Structure-Regularized Attention for Deformable Object Representation

Shenao Zhang[1] Li Shen[2] Zhifeng Li[2] Wei Liu[2]

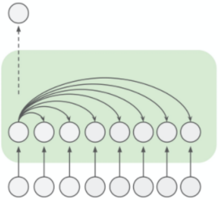
[1] Georgia Institute of Technology [2] Tencent AI Lab

Background

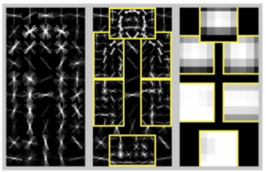
Previous self-attention methods [1] have the problems:

- Lack of structural information: Each position on feature maps attends over all other positions.
- Expensive Computation: The complexity is quadratic to input size (*$O(H^2W^2)$*) for images).

Deformable Object: Deformable object intrinsically has structural dependencies and may require or highly benefit from using the structure prior of data [2].



Self Attention



Deformable Part Model [2]

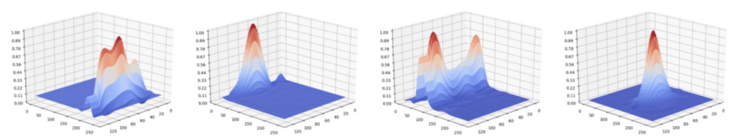
Motivation

Representing deformable objects by modeling structural dependencies the data intrinsically has.

Hypothesis: Structural Factorization

Context Modeling by Structural Factorization

- Project input onto multiple diversified modes (subspaces).
- Generate contextual features for each pixel by a combination of the information derived from every mode.



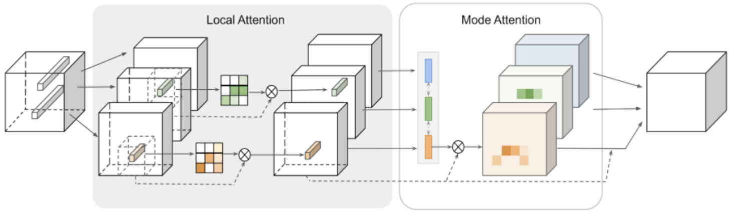
Spatial distributions of high activations on four modes.

Method

Structure-Regularized Attention (StRA)

The contextual feature for node x_i is formulated as a combination of the information derived from each mode $\mathbf{y}_i := \bigcup \mathbf{y}_i^g$

$$\mathbf{y}_i^g = r_{ig} \cdot \mathbf{z}_g, \quad r_{ig} = \gamma(\mathbf{s}_i^g, \mathbf{z}_g)$$



Local Attention projects input onto a set of feature subspaces and simultaneously capture local correlation within the neighborhood.

$$\mathbf{s}_i = \sum_{j \in \mathcal{N}_K(i)} a_{ij} u(\mathbf{x}_j), \; a_{ij} = \sigma_m\left(\omega(\mathbf{x}_i)_j + \nu(\mathbf{x}_j)\right)$$

Mode Attention generates contextual features by modeling relations between nodes and modes, as well as between modes. Each mode is expected to be responsible for the feature distribution of one distinct component.

$$r_{ig} = \gamma(\mathbf{s}_i^g, \mathbf{z}_g) = \sigma(\langle \mathbf{s}_i^g, \mathbf{z}_g \rangle) \qquad \mathbf{z}'_g = \sum_{j=1}^{G} \sigma_m(\langle \mathbf{z}_g, \mathbf{z}_j \rangle) \cdot \mathbf{z}_j$$

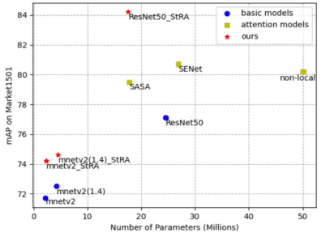
Where the modal vector $\mathbf{z}_g$ is generated through a parameterized function $\xi_g : \mathbf{S}_g \mapsto \mathbf{z}_g$, denoting the intrinsic properties of the mode.

Discussion: The design of correlating nodes to multiple modes is related to soft-clustering and mixture models. The iterative process is substituted by forward and backward propagations, where the associated parameters are learned by gradient descent.

[1] Ashish Vaswani et al. "Attention is all you need". In Advances in neural information processing systems. 2017
[2] Pedro Felzenszwalb et al. "A discriminatively trained, multiscale, deformable part model". In IEEE conference on computer vision and pattern recognition (pp. 1-8). 2008

Experiments

Network	mAP	Rank1	FLOPs
ResNet50 [6]	77.1	90.6	4.05G
SASA [14]	79.5	92.3	3.19G
SENet [7]	80.7	93.3	4.49G
Non-local [22]	80.2	91.9	7.28G
ResNet50_StRA	**84.1**	**93.8**	**3.17G**
mnetv2 [15]	71.7	88.7	**370M**
mnetv2_StRA	74.2	89.3	**370M**
mnetv2(1.4) [15]	72.5	89.0	680M
mnetv2(1.4)_StRA	**74.6**	**89.9**	720M

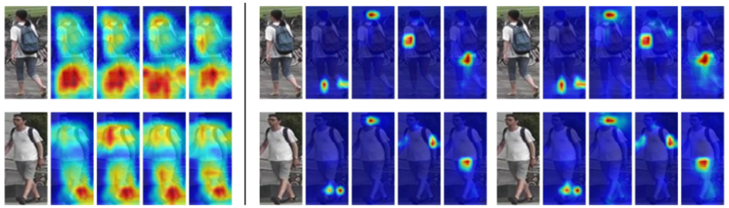


(Left) Comparison on Market 1501. (Right) Model size vs mAP.

Model	Component			mAP	Rank1
	Local Attn.	Mode Attn. w/o Interact.	Mode Attn.		
ResNet50				77.1	90.6
StRAttention	✓			79.9	92.3
	✓	✓		83.3	93.4
	✓	✓	✓	**84.1**	**93.8**

Method	mAP	Rank1	Params	FLOPs
Conv	77.1	90.6	24.6M	4.05G
SASA [15]	79.5	92.3	17.8M	3.19G
Conv + Mode	82.1	93.2	24.6M	4.06G
Group Conv + Mode	82.8	93.3	18.4M	3.26G
SASA + Mode	83.0	93.6	17.8M	3.19G
Local + Mode (ours)	84.1	93.8	17.6M	3.17G

Ablation studies on (left) module components, and (right) Mode Attention.



Local Attention variant Attention coefficients Module outputs

Conclusion

- Introduce a novel module which effectively captures the long-range dependency through the use of **structural factorization** on data.
- The mechanism encourages learning structure-distributed representations.
- The structure prior is assumed to be spatial factorization, and it would be interesting to generalize to disentangled factors.