Modern Saliency Models

Georgii Zhulikov

The definition of saliency

Saliency as low-level visual attention

- Biological process
- Not clear how to measure
    - Hard to split various components of attention
- The best model is the most plausible one, for example the original Itti&Koch
- Modelling is about describing a process

Saliency as gaze prediction measure

- Quantitative metric of attention
- Can be measured by aggregating smoothed fixation data
- The best model is the most accurate one
- Modelling can be rephrased as a mathematical problem

“Learning to Predict Where Humans Look”

- Rephrase the problem
- Collect fixation data
- Build a machine learning model
- Introduce a quantitative metric to compare the prediction with the ground truth

Additional aspects of the modern approach

MIT Saliency Benchmark

- Refine the goal
- Provide clear results
- Inspire competition

SALICON

- Large scale data for large scale training
- Transfer learning
- Easy to work with

Convolutional Neural Networks

- Current best CNNs: VGGNet-16, ResNet-50, DenseNet-161, NasNet-Large
- Pre-trained ImageNet features
- Semantic Segmentation approach

EML-NET

Encoding

- Compute feature maps using a regular classifier CNN

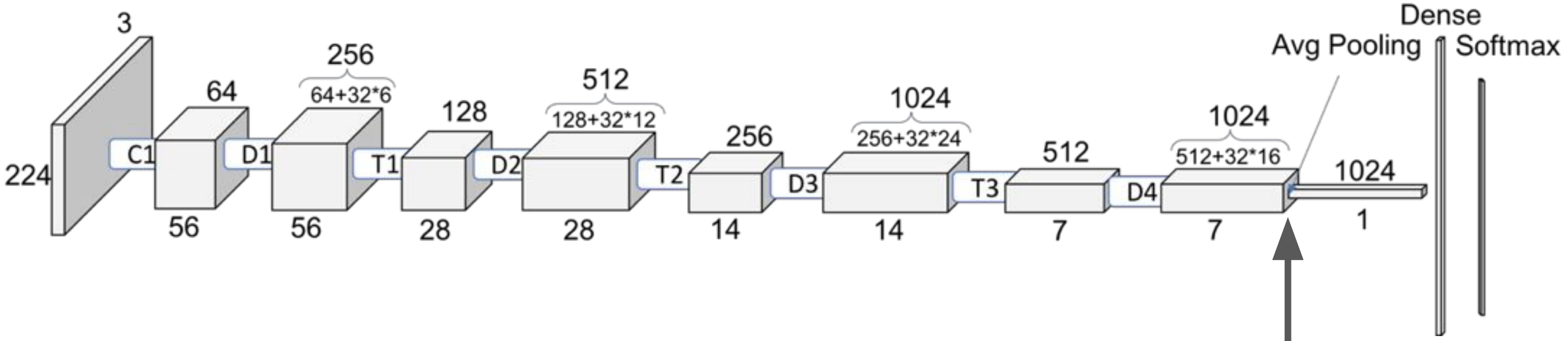
Decoding

- Use 1x1 convolution instead of the Fully Connected layers to combine all the feature maps into a single one
- Upscale bilinearly

EML-NET

Encoding

- Compute **feature maps** using a regular classifier CNN

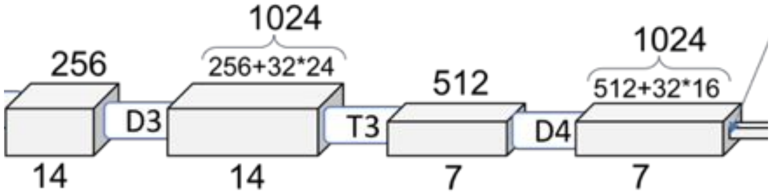


Stop here, no Avg Pooling
and the following layers

EML-NET

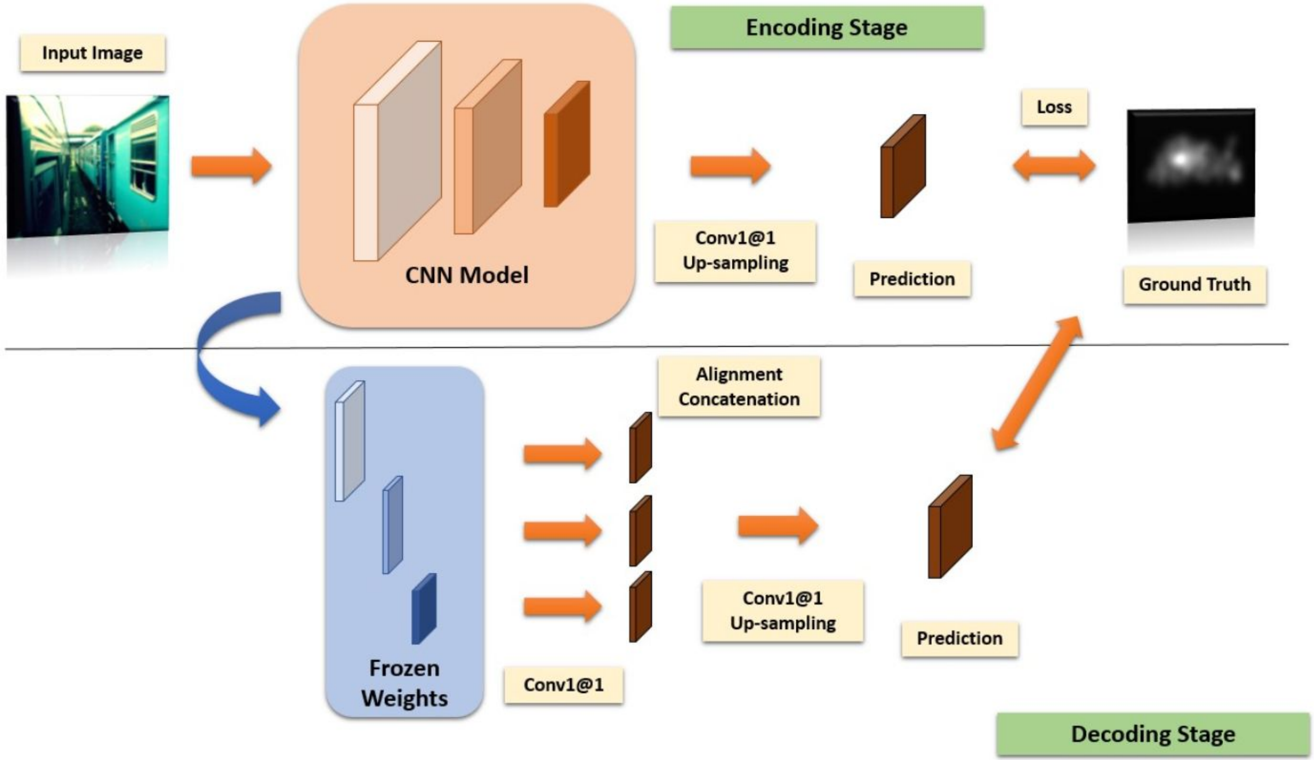
Decoding

- Use 1x1 convolution instead of the Fully Connected layers to combine all the feature maps into a single one
- Upscale bilinearly



- Leave 1024 feature maps
- Concatenate them with maps from other networks
- Sum them up with weights (1x1 convolution)
- Upscale

EML-NET


Input Image
CNN Model
Encoding Stage
Conv1@1
Up-sampling
Prediction
Loss
Ground Truth
Frozen
Weights
Conv1@1
Alignment
Concatenation
Conv1@1
Up-sampling
Prediction
Decoding Stage

EML-NET

- Modular structure: add or remove new feature maps from new encoders
- Computational efficiency: removing FC layers and combining features before upsampling greatly saves space
- Careful metric selection for training: NSS+CC+KLD

- Tested networks are DenseNet, NasNet, DenseNet+NasNet

EML-NET Results

Results on the MIT dataset

Method	AUC-Judd	SIM	EMD	AUC-Borji	sAUC	CC	NSS	KLD
eDN[40]	0.82	0.41	4.56	0.81	0.62	0.45	1.14	1.14
DeepGaze1[38]	0.84	0.39	4.97	0.83	0.66	0.48	1.22	1.23
DeepGaze2[27]	0.88	0.46	3.98	0.86	0.72	0.52	1.29	0.96
BMS[46]	0.83	0.51	3.35	0.82	0.65	0.55	1.41	0.81
iSEEL[37]	0.84	0.57	2.72	0.81	0.68	0.65	1.78	0.65
DVA[41]	0.85	0.58	3.06	0.78	0.71	0.68	1.98	0.64
SalGAN[31]	0.86	0.63	2.29	0.81	0.72	0.73	2.04	1.07
PDP[16]	0.85	0.60	2.58	0.80	0.73	0.70	2.05	0.92
ML-Net[8]	0.85	0.59	2.63	0.75	0.70	0.67	2.05	1.10
Salicon[14]	0.87	0.60	2.62	0.85	0.74	0.74	2.12	0.54
DeepFix[26]	0.87	0.67	2.04	0.80	0.71	0.78	2.26	0.63
SAM-Res[9]	0.87	0.68	2.15	0.78	0.70	0.78	2.34	1.27
DSCLRCN[29]	0.87	0.68	2.17	0.79	0.72	0.80	2.35	0.95
DPN[30]	0.87	0.69	2.05	0.80	0.74	0.82	2.41	0.91
EML-NET	0.88	0.68	1.84	0.77	0.70	0.79	2.47	0.84

EML-NET Results

The CAT2000 dataset contains images of unusual classes while EML-NET was trained on natural scenes

The modular structure allows for easy addition of the new types of images, so these results can be improved

Method	AUC-Judd	SIM	EMD	AUC-Borji	sAUC	CC	NSS	KLD
eDN[40]	0.85	0.52	2.64	0.84	0.55	0.54	1.30	0.97
BMS[46]	0.85	0.61	1.95	0.84	0.59	0.67	1.67	0.83
iSEEL[37]	0.84	0.62	1.78	0.81	0.59	0.66	1.67	0.92
DeepFix[26]	0.87	0.74	1.15	0.81	0.58	0.87	2.28	0.37
SAM-Res[9]	0.88	0.77	1.04	0.80	0.58	0.89	2.38	0.56
EML-NET	0.87	0.74	1.05	0.78	0.58	0.87	2.38	0.95

Results on the CAT2000 dataset

Thank you