Structure-Regularized Attention for Deformable Object Representation

Shenao Zhang
Georgia Institute of Technology
shenao@gatech.edu

Li Shen
Tencent AI Lab
lshen.lsh@gmail.com

Zhifeng Li
Tencent AI Lab
michaelzfli@tencent.com

Wei Liu
Tencent AI Lab
wl2223@columbia.edu

Abstract

Capturing contextual dependencies has proven useful to improve the representational power of deep neural networks. Recent approaches that focus on modeling global context, such as self-attention and non-local operation, achieve this goal by enabling unconstrained pairwise interactions between elements. In this work, we consider learning representations for deformable objects which can benefit from context exploitation by modeling the structural dependencies that the data intrinsically possesses. To this end, we provide a novel structure-regularized attention mechanism, which formalizes feature interaction as structural factorization through the use of a pair of light-weight operations. The instantiated building blocks can be directly incorporated into modern convolutional neural networks, to boost the representational power in an efficient manner. Comprehensive studies on multiple tasks and empirical comparisons with modern attention mechanisms demonstrate the gains brought by our method in terms of both performance and model complexity. We further investigate its effect on feature representations, showing that our trained models can capture diversified representations characterizing object parts without resorting to extra supervision.

1 Introduction

Attention is capable of learning to focus on the most informative or relevant components of input and has proven to be an effective approach for boosting the performance of neural networks on a wide range of tasks [10, 9, 20, 7, 22]. Self-attention [20] is an instantiation of attention which weights the context elements by leveraging pairwise dependencies between the representations of query and every contextual element. The ability of exploiting the entire context with variable length has allowed it to be successfully integrated into the encoder-decoder framework for sequence processing. [22] interprets it as non-local means [2], and adapts it to convolutional neural networks. However, it is computationally expensive where the complexity is quadratic with respect to input length (e.g., spatial dimensions for image and spatial-temporal dimensions for video sequence). The approaches capture long-range association by allowing each node (e.g., a pixel on feature maps) to attend over every other positions, forming a complete and unconstrained graph which may be intractable for extracting informative patterns in practice. Relative position embedding has proven useful in alleviating the issue [1, 16] but the structural information specific to tasks is not effectively exploited.

In this work, we aim to address the visual tasks of representing naturally deformable objects [5, 17], such as face or bicycle, which have appearance and shape variance when encountering different viewing conditions or deformations, and may require or highly benefit from using the structure prior

NeurIPS 2020 Workshop on Object Representations for Learning and Reasoning, Vancouver, Canada.

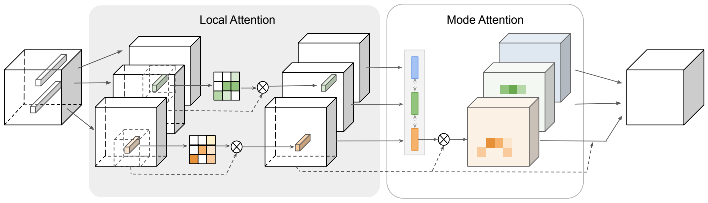

Figure 1: The illustration of the Structure-Regularized Attention Block.

data intrinsically has. Inspired from subspace algorithms [13, 18, 21], the design is built upon the hypothesis of data factorizability, i.e., projecting the data into multiple feature subspaces, defined as modes, which are expected to be more compact and typically represent certain components of objects. A set of parameterized transformations project nodes into multiple modes and capture correlation between nodes and modes, aiming to learn discriminative representations by effectively modeling structural dependencies. We achieve the goal by introducing a novel attention module, which is termed "Structure-Regularized Attention" (StRAttention), formalized as the composition of two-level operations, namely local and mode attentions (in Fig. 1). The local attention, functioning as spatial expansion on local regions. The higher-level contextual information can be accessed through the mode attention, allowing diversified contextual information to be distributed. The mechanism enables each node to attend to (theoretically) global context in a structural manner.

2 Method

We will use "pixel" and "node" interchangeably, and "mode" and "group" interchangeably in the following descriptions. Formally, let $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ denote an input, e.g., feature maps of an image, with spatial dimensions $H \times W$ and channels C, and $\mathbf{x}_i$ denote the feature on pixel x_i, where $i \in \mathcal{N}_G \equiv \{1, \cdots, HW\}$, and $\mathcal{N}_G$ denotes the set of spatial dimensions.

The contextual feature is captured by virtue of the transformation $f : \mathbb{R}^C \times \mathbb{R}^C \to [0, 1]$ [20, 22], $\mathbf{y}_i = \sum_{j \in \mathcal{N}_G} f(\mathbf{x}_i, \mathbf{x}_j) u(\mathbf{x}_j)$, where u represents the unitary transformation on a single node and f captures the pairwise relation between nodes within global context. Here f forms a complete graph in which each node can attend to every other node. It brings about the challenge of a quadratic computational complexity and memory overhead with respect to the size of $\mathcal{N}_G$ [8, 11].

The use of hierarchical structure is believed to play a critical role in capturing the statistics in images independent of learnable parameters [19]. In reality, most data can be assumed to live on low dimensional manifolds. To this end, we formalize the problem as a form of structural factorization. We want to learn a set of transformations to project data onto multiple diversified subspaces, $\Phi := \{\Phi_g\}_{g=1}^G$, where $\Phi_g : \mathcal{X} \to \mathcal{S}_g$ corresponds to the projection from the universal feature space (i.e., input feature maps) onto the g-th subspace which we call "mode" here. The corresponding output of node x_i is represented as $\mathbf{s}_i^g$. Each mode is expected to represent a certain factor the data consists of (e.g., discriminative parts), denoted by modal vectors $\mathbf{Z} = \{\mathbf{z}_g\}_{g=1}^G$ for input $\mathbf{X}$. The modal vectors are generated by integrating the projections through the function $\xi_g : \mathbf{S}_g \mapsto \mathbf{z}_g$ where $\mathbf{S}_g = \{\mathbf{s}_i^g\}_{i \in \mathcal{N}_G}$. Let $r_{ig} \in [0, 1]$ indicate the matching degree of node x_i with respect to the g-th mode, which we term *attention coefficients*. Then the context for node x_i is formulated as a combination of the information derived from each mode $\mathbf{y}_i := \bigcup \mathbf{y}_i^g$, and

$$\mathbf{y}_i^g = r_{ig} \cdot \mathbf{z}_g, \quad r_{ig} = \gamma(\mathbf{s}_i^g, \mathbf{z}_g), \tag{1}$$

where the information between modes can be further correlated through a function $\rho : \mathcal{Z} \to \mathcal{Z}$ that captures the relation between modal vectors and propagates such higher-level context to each node.

Local Attention. The transformation onto the g-th subspace Φ_g can be implemented by convolutions. The index g is omitted for simplification. We propose an alternative which is defined as,

$$\mathbf{s}_i = \sum_{j \in \mathcal{N}_K(i)} a_{ij} u(\mathbf{x}_j), \; a_{ij} = \sigma_m \left(\omega(\mathbf{x}_i)_j + \nu(\mathbf{x}_j) \right), \tag{2}$$

where σ_m denotes the softmax function, $\omega : \mathbb{R}^C \to \mathbb{R}^{K \cdot K}$ and $\nu : \mathbb{R}^C \to \mathbb{R}$. The affinity matrix $A_i = \{a_{ij}\}_{j \in \mathcal{N}_K(i)} \in [0, 1]^{K \times K}$ is expected to generate a proper *data-dependent* local softmask on the $K \times K$ neighbourhood of each node x_i for local context aggregation.

Table 1: Comparison on Market1501.

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

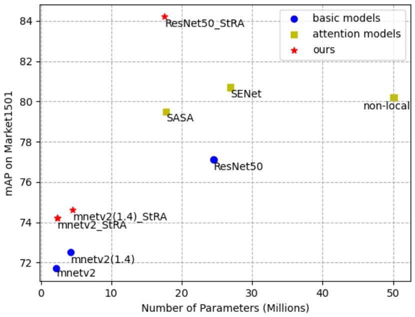

Figure 2: Model size vs mAP.

Mode Attention. A deformable object can be effectively described by a combination of representations towards different parts [5]. We expect each mode responsible for the feature distribution of one distinct component whose intrinsic properties are described by modal vectors, *i.e.*, ξ_g is realized by mean features (averaging over the local attention output $\mathbf{S}_g$) or centroid features (averaging over representative nodes). The attention coefficient r_{ig} in (1) is then measured by inner product between the corresponding feature vector for node x_i and the modal vector:

$$r_{ig} = \gamma(\mathbf{s}_i^g, \mathbf{z}_g) = \sigma(\langle \mathbf{s}_i^g, \mathbf{z}_g \rangle). \tag{3}$$

σ denotes the gating, which can be defined as either softmax or sigmoid function, representing that the relation is modelled in a mutually exclusive or independently manner. Mode interaction $\rho : \mathcal{Z} \to \mathcal{Z}$ can be conveniently achieved by, $\mathbf{z}'_g = \sum_{j=1}^{G} \sigma_m(\langle \mathbf{z}_g, \mathbf{z}_j \rangle) \cdot \mathbf{z}_j$. The updated $\mathbf{Z}$ of across-mode interactions can substitute that in (3), which is complementary (added) to the output of local attention. Detailed implementation and module schema can be found in the Appendix.

Discussion. The design of correlating nodes to multiple modes is related to soft-clustering and mixture models [12] which learn clusters by updating central vectors and node assignments iteratively through Expectation-Maximization algorithm [4]. Such an iterative process is substituted by forward and backward propagations in the framework where the associated parameters are learned by gradient descent. During inference the modal vectors and the attention coefficients are computed once, which is more efficient and suitable for neural network paradigms.

3 Experiments

To validate the effectiveness of the proposed StRA, we conduct experiments on two types of widely studied deformable objects: human body and human face. We will focus on three tasks: person re-identification (ReID), face recognition and facial expression recognition.

3.1 Human Body

We evaluate the method mainly on the Market1501 dataset[23] of the person ReID task. Model comparison is conducted on two widely used backbone architectures ResNet [6] and MobileNetV2 [15] in terms of both performance and model complexity. The results in Table 1 show that our method outperforms the baseline and other attention networks by a large margin on both metrics with the highest efficiency (i.e., Flops). The comparison of parameter sizes (in Fig.2) shows that our method achieves the best trade-off between performance and model complexity on this task.

3.2 Face

Table 2: Scalability on face recognition (performance %).

Dataset	Network	LFW	CFP-FP	CPLFW	CALFW	AgeDB-30
Medium	ResNet50	99.1	94.4	82.0	89.1	93.1
	ResNet50-StRA	99.1	95.0	82.4	89.5	93.3
Large	ResNet50	99.5	97.3	87.2	89.9	93.9
	ResNet50-StRA	99.7	97.5	88.0	91.8	94.4

Face Recognition. Challenges of face recognition may come from various factors, e.g., variations in pose, expression and illumination. We conduct the experiments to assess the scalability of the method, based on standard classification loss, i.e., softmax loss, for training. The attention modules are used at the last stage of

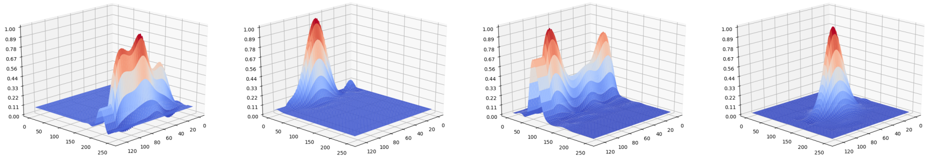

Figure 4: Spatial distributions of high activations (i.e., attention coefficients) on the four modes. Higher peaks indicate that more samples are focusing on the corresponding locations.

the architecture. The results in Table 2 show that the scalability of the method, i.e., it can consistently enhance the representational power of networks benefiting from increased dataset scale.

Facial Expression Recognition. Facial expressions explicitly correspond to the deformation of discriminative part/landmarks [3]. With ResNet as backbone, our model can achieve 73.2% accuracy on test set and 71.5% on the public validation set, outperforming the baseline performance, 71.5% and 69.9%, by a large margin (1.7% and 1.6%) respectively.

4 Interpretation and Discussion

Activation of structure-distributed representations. We present the examples of activation visualization (i.e., pixel-wise magnitude on feature maps) for the four modes at the stage 5-1 of ResNet50_StRA, and compare three types of activations in the Fig. 3, *i.e.*, the output of local attention variant running only with local attention, the attention coefficients derived from the mode attention and the final output of the module. The difference between the heatmaps generated by local attention only variant is marginal. Although the multi-head transformations are assumed to detect distinct patterns, the diversity between groups is still difficult to achieve in practice. In contrast, incorporating the regularization of the mode attention unit can diversify feature learning on different groups and encourage exciting features corresponding to discriminative parts of objects, realizing the learning of effective structure-distributed representations.

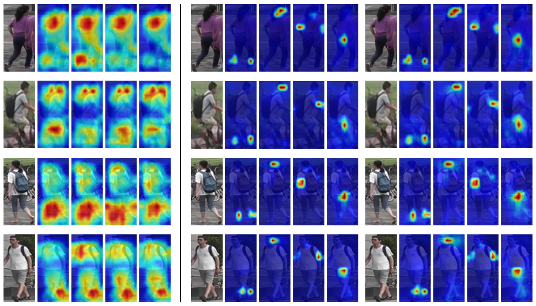

Figure 3: Visualization for the activations of Local Attention variant, attention coefficients in *Mode Attention* and the module outputs.

We also provide the spatial distribution of the four modes (*i.e.*, statistics of attention coefficients with respect to spatial locations across a set of samples) in Fig. 4, showing that nodes (pixels) tend to be projected and highly correlated with certain modes. It demonstrates that such structural regularization mechanism encourages capturing structure-distributed features for deformable objects in a factorized manner, which potentially provides interpretable features.

5 Conclusion

In this work we introduced a novel attention module which can effectively capture the long-range dependency for deformable objects through the use of structural factorization on data. The comprised components, i.e., local attention and mode attention, are complementary for capturing the informative patterns and the combination is capable of improving the discriminative power of models.

The proposed mechanism encourages learning structure-distributed representations which are realized by regularizing information flow conditioned on feature space factorization. The structure prior is assumed to be spatial factorization in the work, where part components are learned in an unsupervised manner (without the need of extra supervision). It would be interesting to generalize to disentangle factors (e.g., describing the factors of age and emotions for face perception) which would widely benefit the representation learning for generative models.

6 Broader Impact

The work presented an insight that representation learning for deformable objects can strongly benefit from exploiting the prior knowledge the data consists of, though recent progress on general network architectures has shown to be transferable to this kind of data. The experiments are conducted on the tasks related to face and human bodies, as they are typical examples of deformable objects. Such tasks may incur some concerns on privacy. The contribution of properly modeling structural dependencies is not confined to such certain task.

We hope that the work can encourage research on network architectures and representation learning for better modeling structural dependencies which will potentially facilitate research on disentangling and interpretable features. The effectiveness of the method may be obstructed by the given subspace number (analogous to subspace number for subspace segmentation methods), especially when the number is difficult to estimate in advance or needs gradually increase during training, while the issue may be addressed from the direction of incorporating network evolution in a nonparametric manner.

References

[1] Irwan Bello, Barret Zoph, Ashish Vaswani, Jonathon Shlens, and Quoc V Le. Attention augmented convolutional networks. In *ICCV*, 2019.

[2] Antoni Buades, Bartomeu Coll, and Jean-Michel Morel. Non-local means denoising. *Image Processing On Line*, 1:208–212, 2011.

[3] Jeffrey F Cohn, Zara Ambadar, and Paul Ekman. Observer-based measurement of facial expression with the facial action coding system. *The handbook of emotion elicitation and assessment*, 1(3):203–221, 2007.

[4] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[5] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable part model. In *CVPR*, 2008.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[7] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[8] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

[9] L. Itti and C. Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2001.

[10] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[11] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9167–9176, 2019.

[12] Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 84. M. Dekker New York, 1988.

[13] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *Acm Sigkdd Explorations Newsletter*, 2004.

[14] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. Stand-alone self-attention in vision models. In *NeurIPS*, 2019.

[15] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.

[16] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[17] James Thewlis, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks by factorized spatial embeddings. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5916–5925, 2017.

[18] Michael E Tipping and Christopher M Bishop. Mixtures of probabilistic principal component analyzers. *Neural computation*, 11(2):443–482, 1999.

[19] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, 2018.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[21] Rene Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1945–1959, 2005.

[22] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018.

[23] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.