Learning Visual Context by Comparison

Minchul Kim al., Lunit.

ECCV 2020 spotlight paper

2020-10-26
MI2RL
Kyuri Kim

Introduction

- Contributions

- We present Attend-and Compare Module (ACM) for capturing the difference between an object of interest and its corresponding context. (**The necessity of comparison between related regions in an image**)

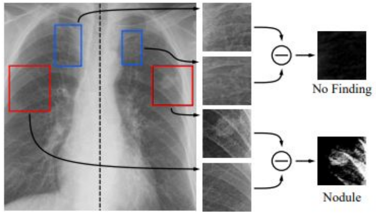


Fig. 1: An example of a comparison procedure for radiologists. Little differences indicate no disease (blue), the significant difference is likely to be a lesion (red).

(1) We propose a novel context module called ACM that explicitly compares different regions, following the way radiologists read chest X-rays.

(2) The proposed ACM captures multiple comparative self-attentions whose difference is beneficial to recognition tasks.

(3) We demonstrate the effectiveness of ACM on three chest X-ray datasets and COCO detection & segmentation dataset with various architectures.

Related Work

- Context Modeling

In the visual recognition domain recent self-attention mechanisms,

- SENet(Squeeze and excitation networks) - CVPR 2018
- A Style-based Re-calibration Module (SRM) (Lunit)
- Convolutional block attention module (CBAM) - ECCV 2018 (Lunit)

Works that explicitly tackle the problem of using context stem from using pixel-level pairwise relationships,

- Non-local neural networks (NL) - CVPR 2018 (Kaiming He)
- Global-Context network (GC) - 2019 (Microsoft)
- Criss-cross attention (CC) - IEEE TPAMI 2020 & ICCV 2019

Related Work

- Context Modeling

- SENet(Squeeze and excitation networks)
  - *“ Our goal is to improve the representational power of a network by explicitly modelling the interdependencies between the channels of its convolutional features. ”*
  - Learns to model channel-wise attention using the spatially averaged feature.

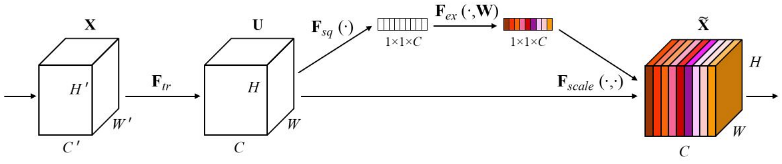


Figure 1: A Squeeze-and-Excitation block.

- Non-local neural networks (NL)
  - Calculate pixel-level pairwise relationship weights and aggregate (weighted average) the features from all locations according to the weights.

Weighted sum of **All** pixels with **similarity**

$$y_i = \frac{1}{C(x)} \sum_j \boxed{f(x_i, x_j)} \boxed{g(x_j)}$$

$$g(x_j) = W_g x_j$$

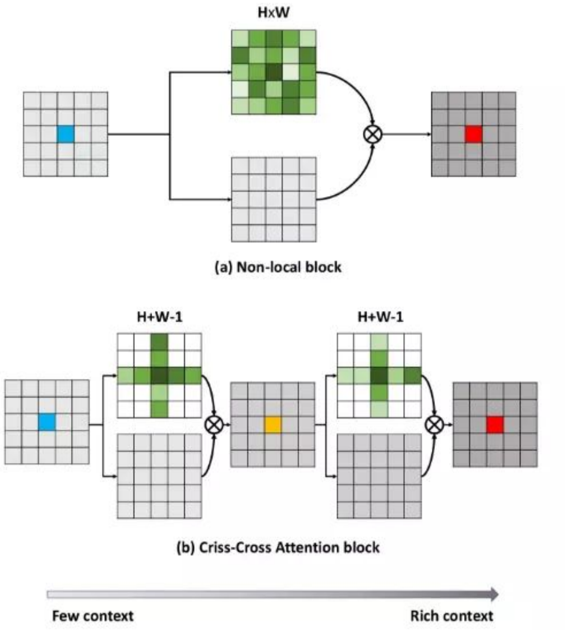
weight*input pixel

[Squeeze-and-Excitation Networks] https://arxiv.org/abs/1709.01507
[Non-local Neural Networks] https://arxiv.org/pdf/1711.07971.pdf , https://blog.lunit.io/2018/01/19/non-local-neural-networks/

Related Work

- Context Modeling

- Criss-cross attention (CC)
  - For semantic segmentation reduces the computation cost of NL by replacing the pairwise relationship attention maps with criss-cross attention block which considers only horizontal and vertical directions separately.
  - NL and CC explicitly model the pairwise relationship between regions with affinity metrics, but the qualitative results in demonstrate a tendency to aggregate features only among foreground objects or among pixels with similar semantics.



[CCNet: Criss-Cross Attention for Semantic Segmentation] https://arxiv.org/pdf/1811.11721.pdf

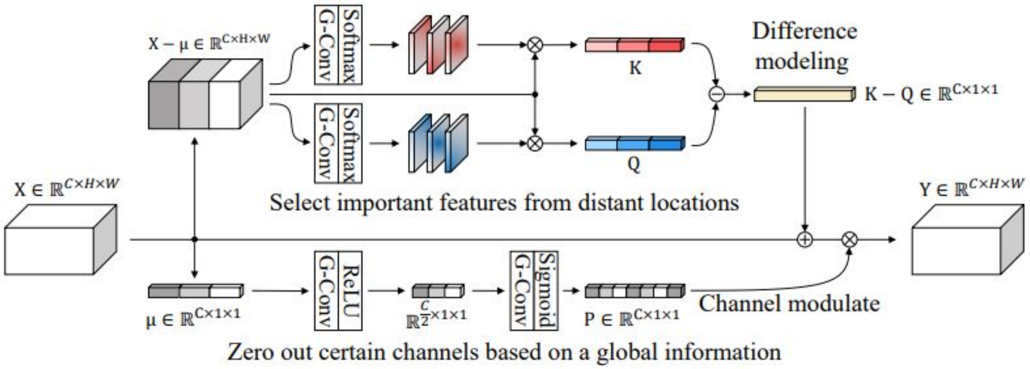
Method

- Attend-and-Compare Module

1. Overview

   - Attend-and-Compare Module (ACM) extracts an object of interest and the corresponding context to compare, and enhances the original image feature with the comparison result.

$$Y = f_{\mathrm{ACM}}(X) = P(X + (K - Q)), \qquad (1)$$



Method

- Attend-and-Compare Module

2. Components of ACM

- Object of Interest and Corresponding Context

$$K = \sum_{i,j \in H,W} \frac{\exp(W_K X_{i,j})}{\sum_{H,W} \exp(W_K X_{h,w})} X_{i,j}, \quad (2)$$

① 1x1 conv for single ch.
② softmax for Normalize
③ Apply weighted avg.

- Channel Re-calibration

$$P = \sigma \circ \mathrm{conv}_2^{1\times1} \circ \mathrm{ReLU} \circ \mathrm{conv}_1^{1\times1}(\mu), \quad (3)$$

Method

- Attend-and-Compare Module

2. Components of ACM

- Group Operation

$$K^g = \sum_{i,j \in H,W} \frac{\exp(W_K^g X_{i,j}^g)}{\sum_{H,W} \exp(W_K^g X_{h,w}^g)} X_{i,j}^g, \quad (4)$$

- group conv. (#G)
- K= [K1,…,KG]

- Loss Function

$$\ell_{\text{orth}}(K,Q) = \frac{K \cdot Q}{C}, \quad (5)$$

$$\ell_{\text{task}} + \lambda \sum_{m}^{M} \ell_{\text{orth}}(K_m, Q_m), \quad (6)$$

- C= channel num.

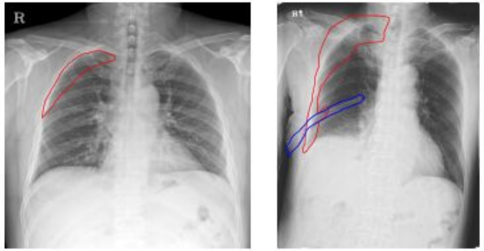
- Placement of ACMs

Experiments

- Experiment Dataset

1) Emergency-Pneumothorax (Em-Ptx) and Nodule (Ndl) datasets for lesion localization in chest X-rays
2) Chest X-ray 14 dataset for multi-label classification
3) COCO 2017 dataset for object detection and instance segmentation.

〈Emergency-Pneumothorax (Em-Ptx) Datasets〉



(a) Emergency (b) Non-emergency

Table 1: Results on Em-Ptx dataset. Average of 5 random runs are reported for each setting with standard deviation. RN stands for ResNet [14].

Method	AUC-ROC	JAFROC	Method	AUC-ROC	JAFROC
RN-50	86.78±0.58	81.84±0.64	RN-101	89.75±0.49	85.36±0.44
RN-50 + SE [15]	93.05±3.63	89.19±4.38	RN-101 + SE [15]	90.36±0.83	85.54±0.85
RN-50 + NL [36]	94.63±0.39	91.93±0.68	RN-101 + NL [36]	94.24±0.34	91.70±0.83
RN-50 + CC [17]	87.73±8.66	83.32±10.36	RN-101 + CC [17]	92.57±0.89	89.75±0.89
RN-50 + ACM	**95.35**±0.12	**94.16**±0.21	RN-101 + ACM	**95.43**±0.14	**94.47**±0.10

Results

Table 2: Performance with respect to varying module architectures and hyper-parameters on Em-Ptx dataset. All the experiments are based on ResNet-50.

Module	AUC-ROC	JAFROC
None	86.78±0.58	81.84±0.64
$X + (K - Q)$	94.25±0.31	92.94±0.36
PX	87.16±0.42	82.05±0.30
$P(X + K)$	94.96±0.15	93.59±0.24
$P(X + (K - Q))$	**95.35**±0.12	**94.16**±0.21

(a) Ablations on K, Q and P.

#groups	AUC-ROC	JAFROC
8	90.96±1.88	88.79±2.23
32	**95.35**±0.12	**94.16**±0.21
64	95.08±0.25	93.73±0.31
128	94.89±0.53	92.88±0.53

(b) Ablations on number of groups.

λ	AUC-ROC	JAFROC
0.00	95.11±0.20	93.87±0.20
0.01	95.29±0.34	94.09±0.41
0.10	**95.35**±0.12	**94.16**±0.21
1.00	95.30±0.17	94.04±0.11

(c) Ablations on orthogonal loss weight λ.

〈Nodule (Ndl) Datasets〉

Table 3: Results on Ndl dataset. Average of 5 random runs are reported for each setting with standard deviation.

Method	AUC-ROC	JAFROC
ResNet-50	87.34±0.34	77.35±0.50
ResNet-50 + SE [15]	87.66±0.40	77.57±0.44
ResNet-50 + NL [36]	88.35±0.35	80.51±0.56
ResNet-50 + CC [17]	87.72±0.18	78.63±0.40
ResNet-50 + ACM	**88.60**±0.23	**83.03**±0.24

Results

〈Chest X-ray 14 Datasets〉

Table 5: Performance in average AUC of various methods on CXR14 dataset. The numbers in the bracket after model names are the input sizes.

Modules	DenseNet121(448)	ResNet-50(448)
None	(CheXNet [29]) 84.54	84.19
SE [15]	84.95	84.53
NL [36]	84.49	85.08
CC [17]	84.43	85.11
ACM	**85.03**	**85.39**

〈COCO Datasets〉

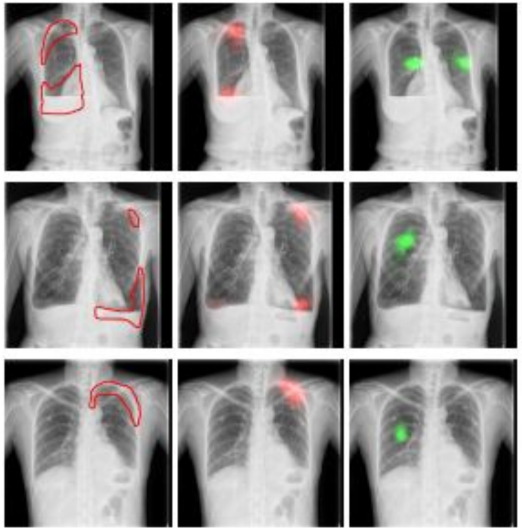
Table 6: Results on COCO dataset. All experiments are based on Mask-RCNN [13].

Method	AP^{bbox}	AP^{bbox}_{50}	AP^{bbox}_{75}	AP^{mask}	AP^{mask}_{50}	AP^{mask}_{75}
ResNet-50	38.59	59.36	42.23	35.24	56.24	37.66
ResNet-50+SE [15]	39.10	60.32	42.59	35.72	57.16	38.20
ResNet-50+NL [36]	39.40	60.60	43.02	35.85	57.63	38.15
ResNet-50+CC [17]	39.82	60.97	42.88	36.05	57.82	38.37
ResNet-50+ACM	**39.94**	**61.58**	**43.30**	**36.40**	**58.40**	**38.63**
ResNet-101	40.77	61.67	44.53	36.86	58.35	39.59
ResNet-101+SE [15]	41.30	62.36	45.26	37.38	59.34	40.00
ResNet-101+NL [36]	41.57	62.75	45.39	37.39	59.50	40.01
ResNet-101+CC [17]	**42.09**	63.21	**45.79**	**37.77**	59.98	**40.29**
ResNet-101+ACM	41.76	**63.38**	45.16	37.68	**60.16**	40.19
ResNeXt-101	43.23	64.42	47.47	39.02	61.10	42.11
ResNeXt-101+SE [15]	43.44	64.91	47.66	39.20	61.92	42.17
ResNeXt-101+NL [36]	43.93	65.44	48.20	39.45	61.99	42.33
ResNeXt-101+CC [17]	43.86	65.28	47.74	39.26	62.06	41.97
ResNeXt-101+ACM	**44.07**	**65.92**	**48.33**	**39.54**	**62.53**	**42.44**

Results

- Qualitative Results

(L) Utilize pneumothorax regions as objects of interest and normal lung regions as the corresponding context.
(R) Utilize the object of interest and the corresponding context in the natural image domain.



GT　　K map　　Q map

GT　　K map　　Q map

Conclusions

- Key idea is to extract an object of interest and a corresponding context and explicitly compare them to make the image representation more distinguishable.
- The qualitative analysis shows that ACM automatically learns dynamic relationships. The objects of interest and corresponding contexts are different yet contain useful information for the given task.