How much Position Information Do Convolutional Neural Networks Encode?

Published at ICLR 2020, spotlight

Md Amirul Islam, Sen Jia, Neil D. B. Bruce

Presenter: Jiaru Zhang

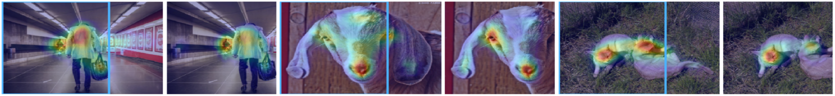
2020.5.16

Introduction

- CNNs are considered to be spatially-agnostic. *Capsule* and *recurrent networks* are proposed to model spatial information.

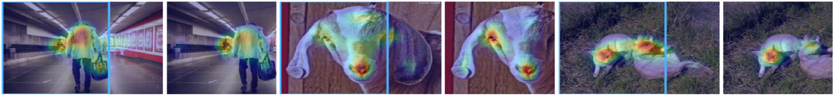
Introduction

- CNNs are considered to be spatially-agnostic. *Capsule* and *recurrent networks* are proposed to model spatial information.
- The regions determined to be most salient by CNNs tend to be near the center of an image.



Introduction

- CNNs are considered to be spatially-agnostic. *Capsule* and *recurrent networks* are proposed to model spatial information.
- The regions determined to be most salient by CNNs tend to be near the center of an image.



- This paper examines the role of absolute position information and reveal where position information comes from.

Position Information in CNNs

Problem Formulation

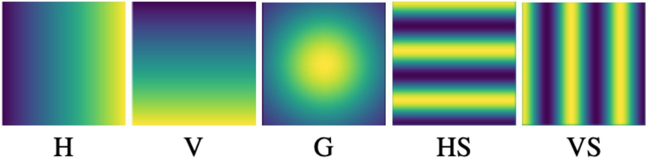
Given an input image $\mathcal{I}_m \in \mathbb{R}^{h \times w \times 3}$, our goal is to predict a gradient-like position information mask $\hat{f}_p \in \mathbb{R}^{h \times w}$ where each pixel value defines the absolute coordinates of an pixel from left $\rightarrow$ right or top $\rightarrow$ bottom.

Position Information in CNNs

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.

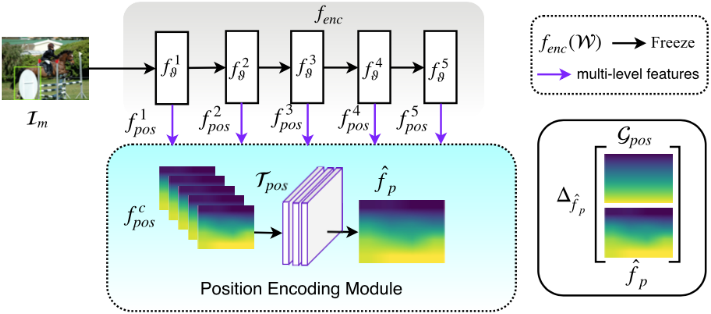
Here are some sample position maps:



Position Encoding Network

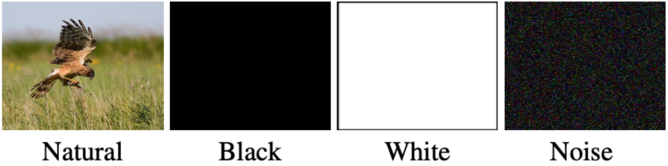
Position Encoding Network (PosENet) consists of f_{enc} and f_{pem}.

- **Encoder** f_{enc}: ResNet and VGG based architectures without average pooling layer and the last layer, frozen when probing the encoding network.
- **Position Encoding Module** f_{pem}: It takes features from f_{enc} as input and generates the desired position map.



Experiments Setting

- **Datasets**: Natural images from DUT-S and PASCAL-S, and synthetic images.



- **Evaluation Metrics**: Spearmen Correlation (SPC) and Mean Absoute Error (MAE). Higher SPC and lower MAE mean better performance.

Existing of Position Information

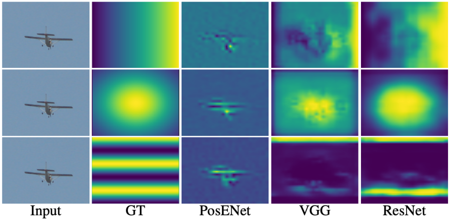
Models: VGG and ResNet based networks and PosENet without using any pretrained model.

- PosENet (VGG and ResNet) can extract position information from the pretrained CNN models.
- Training PosENet separately achieves much lower scores.

	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

Models: VGG and ResNet based networks and PosENet without using any pretrained model.

- PosENet (VGG and ResNet) can extract position information from the pretrained CNN models.
- Training PosENet separately achieves much lower scores.



Analyzing PosENet

The PosENet used has only one convolutional layer with a kernel size of 3×3. What about changing it?

- Applying more layers in the PosENet can improve the readout of position information for all the networks.
- A reason could be that the effective receptive field becomes larger.

	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)

Analyzing PosENet

The PosENet used has only one convolutional layer with a kernel size of 3×3. What about changing it?

- Larger kernel sizes are likely to capture more position information compared to smaller sizes.
- This also supports that a larger receptive field can better resolve position information.

	Kernel	PosENet SPC	PosENet MAE	VGG SPC	VGG 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)

Where is tha Position Information Stored?

It is also interesting to see whether position information is equally distributed across the layers.

- VGG based PosENet with top f_5^{pos} features achieves higher performance compared to bottom features.
- This is partially a result of more feature maps, 512 vs. 64.
- f_5^{pos} achieves better results than f_4^{pos}, suggests that the deeper feature contains more position information.

	Method	f_{pos}^1	f_{pos}^2	f_{pos}^3	f_{pos}^4	f_{pos}^5	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

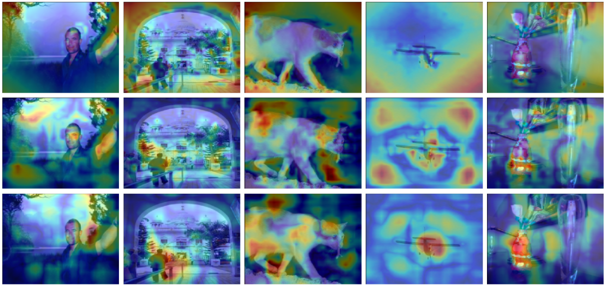
Where does Position Information Come from?

- The authors believe that the padding near the border delivers position information to learn.
- The VGG16 model without zero-padding achieves much lower performance than the default setting (padding=1) on the natural images.
- PosENet with larger padding achieves higher performance.
- This is also the reason why padding is not used in previous experiments.

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

Case Study

- The semantics within an image may affect the position map as shown in Page 6.
- The heatmaps of PosENet have larger content loss around the corners, and the heatmaps of VGG and ResNet correlate more with the semantic content.
- This visualization can be used to show which regions a model focuses on, especially in the case of ResNet.



Zero-Padding Driven Position Information

- Saliency Detection and Semantic Segmentation are two position-dependent tasks.
- VGG without padding achieves much worse results on both tasks, which further validates the findings that zero-padding is the key source of position information.

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)

Zero-Padding Driven Position Information

- CNN models pretrained on these two tasks can learn more position information than classification task.

	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

Conclusion

- This paper shows that absolute position information is implicitly encoded in convolutional neural networks.
- These results demonstrate a fundamental property of CNNs that was unknown to date.

Comments:

- The idea is natural and the experiments are not difficult because there is no comparison with sota methods.
- Maybe it is feasible to explore more on it, e.g. doing more theoretical analysis.