Deep Attentional Structured Representation Learning for Visual Recognition

Krishna Kanth Nakka and Mathieu Salzmann
CVLAB, EPFL
ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
IC COMPUTER AND COMMUNICATION SCIENCES

Goal

Incorporate attention into deep structured-representation architectures

Contribution

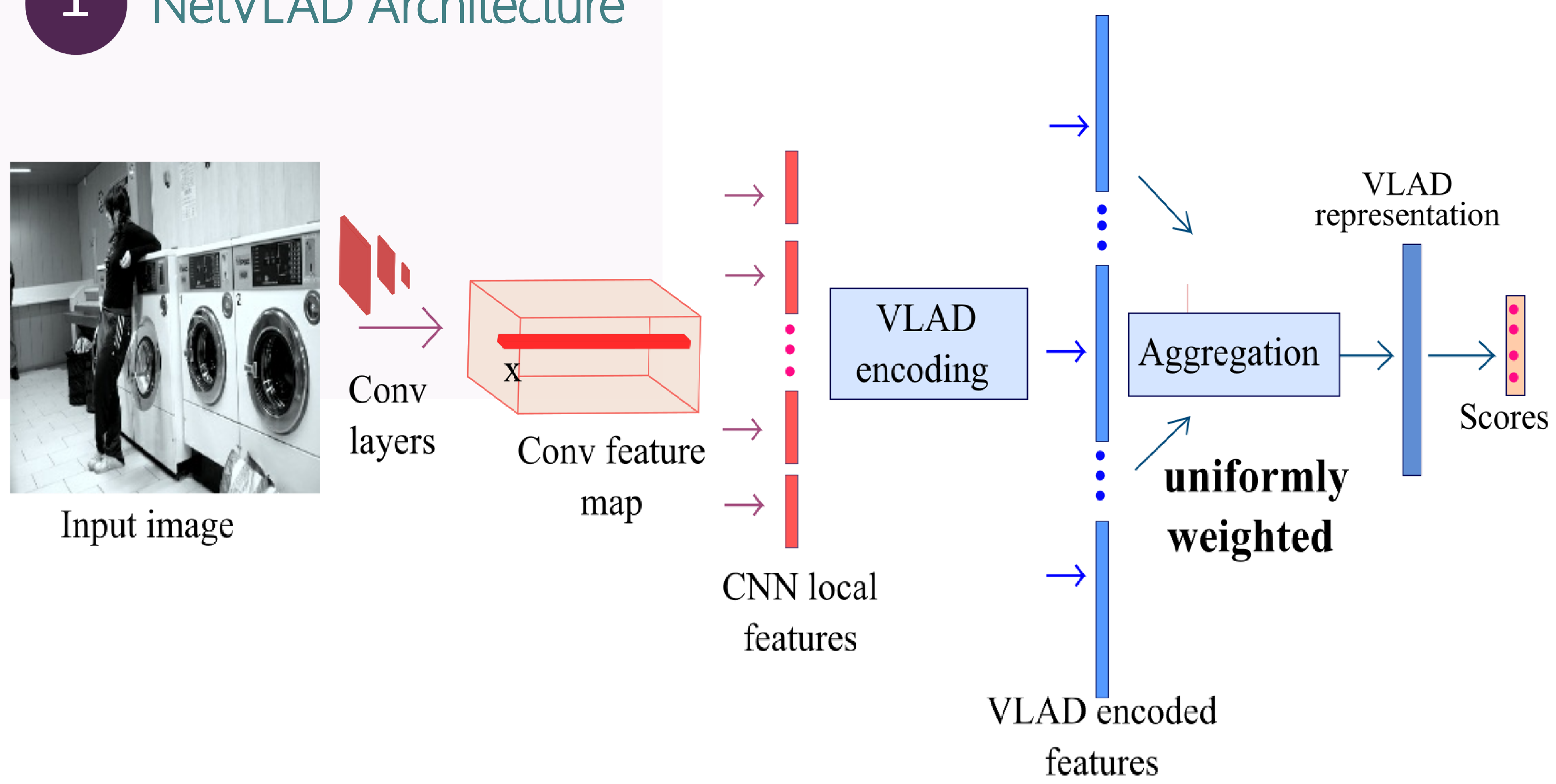
An attentional structured representation learning framework that incorporates an image-specific attention mechanism

Results

Improvement across various recognition tasks: scene recognition, fine-grained categorization.

Standard Structured Representation Architecture

1 NetVLAD Architecture



Standard structured-representation network [1]

2 Uniformly Weighted Feature Aggregation

Regions from irrelevant classes (e.g., Person) contribute equally as other regions, thereby reducing the discriminative power of the structured descriptor

All local descriptors are weighted equally in the aggregation process

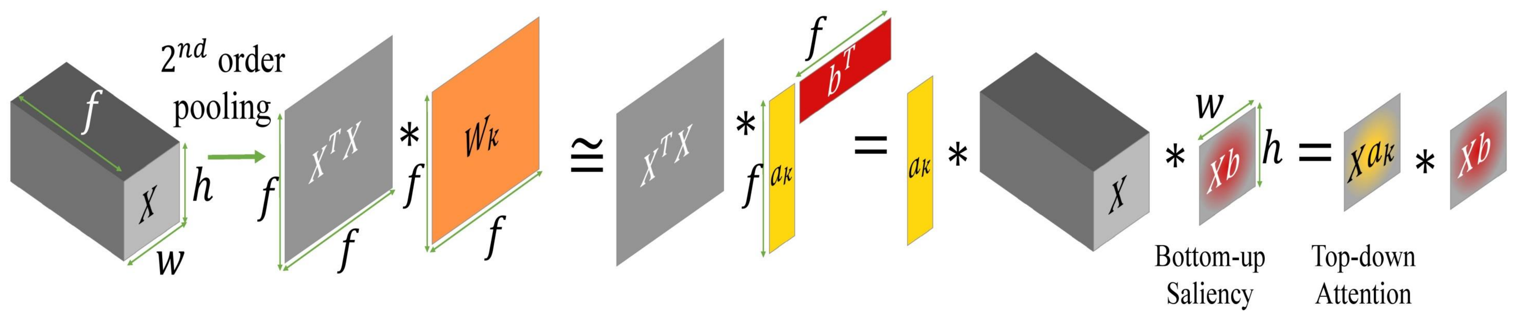
Attention-Aware Structured Representation: Incorporating top-down and bottom-up information

Main Contribution:

We incorporate Attention within the feature aggregation process

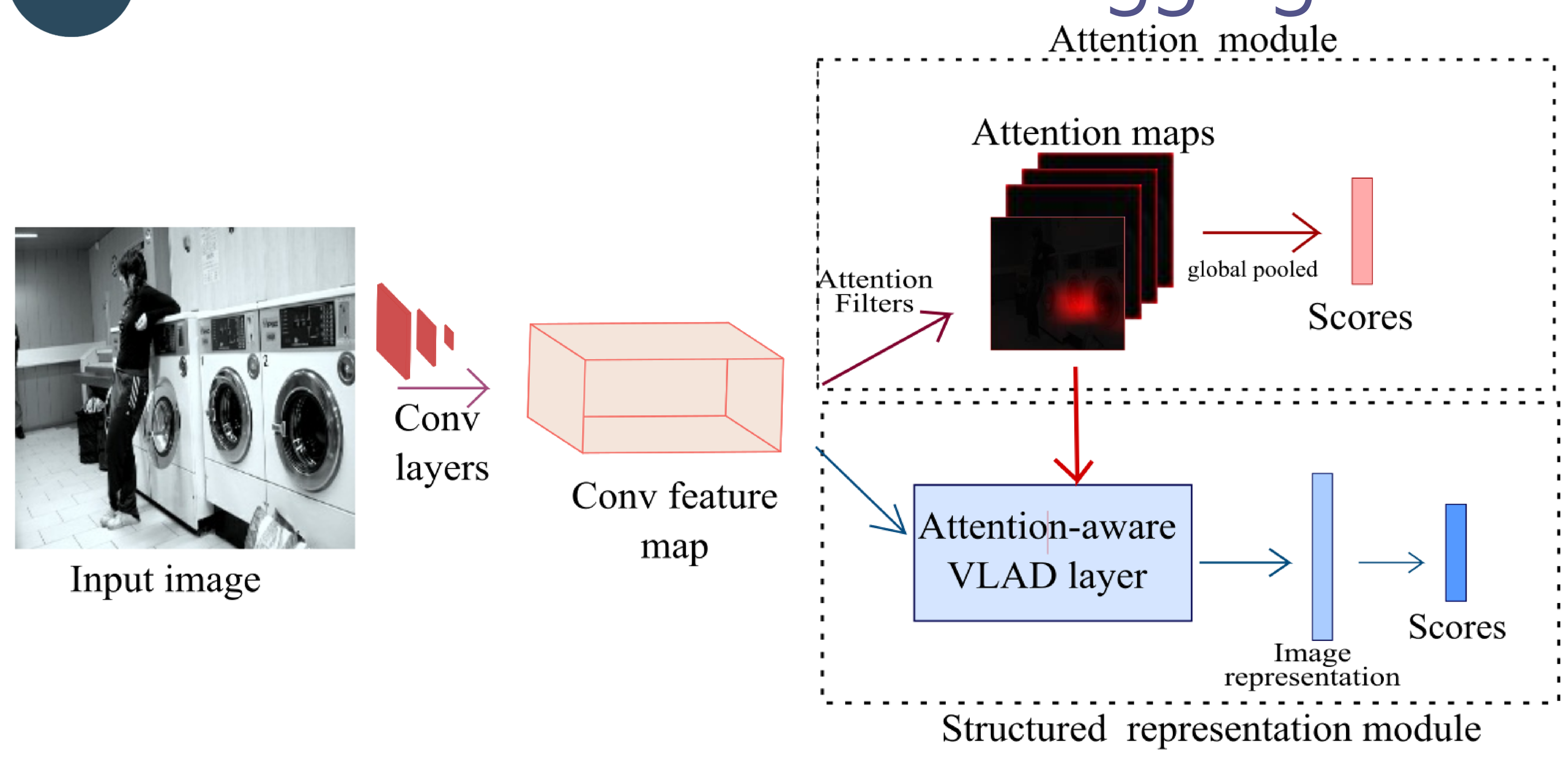
1 Attention Module

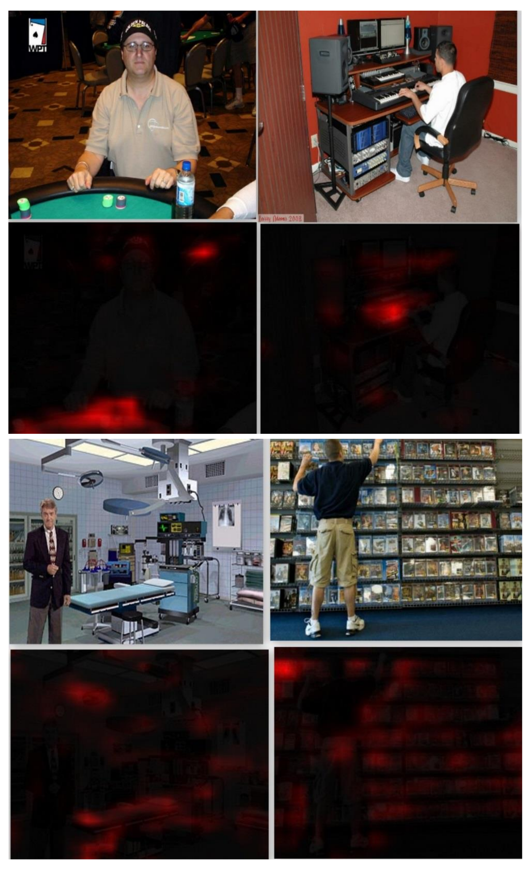
- Generates class-specific spatial attention maps from final feature map
- Combines top-down attention with bottom-up saliency



Attention framework[2]

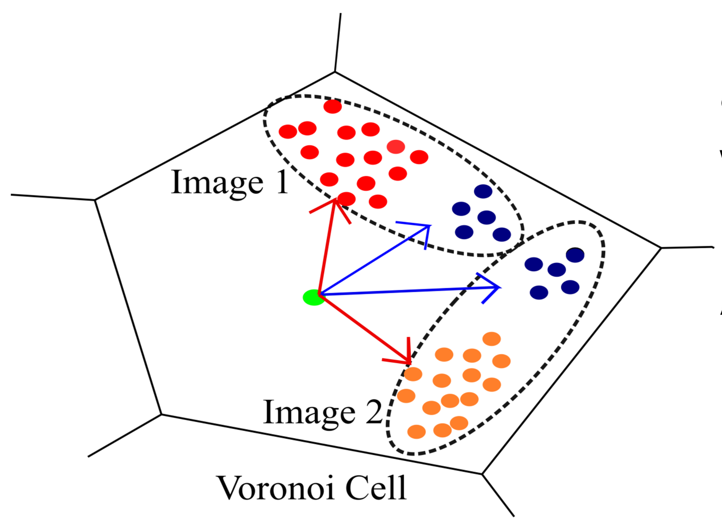
2 Attention-Aware Feature Aggregation





3 Geometric Interpretation of Attention

Blue – high attention descriptors
Red – Low attention descriptors



Ignoring attention would yield residual vectors pointing in almost opposite directions, shown with red arrows

Attention-aware aggregation produces residual vectors with high cosine similarity, shown with blue arrows

Experiments: Impact of Attention into Structured Representation

Attentional Structured Pooling Scheme

Pooling	Anno.	Birds	Cars	Aircrafts
VGG-16	✓	79.9	88.4	86.9
Attention	✓	77.2	90.3	85.0
NetBoW	✓	74.4	89.1	85.6
Attentional NetBoW	✓	80.5	91.2	89.3
NetVLAD	✓	82.4	89.8	88.0
Attentional NetVLAD	✓	85.5	93.5	89.2

+ With bounding box information

Pooling	Anno.	Birds	Cars	Aircrafts	MIT-Indoor
VGG-16	-	76.0	82.8	82.3	76.6
Attention	-	77.0	87.4	81.4	77.2
NetBoW	-	68.9	85.2	79.9	76.1
Attentional NetBoW	-	76.9	90.6	88.3	76.6
NetVLAD	-	80.6	89.4	86.4	79.2
Attentional NetVLAD	-	84.3	92.8	88.8	81.2

+ Without bounding box information

Comparison with State of the Art

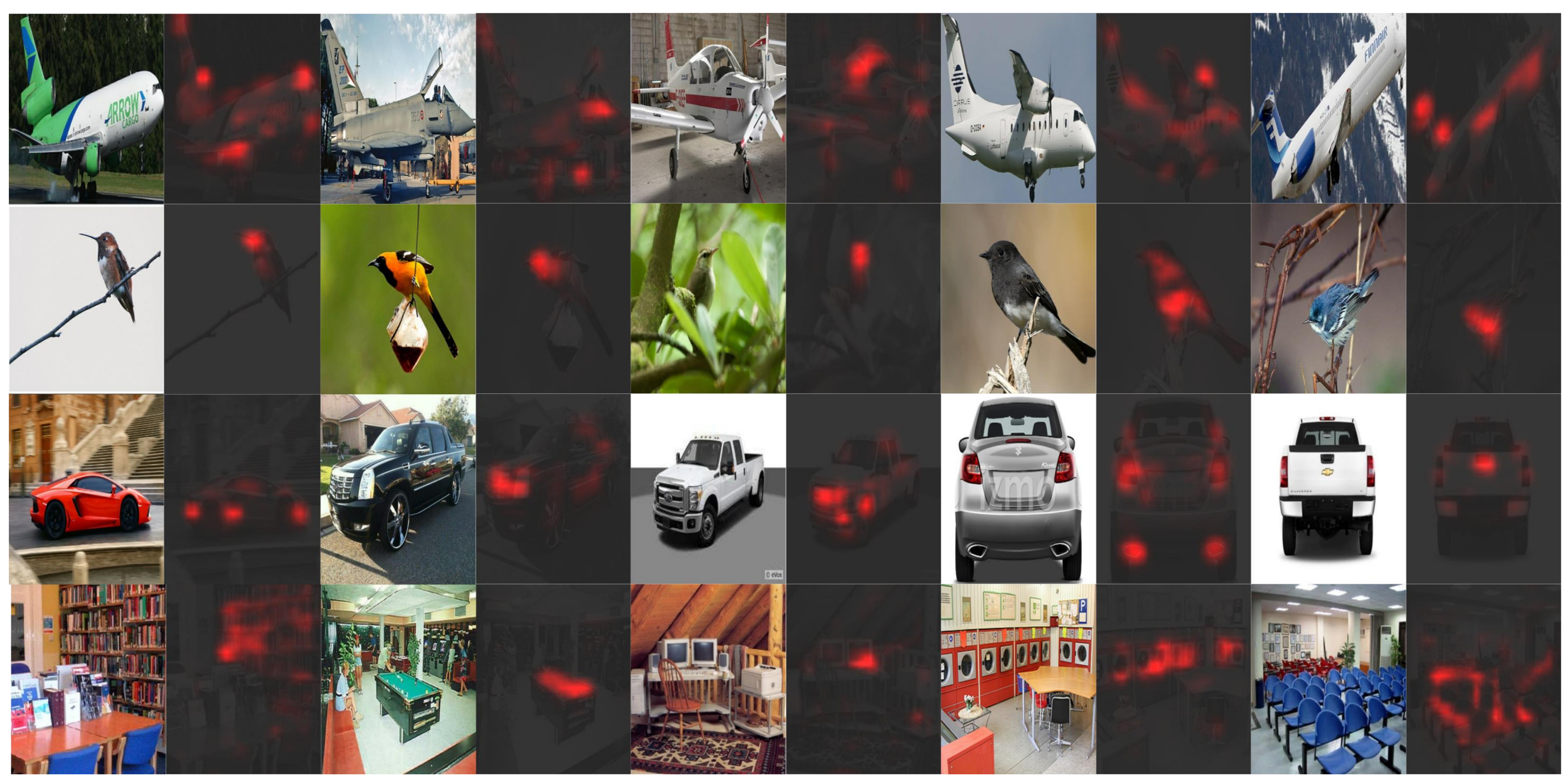
MIT-Indoor Scene Dataset

Method	Birds
Deep FisherNet	76.5
CBN	77.6
NetVLAD	79.1
H-Sparse	79.5
B-CNN	79.7
FV+FC	81.0
MFAFVNet	81.1
Ours	81.2

Fine-Grained Datasets

Pooling	Anno.	Birds	Cars	Aircrafts
MG-CNN	✓	83.0	-	86.6
B-CNN	✓	85.1	-	-
PA-CNN	✓	82.8	92.8	-
Mask-CNN	✓	85.4	-	-
MDTP	✓	-	92.6	88.4
Ours	✓	85.5	93.5	89.2
KP	-	86.2	92.4	86.9
Boost-CNN	-	86.2	92.1	88.5
Imp. B-CNN	-	85.8	92.0	88.5
alpha-pooling	-	85.8	92.0	88.5
RA-CNN	-	84.1	92.5	88.2
MA-CNN	-	86.5	92.8	88.9
Ours	-	84.3	92.8	88.8

Resulting Attention Maps



Our method is able to localize discriminative parts of birds (tail, beak), aircrafts (engine, landing gear) and cars (lights, logo).

References

1. Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: CNN architecture for weakly supervised place recognition. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5297–5307, 2016
2. Rohit Girdhar and Deva Ramanan. Attentional pooling for action recognition. In Advances in Neural Information Processing Systems, pages 33–44, 2017