How much position information do convolutional neural network encode?

Md Amirul Islam, Sen Jia, Neil D. B. Bruce
Ryerson University
Vector Institute for Artificial Intelligence

ICML2020
PRESENTER: ZEJIA FAN

STRUCT GROUP SEMINAR
2021/10/21

Outline

- Authors
- Motivation
- Method
- Experiments
- Conclusion

Motivation

- CNN lacks of interpretability
- CNN localization
- Absolute spatial information is important for position-dependent tasks: semantic segmentation, object detection

Motivation

- CNN lacks of interpretability
- CNN localization
- Absolute spatial information is important for position-dependent tasks: semantic segmentation, object detection

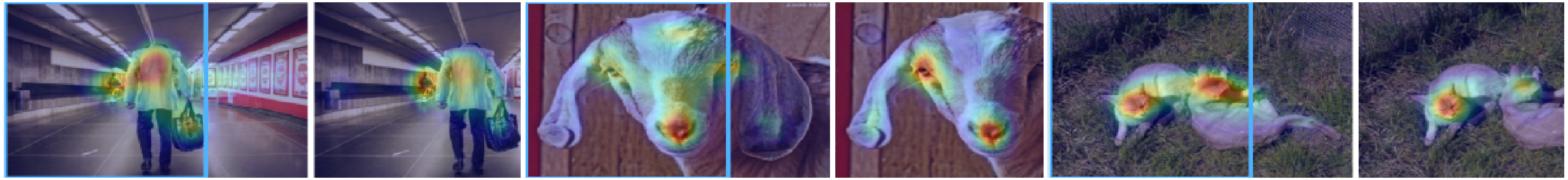


Figure 1: Sample predictions for salient regions for input images (left), and a slightly cropped version (right). Cropping results in a shift in position rightward of features relative to the centre. It is notable that this has a significant impact on output and decision of regions deemed salient despite no explicit position encoding and a modest change to position in the input.

Problem Formulation

Problem Formulation: Given an input image $\mathcal{I}_m \in \mathbb{R}^{h \times w \times 3}$, our goal is to predict a gradient-like position information mask $\hat{f}_p \in \mathbb{R}^{h \times w}$ where each pixel value defines the absolute coordinates of an pixel from left$\rightarrow$right or top$\rightarrow$bottom. We generate gradient-like masks $\mathcal{G}_{pos} \in \mathbb{R}^{h \times w}$ (Sec. 2.2) for supervision in our experiments, with weights of the base CNN archetypes being fixed.

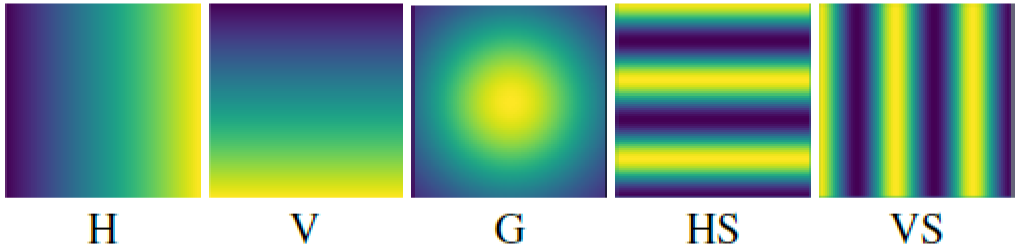


Figure 3: Sample images and generated gradient-like ground-truth position maps.

Position Encoding Network

- Backbone as ResNet, VGG, weight frozen
- Resize, concate, convolution

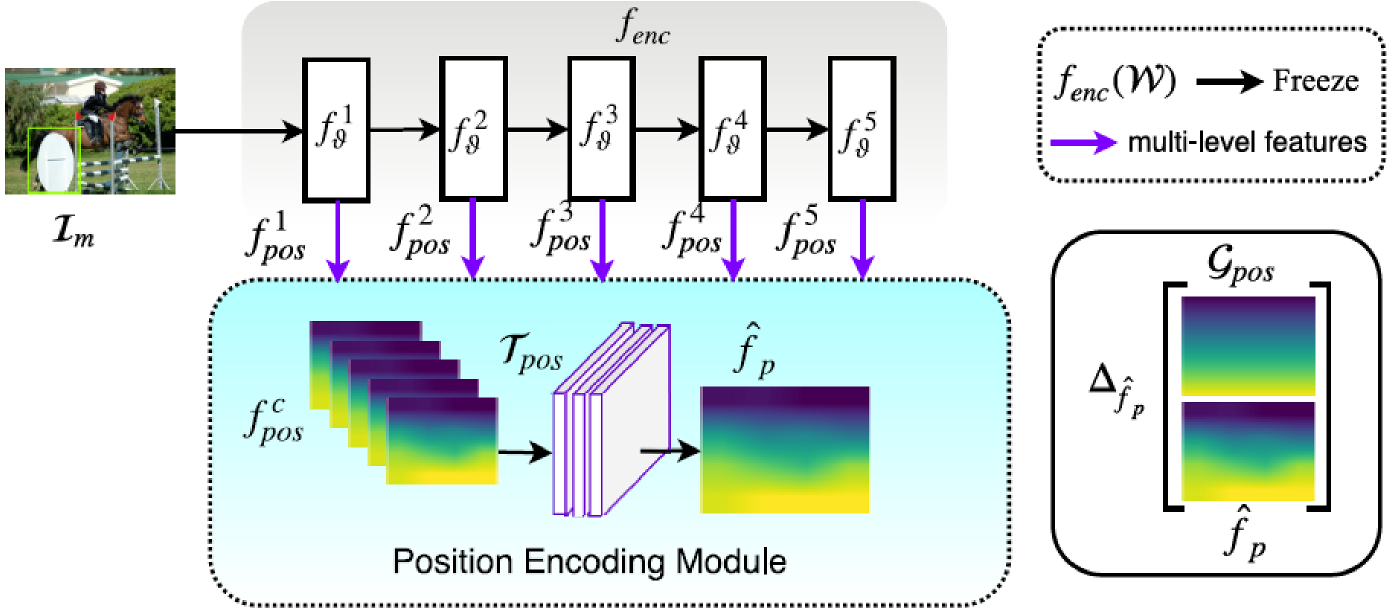


Figure 2: Illustration of PosENet architecture.

Experiment Setting

- Add data content independent case
- Test if contain 2D absolute position information
- MSE loss

$$\Delta_{\hat{f}_p} = \frac{1}{2n}\sum_{i=1}^{n}(x_i - y_i)^2$$

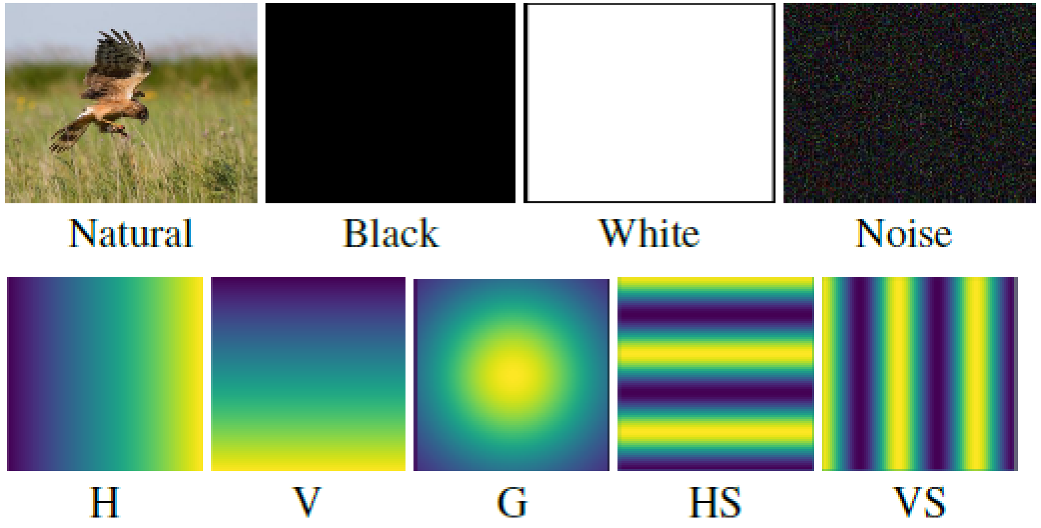


Figure 3: Sample images and generated gradient-like ground-truth position maps.

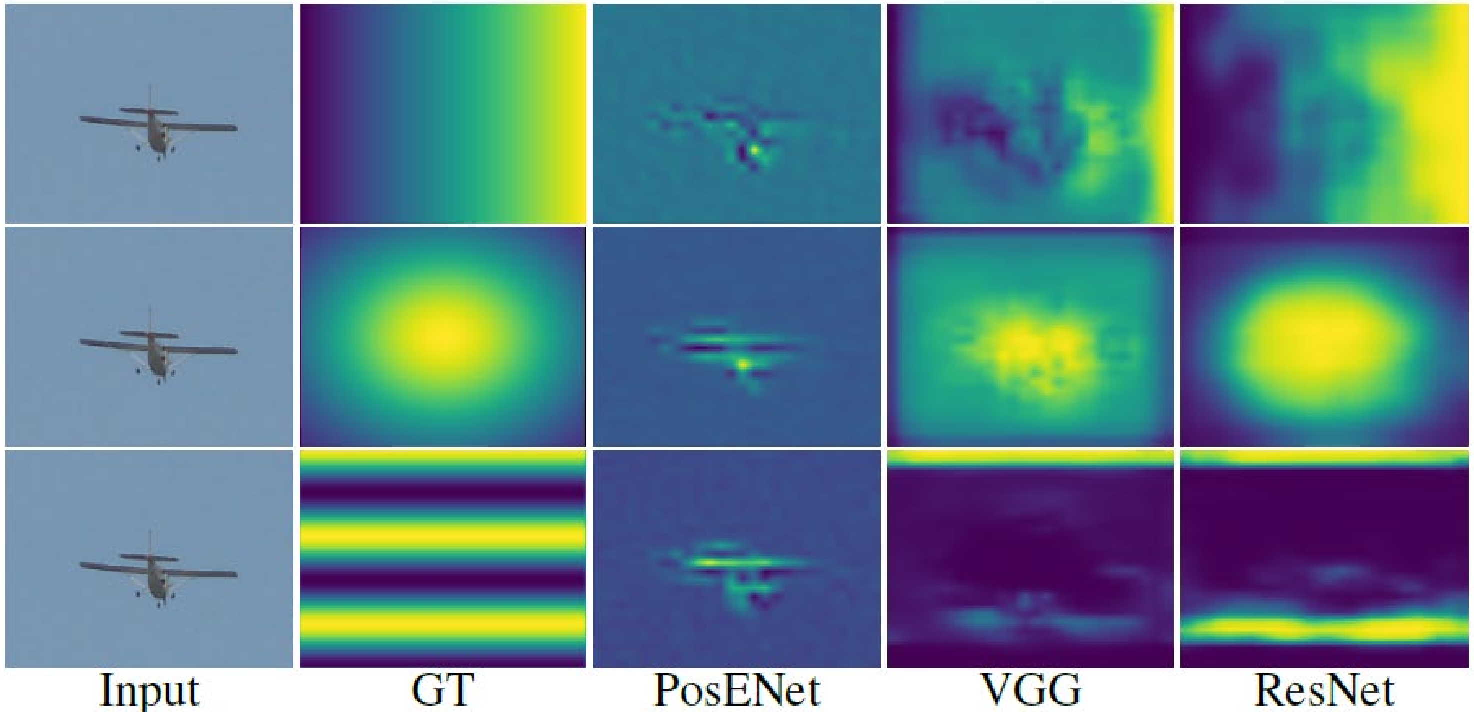
Experiment Results

- SPC for correlation, MAE for Mean Absolute Error
- VGG16, ResNet-152

	Model	PASCAL-S		Black		White		Noise	
		SPC	MAE	SPC	MAE	SPC	MAE	SPC	MAE
H	PosENet	.012	.251	.0	.251	.0	.251	.001	.251
	VGG	.742	.149	.751	.164	.873	.157	.591	.173
	ResNet	.933	.084	.987	.080	.994	.078	.973	.077
V	PosENet	.131	.248	.0	.251	.0	.251	.053	.250
	VGG	.816	.129	.846	.146	.927	.138	.771	.150
	ResNet	.951	.083	.978	.069	.979	.072	.968	.074
G	PosENet	-.001	.233	.0	.186	.0	.186	-.034	.214
	VGG	.814	.109	.842	.123	.898	.116	.762	.129
	ResNet	.936	.070	.953	.068	.964	.064	.971	.055
HS	PosENet	-.001	.712	-.055	.704	.0	.704	.023	.710
	VGG	.405	.556	.532	.583	.576	.574	.375	.573
	ResNet	.534	.528	.566	.518	.562	.515	.471	.530
VS	PosENet	.006	.723	.081	.709	.081	.709	.018	.714
	VGG	.374	.567	.538	.575	.437	.578	.526	.566
	ResNet	.520	.537	.574	.523	.593	.514	.523	.545

Experiment Results

- Qualitative results of PosENet based networks corresponding to different ground-truth patterns.



Ablation Study

- Different receptive field
- VGG improves, PosENet keeps the same

	Layers	PosENet		VGG	
		SPC	MAE	SPC	MAE
H	1 Layer	.012	.251	.742	.149
	2 Layers	.056	.250	.797	.128
	3 Layers	.055	.250	.830	.117
G	1 Layer	-.001	.233	.814	.109
	2 Layers	.067	.187	.828	.105
	3 Layers	.126	.186	.835	.104
HS	1 Layer	-.001	.712	.405	.556
	2 Layers	-.006	.628	.483	.538
	3 Layers	.003	.628	.491	.540

(a)

	Kernel	PosENet		VGG	
		SPC	MAE	SPC	MAE
H	1×1	.013	.251	.542	.196
	3×3	.012	.251	.742	.149
	7×7	.060	.250	.828	.120
G	1×1	.017	.188	.724	.127
	3×3	-.001	.233	.814	.109
	7×7	.068	.187	.816	.111
HS	1×1	-.004	.628	.317	.576
	3×3	-.001	.723	.405	.556
	7×7	.002	.628	.487	.532

(b)

Table 2: Quantitative comparison on the PASCAL-S dataset in terms of SPC and MAE with varying (a) number of layers and (b) kernel sizes. Note that (a) the kernel size is fixed to 3×3 but different numbers of layers are used in the PosENet. (b) Number of layers is fixed to one but we use different kernel sizes in the PosENet.

Ablation Study

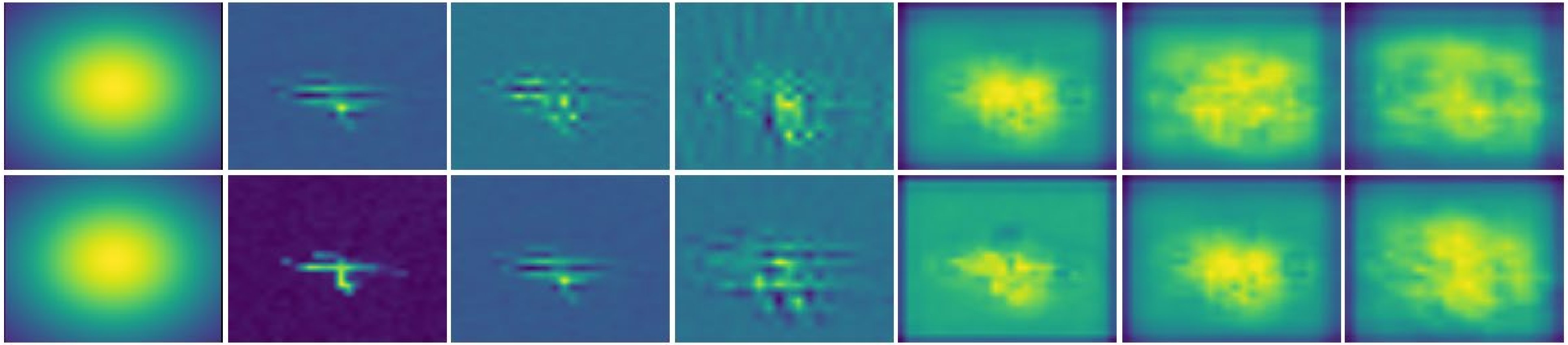


Figure 5: The effect of more **L**ayers (Top row) and varying **K**ernel **S**ize (bottom row) applied in the PoseNet. Order (left → right): GT (**G**), PosENet (L=1, KS=1), PosENet (L=2, KS=3), PosENet (L=3, KS=7), VGG (L=1, KS=1), VGG (L=2, KS=3), VGG (L=3, KS=7).

Ablation Study

- The power of different layer feature, deeper one encodes more position information.

	Method	f^1_{pos}	f^2_{pos}	f^3_{pos}	f^4_{pos}	f^5_{pos}	SPC	MAE
H	**VGG**	✓					.101	.249
			✓				.344	.225
				✓			.472	.203
					✓		.610	.181
						✓	.657	.177
		✓	✓	✓	✓	✓	.742	.149
G	**VGG**	✓					.241	.182
			✓				.404	.168
				✓			.588	.146
					✓		.653	.138
						✓	.693	.135
		✓	✓	✓	✓	✓	.814	.109

Table 3: Performance of VGG on natural images with a varying extent of the reach of different feed-forward blocks.

Effect – Zero Padding

- How zero padding affects the position information encoding.

Model	H		G		HS	
	SPC	MAE	SPC	MAE	SPC	MAE
PosENet	.012	.251	-.001	.233	-.001	.712
PosENet with *padding*=1	.274	.239	.205	.184	.148	.608
PosENet with *padding*=2	.397	.223	.380	.177	.214	.595
VGG16	.742	.149	.814	.109	.405	.556
VGG16 w/o. *padding*	.381	.223	.359	.174	.011	.628

Table 4: Quantitative comparison subject to padding in the convolution layers used in PosENet and VGG (w/o and with zero padding) on natural images.

Effect – Zero Padding

- How zero padding affects the position

Model	H		
	SPC	MAE	S
PosENet	.012	.251	-.(
PosENet with *padding*=1	.274	.239	.2
PosENet with *padding*=2	.397	.223	.3
VGG16	.742	.149	.8
VGG16 w/o. *padding*	.381	.223	.3

ConvNet Configuration					
A	A-LRN	B	C	D	E
11 weight layers	11 weight layers	13 weight layers	16 weight layers	16 weight layers	19 weight layers
input (224 × 224 RGB image)					
conv3-64	conv3-64 **LRN**	conv3-64 **conv3-64**	conv3-64 conv3-64	conv3-64 conv3-64	conv3-64 conv3-64
maxpool					
conv3-128	conv3-128	conv3-128 **conv3-128**	conv3-128 conv3-128	conv3-128 conv3-128	conv3-128 conv3-128
maxpool					
conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256	conv3-256 conv3-256 **conv1-256**	conv3-256 conv3-256 **conv3-256**	conv3-256 conv3-256 conv3-256 **conv3-256**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512	conv3-512 conv3-512 **conv1-512**	conv3-512 conv3-512 **conv3-512**	conv3-512 conv3-512 conv3-512 **conv3-512**
maxpool					
FC-4096					
FC-4096					
FC-1000					
soft-max					

Table 4: Quantitative comparison subject to padding in the convolution layers used in PosENet and VGG (w/o and with zero padding) on natural images.

Effect – Zero Padding

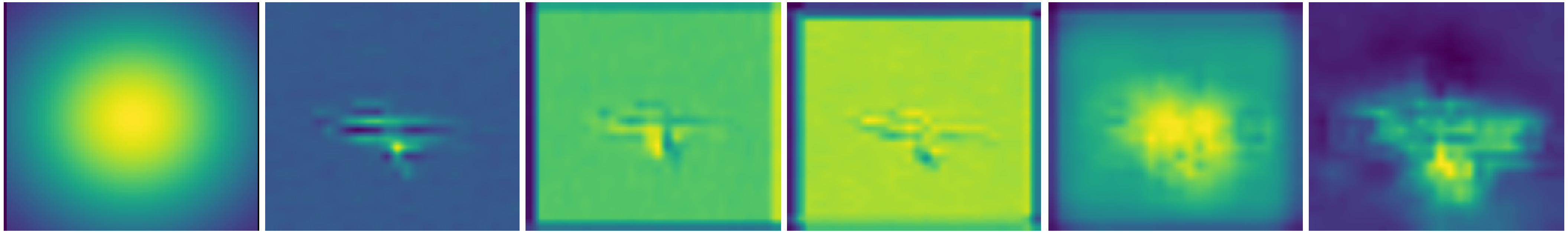


Figure 6: The effect of zero-padding on Gaussian pattern. Left to right: GT (**G**), Pad=0 (.286, .186), Pad=1 (.227, .180), Pad=2 (.473, .169), VGG Pad=1 (.928, .085), VGG Pad=0(.405, .170).

Error Heat Map

$$\mathcal{L} = \frac{|(\mathcal{G}^h_{pos} - \hat{f}^h_p)| + |(\mathcal{G}^v_{pos} - \hat{f}^v_p)| + |(\mathcal{G}^g_{pos} - \hat{f}^g_p)|}{3}$$

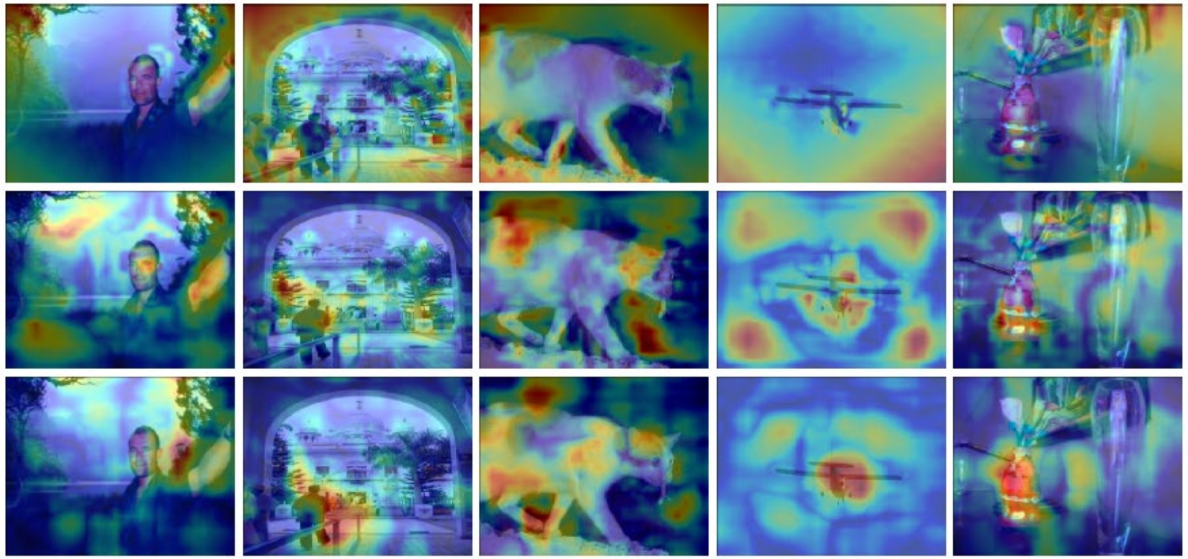


Figure 7: Error heat maps of PosENet (1st row), VGG (2nd row) and ResNet (3rd row).

SOD & SS

- Zero padding effects Saliency object detection and Semantic segmentation.

Model	ECSSD		PASCAL-S		DUT-OMRON	
	Fm	MAE	Fm	MAE	Fm	MAE
VGG w/o padding	.36	.48	.32	.48	.25	.48
VGG	.78	.17	.66	.21	.63	.18

(a)

Model	mIoU (%)
VGG w/o padding	12.3
VGG	23.1

(b)

Table 5: VGG models with and w/o zero-padding for (a) SOD and (b) semantic segmentation.

SOD & SS

- Saliency object detection and Semantic segmentation have higher requirements on position information than classification.
- (high-low combination)

	Model	PASCAL-S		BLACK		WHITE		NOISE	
		SPC	MAE	SPC	MAE	SPC	MAE	SPC	MAE
H	VGG	.742	.149	.751	.164	.873	.157	.591	.173
	VGG-SOD	.969	.055	.857	.099	.938	.087	.965	.060
	VGG-SS	.982	.038	.990	.030	.985	.032	.985	.033
G	VGG	.814	.109	.842	.123	.898	.116	.762	.129
	VGG-SOD	.948	.067	.904	.086	.907	.085	.912	.077
	VGG-SS	.971	.055	.984	.050	.989	.046	.982	.051
HS	VGG	.405	.556	.532	.583	.576	.574	.375	.573
	VGG-SOD	.667	.476	.699	.506	.709	.482	.668	.489
	VGG-SS	.810	.430	.802	.426	.810	.426	.789	.428

Table 6: Comparison of VGG models pretrained for classification, SOD, and semantic segmentation.

Conclusion

- Absolute position information is implicitly encoded in convolutional neural networks.
- Position information is encoded through zero-padding to some degree.
- High-low combination tasks may rely more on position information.

Thanks for your listening!