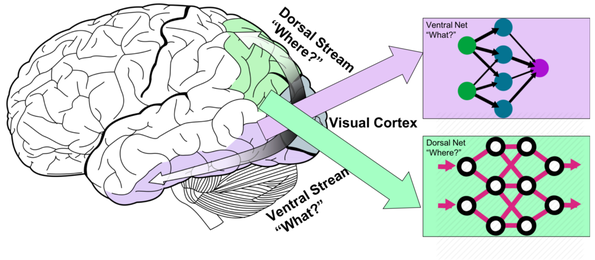
Ventral-Dorsal Neural Networks: Object Detection via Selective Attention

Mohammad K. Ebrahimpour[1] , Jiayun Li[2], Yen-Yun Yu[3], Jackson L. Reese [3], Azadeh Moghtaderi [3], Ming-Hsuan Yang[1], David C. Noelle[1]

University of California, Merced [1], University of California, Los Angles[2], Ancestry.com [3]

Brain Inspirations

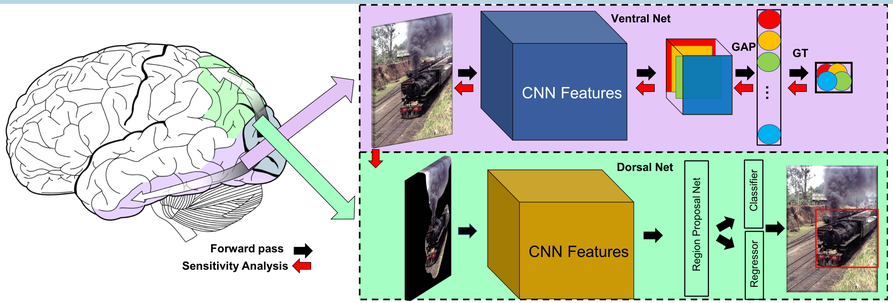
Vision Streams: Vision in the brain has two main streams called Ventral and Dorsal.



Brain Inspiration:

- Ventral Pathway: The ventral pathway from primary visual cortex, entering the temporal lobe, is dominated by "what" information.
- Dorsal Stream: The dorsal pathway, into the parietal lobe, is dominated by "where" information.
- Inspired by this structure, we propose the integration of a "Ventral Network" and a "Dorsal Network", which are complementary. Information about object identity can guide localization, and location information can guide attention to relevant image regions, improving object recognition.

Dual Networks: Ventral and Dorsal



Gestalt Total (GT):

For a given image, let $f(x, y, k)$ denote the activation of the last convolutional layer at spatial location (x, y). The "Gestalt Total" activation is defined as:

$$GT = \sum_{k} \sum_{x,y} f(x, y, k) \tag{1}$$

Sensitivity Analysis:

$$S = \frac{\partial\, GT}{\partial\, X}\bigg|_{X=I} \tag{2}$$

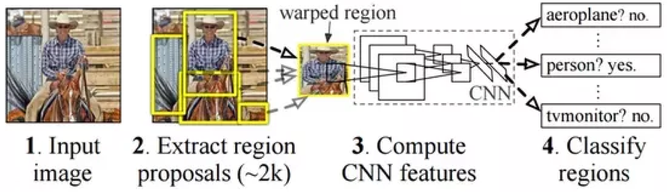
where I denotes the current image.

Loss Function:

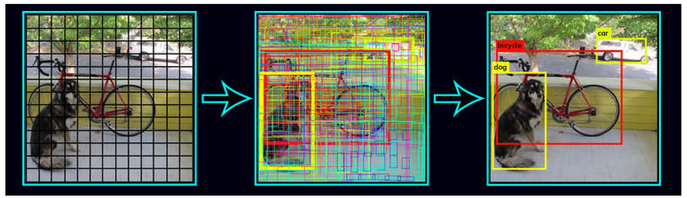
$$L(p_i, t_i) = \frac{1}{N_{cls}} \sum_i L_{cls}(p_i, p_i^*) + \lambda \frac{1}{N_{reg}} \sum_i p_i^* L_{reg}(t_i, t_i^*). \tag{3}$$

Literature Review

Region Proposal Based Framework



Single Shot Detectors



Experiments & Results

Object Detection Results:

Figure 3. VDNet Performance on Some PASCAL VOC 2007 Validation Images

mAP results on VOC 2007:

Method	Network	mAP	areo	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
Faster [25]	VGG	73.2	76.5	79	70.9	65.5	52.1	83.1	84.7	86.4	52	81.9	65.7	84.8	84.6	77.5	76.7	38.8	73.6	73.9	83	72.6
ION [2]	VGG	75.6	79.2	83.1	77.6	65.6	54.9	85.4	85.1	87	54.4	80.6	73.8	85.3	82.2	82.2	74.4	47.1	75.8	72.7	84.2	80.4
Faster [15]	Residual-101	76.4	79.8	80.7	76.2	68.3	55.9	85.1	85.3	89.8	56.7	87.8	69.4	88.3	88.9	80.9	78.4	41.7	78.6	79.8	85.3	72
MR-CNN [10]	VGG	78.2	80.3	84.1	78.5	70.8	68.5	88	85.9	87.8	60.3	85.2	73.7	87.2	86.5	85	76.4	48.5	76.3	75.5	85	81
R-FCN [4]	Residual-101	80.5	79.9	87.2	81.5	72	69.8	86.8	88.5	89.8	67	88.1	74.5	89.8	90.6	79.9	81.2	53.7	81.8	81.5	85.9	79.9
SSD300 [21]	VGG	77.5	79.5	83.9	76	69.6	50.5	87	85.7	88.1	60.3	81.5	77	86.1	87.5	83.9	79.4	52.3	77.9	79.5	87.6	76.8
SSD512 [21]	VGG	79.5	84.8	85.1	81.5	73	57.8	87.8	88.3	87.4	63.5	85.4	73.2	86.2	86.7	83.9	82.5	55.6	81.7	79	86.6	80
DSSD321 [8]	Residual-101	78.6	81.9	84.9	80.5	68.4	53.9	85.6	86.2	88.9	61.1	83.5	78.7	86.7	88.7	86.7	79.7	51.7	78	80.9	87.2	79.4
DSSD513 [8]	Residual-101	81.5	86.6	86.2	82.6	74.9	62.5	89	88.7	88.8	65.2	87	78.7	88.2	89	87.5	83.7	51.1	86.3	81.6	85.7	83.7
STDN300 [38]	DenseNet-169	78.1	81.1	86.9	76.4	69.2	52.4	87.7	84.2	88.3	60.2	81.3	77.6	86.6	88.9	87.8	76.8	51.8	78.4	81.3	87.5	77.8
STDN321 [38]	DenseNet-169	79.3	81.2	88.3	78.1	72.2	54.3	87.6	86.5	88.8	63.5	83.2	79.4	86.1	89.3	88.0	77.3	52.5	80.3	80.8	86.3	82.1
STDN513 [38]	DenseNet-169	80.9	86.1	89.3	79.5	74.3	61.9	88.5	88.3	89.4	67.4	86.5	79.5	86.4	89.2	88.5	79.3	53.0	77.9	81.4	86.6	85.5
VDNet	Resnet-101	**86.2**	95.8	98.1	98.4	65.1	94.6	90.1	96.2	71.7	72.3	54.6	97.9	95.6	89.2	90.1	93.2	69.1	89.2	82.1	93.4.6	74.0

Table 1. PASCAL VOC 2007 Test Detection Results. Note that the minimum dimension of the input image for Faster and R-FCN is 600, and the speed is less than 10 frames per second. SSD300 indicates the input image dimension of SSD is 300 × 300. Large input sizes can lead to better results, but this increases running times. All models were trained on the union of the trainval set from VOC 2007 and VOC 2012 and tested on the VOC 2007 test set.

mAP results on VOC 2012:

Method	mAP	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
HyperNet-VGG [18]	71.4	84.2	78.5	73.6	55.6	53.7	78.7	79.8	87.7	49.6	74.97	52.1	86.0	81.7	83.3	81.8	48.6	73.5	59.4	79.9	65.7
HyperNet-SP [18]	71.3	84.1	78.3	73.3	55.5	53.6	78.6	79.6	87.5	49.5	74.9	52.1	85.6	81.6	83.2	81.6	48.4	73.2	59.3	79.7	65.6
Fast R-CNN + YOLO [23]	70.7	83.4	78.5	73.5	55.8	43.4	79.1	73.1	89.4	49.4	75.5	57.0	87.5	80.9	81.0	74.7	41.8	71.5	68.5	82.1	67.2
MR-CNN-S-CNN [11]	70.7	85.0	79.6	71.5	55.3	57.7	76.0	73.9	84.6	50.5	74.3	61.7	85.5	79.9	81.7	76.4	41.0	69.0	61.2	77.7	72.1
Faster R-CNN [25]	70.4	84.9	79.8	74.3	53.9	49.8	77.5	75.9	88.5	45.6	77.1	55.3	86.9	81.7	80.9	79.6	40.1	72.6	60.9	81.2	61.5
NoC [26]	68.8	82.8	79.0	71.6	52.3	53.7	74.1	69.0	84.9	46.9	74.3	53.1	85.0	81.3	79.5	72.2	38.9	72.4	59.5	76.7	68.1
VDNet	**73.2**	85.1	82.4	73.6	57.7	61.2	79.2	77.1	85.5	54.9	79.8	61.4	87.1	83.6	81.7	77.9	45.6	74.1	64.9	80.3	73.1

Table 2. PASCAL VOC 2012 Test Detection Results. Note that the performance of VDNet is about 3% better than baseline Faster-RCNN.

Object Detection Results:

VDNet Component	Deep Network	mAP
Dorsal Net	Inception	63.1
Dorsal Net	ResNet50	71.6
Dorsal Net	ResNet101	86.2

Table 3. PASCAL VOC 2007 Test Results for Different Network Architectures

Selective Attention:

Ventral -Dorsal Networks on Github: