crowd count estimate

8x conv layers
2x max pooling

Marsden et al. **Fully convolutional crowd counting on highly congested scenes.** VISAPP 2017
https://arxiv.org/abs/1612.00220


True count: 1544
Predicted count: 1566

Benchmark results: UCF CC 50 dataset

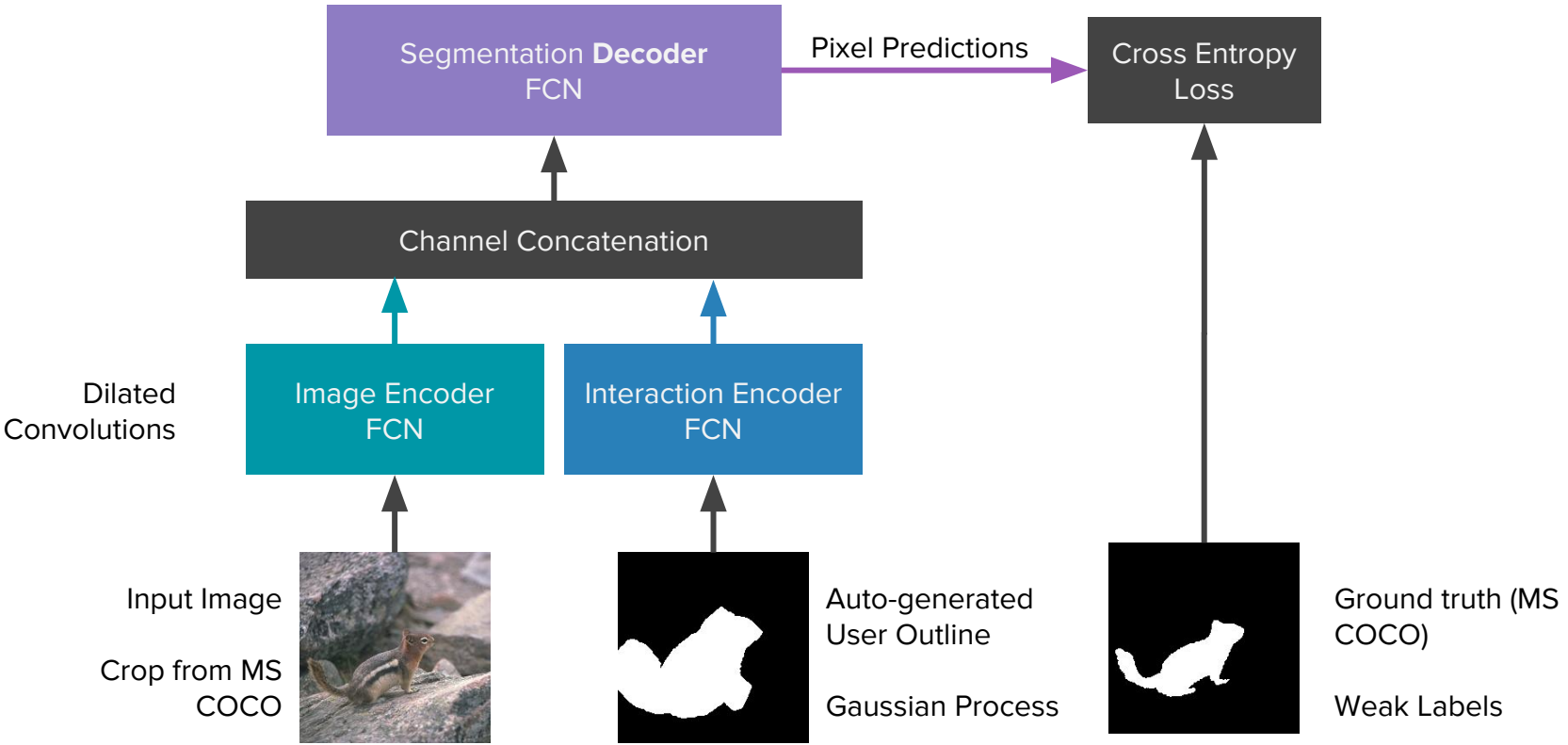
Method	Mean Absolute Error	Mean Squared Error
(Rodriguez et al., 2011)	655.7	697.8
(Lemiptsky and Zisserman, 2010)	493.4	487.1
(Idress et al., 2013)	419.5	541.6
(Zhang et al.,2015)	467	498.6
(Zhang et al.,2016)	377.6	509.1
Our Approach	**338.6**	**425.5**

45-45,000 people per image
State of the art improved by 11% (MSE) and 13% (MSE)

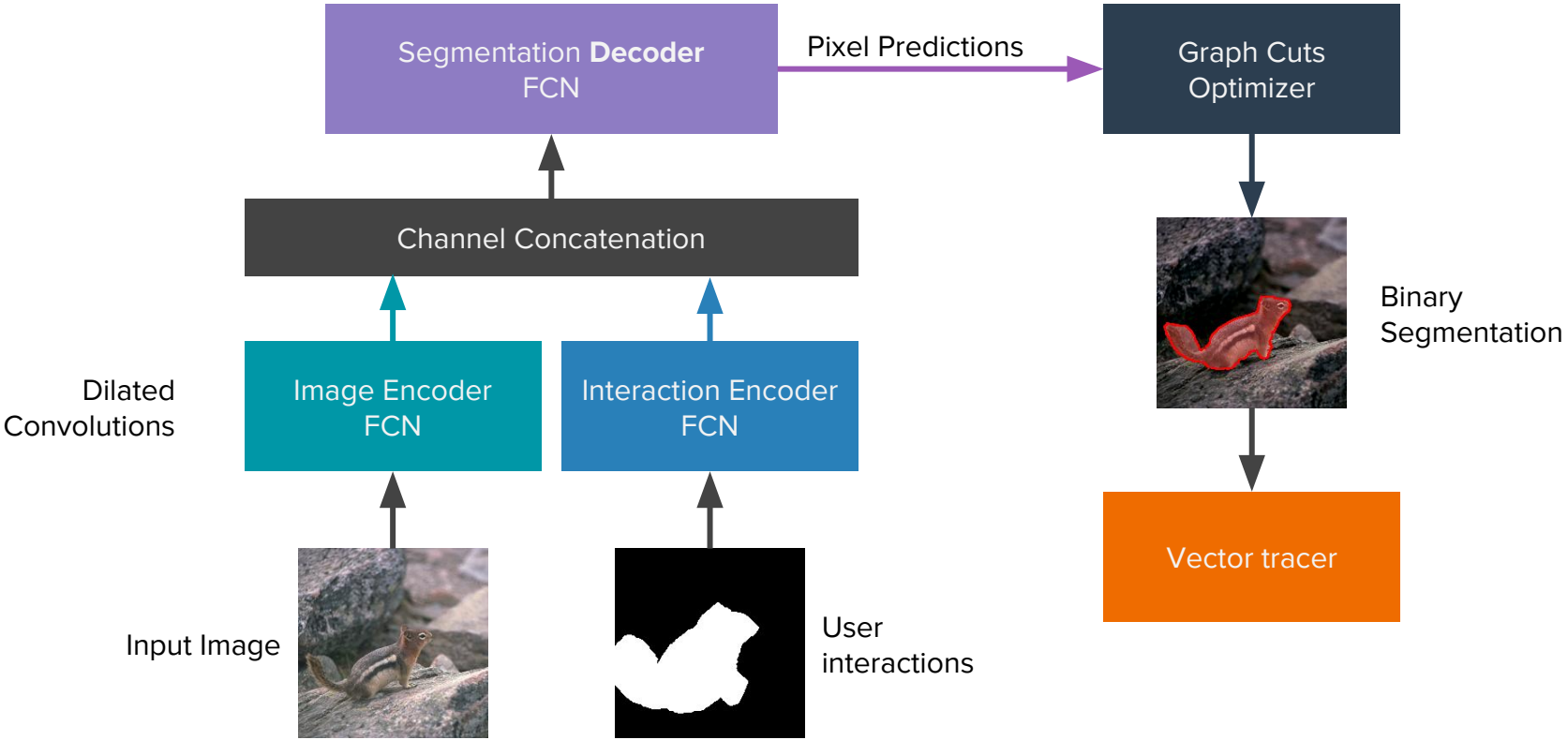
Interactive Deep Vision

Interactive image segmentation

DeepClip: training



DeepClip: prediction



Closing remarks

- Deep learning has completely revolutionized computer vision
- Human visual attention is important! Incorporating visual attention models helps in many tasks
- You don’t need a huge amount of training data to train an effective deep model
- Simulation techniques are effective for data generation
- Multi-task deep learning is an effective way of providing “more signal” during training.

Questions?