All the attention you need: Global-local, spatial-channel attention for image retrieval

Chull Hwan Song[1], Hye Joo Han[1], Yannnis Avrithis[2]

[1]Odd Concepts, INC [2]Athena RC

Introduction

- **Goal:** Learning representations using attention mechanisms for instance-level image retrieval
- Introducing the **Global-Local Attention Module (GLAM)** attached at the end of backbone
- GLAM uses all four forms of attention: **local** and **global**, **spatial** and **channel**

Contributions

- First study employing all four forms of attention, global and local, channel and spatial
- Empirical evidence of the interaction of all forms of attention
- State of the art on global descriptors without re-ranking for instance-level image retrieval

Related work on attention

Method	Local		Global		Lrn	Ret
	Spatial	Channel	Spatial	Channel		
SENet [20]		✓			✓	
ECA-Net [51]		✓			✓	
GCNet [6]		✓			✓	
CBAM [54]	✓	✓			✓	
GE [19]	✓				✓	
NL-Net [52]			✓		✓	
AA-Net [4]			✓		✓	
SAN [59]			✓		✓	
N^3Net [34]			✓		✓	
A^2-Net [9]				✓	✓	
GSoP [14]				✓	✓	
OnA [23]	✓					✓
AGeM [17]	✓					✓
CroW [24]	✓	✓				✓
CRN [25]	✓				✓	✓
DELF [29]	✓				✓	✓
DELG [5]	✓				✓	✓
Tolias *et al.* [47]	✓				✓	✓
SOLAR [28]			✓		✓	✓
Ours	✓	✓	✓	✓	✓	✓

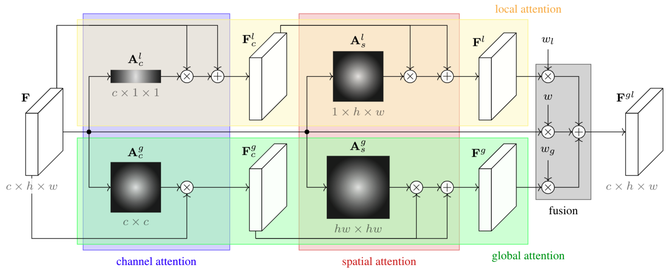
LRN: learned; RET: applied to instance-level image retrieval.

Datasets and implementation details

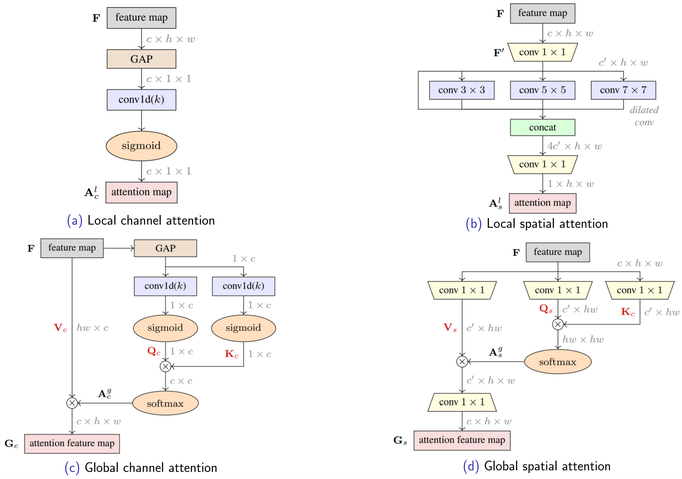
- ResNet101+GeM pooling
- Global descriptor only without re-ranking
- Training : Google Landmarks v2 clean (GLDv2-clean) by ArcFace loss
- Test set: Oxford5k, Paris6k, Revisited Oxford (ROxf)/Paris (RPar) with and without 1M distractors
- Metrics: mean average precision (mAP), mean precision at 10 (mP)
- Mini-batch samples with similar aspect ratios resized together
- Multi-resolution representation for the final feature at inference

Global-local attention module (GLAM)

- **Local attention** weights channels or locations independently
- **Global attention** captures pairwise interaction within channels or within spatial locations.
- Four attention maps used: **local channel** ($\mathbf{A}_c^l$), **local spatial** ($\mathbf{A}_s^l$), **global channel** ($\mathbf{A}_c^g$) and **global spatial** ($\mathbf{A}_s^g$).
- The input feature map $\mathbf{F}$ is weighted into local ($\mathbf{F}^l$) and global ($\mathbf{F}^g$) attention feature maps, which are fused with $\mathbf{F}$ to yield the **global-local attention feature map** $\mathbf{F}^{gl}$.



Attention Layers



Ablation study

- mAP comparison of spatial and channel variants of our local and global attention modules to the baseline.

Method	Oxf5k	Par6k	$\mathcal{R}$Medium		$\mathcal{R}$Hard	
			$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GLAM baseline	91.9	94.5	72.8	84.2	49.9	69.7
+local-channel	91.3	95.3	72.2	85.8	48.3	73.1
+local-spatial	91.0	95.1	72.1	85.3	48.3	71.9
+local	91.2	95.4	73.7	86.5	52.6	75.0
+global-channel	92.5	94.4	73.3	84.4	49.8	70.1
+global-spatial	92.4	95.1	73.2	86.3	50.0	72.7
+global	92.3	95.3	77.2	86.7	57.4	75.0
+global+local	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

- mAP comparison of our best model trained on different training sets against SOTA. Red: best results. Blue: GLAM higher than DELG on GLDv1-noisy.

Method	Train Set	dim	Oxf5k	Par6k	$\mathcal{R}$Medium		$\mathcal{R}$Hard	
					$\mathcal{R}$Oxf	$\mathcal{R}$Par	$\mathcal{R}$Oxf	$\mathcal{R}$Par
GeM-Siamese [37, 35]	SfM-120k	2048	87.8	92.7	64.7	77.2	38.5	56.3
SOLAR [28]	GLDv1-noisy	2048	–	–	69.9	81.6	47.9	64.5
DELG [5]	GLDv1-noisy	2048	–	–	73.2	82.4	51.2	64.7
GLDv2 [53]	GLDv2-clean	2048	–	–	74.2	84.9	51.6	70.3
GLAM (Ours)	NC-clean	512	77.8	85.8	51.6	68.1	20.9	44.7
	GLDv1-noisy	512	92.8	95.0	**73.7**	**83.5**	49.8	**69.4**
	GLDv2-noisy	512	93.3	95.3	75.7	86.0	53.1	73.8
	GLDv2-clean	512	**94.2**	**95.6**	**78.6**	**88.5**	**60.2**	**76.8**

State of the art comparisons

Method	Train Set	Dim	Base		Medium						Hard					
			Ox5k	Par6k	$\mathcal{R}$Oxf		+$\mathcal{R}$1M		$\mathcal{R}$Par		+$\mathcal{R}$1M		$\mathcal{R}$Oxf		+$\mathcal{R}$1M		$\mathcal{R}$Par		+$\mathcal{R}$1M	
			mAP	mAP	mAP	mP	mAP	mP	mAP	mP	mAP	mP	mAP	mP	mAP	mP	mAP	mP	mAP	mP
SPoC-V16 [2, 35]	[O]	512	53.1*	–	38.0	54.6	17.1	33.3	59.8	93.0	30.3	83.0	11.4	20.9	0.9	2.9	32.4	69.7	7.6	30.6
SPoC-R101 [35]	[O]	2048	–	–	39.8	61.0	21.5	40.4	69.2	96.7	41.6	92.0	12.4	23.8	2.8	5.6	44.7	78.0	15.3	54.4
CroW-V16 [24, 35]	[O]	512	70.8	79.7	41.4	58.8	22.5	40.5	62.9	94.4	34.1	87.1	13.9	25.7	3.0	6.6	36.9	77.9	10.3	45.1
CroW-R101 [35]	[O]	2048	–	–	42.4	61.9	21.2	39.4	70.4	97.1	42.7	92.9	13.3	27.7	3.3	9.3	47.2	83.6	16.3	61.6
MAC-V16-Siamese [36, 35]	[O]	512	80.0	82.9	37.8	57.8	21.8	39.7	59.2	93.3	33.6	87.1	14.6	27.0	7.4	11.9	35.9	78.4	13.2	54.7
MAC-R101-Siamese [35]	[O]	2048	–	–	41.7	65.0	24.2	43.7	66.2	96.4	40.8	93.0	18.0	32.9	5.7	14.4	44.1	86.3	18.2	67.7
RMAC-V16-Siamese [36, 35]	[O]	512	80.1	85.0	42.5	62.8	21.7	40.3	66.2	95.4	39.9	88.9	12.0	26.1	1.7	5.8	40.9	77.1	14.8	54.0
RMAC-R101-Siamese [35]	[O]	2048	–	–	49.8	68.9	29.2	48.9	74.0	97.7	49.3	93.7	18.5	32.2	4.5	13.0	52.1	87.1	21.3	67.4
RMAC-R101-Triplet [16, 35]	NC-clean	2048	86.1	**94.5**	60.9	78.1	39.3	62.1	78.9	96.9	54.8	93.9	32.4	50.0	12.5	24.9	59.4	86.1	28.0	70.0
GeM-R101-Siamese [37, 35]	SfM-120k	2048	**87.8**	92.7	64.7	84.7	45.2	71.7	77.2	**98.1**	52.3	**95.3**	38.5	53.0	19.9	34.9	56.3	89.1	24.7	73.3
AGeM-R101-Siamese [17]	SfM-120k	2048	–	–	67.0	–	–	–	78.1	–	–	–	40.7	–	–	–	57.3	–	–	–
SOLAR-GeM-R101-Triplet/SOS [28]	GLDv1-noisy	2048	–	–	69.9	**86.7**	53.5	**76.7**	81.6	97.1	59.2	94.9	47.9	**63.0**	29.9	**48.9**	64.5	**93.0**	33.4	**81.6**
DELG-GeM-R101-ArcFace [5]	GLDv1-noisy	2048	–	–	73.2	–	**54.8**	–	82.4	–	**61.8**	–	51.2	–	**30.3**	–	64.7	–	**35.5**	–
GeM-R101-ArcFace [53]	GLDv2-clean	2048	–	–	**74.2**	–	–	–	**84.9**	–	–	–	**51.6**	–	–	–	**70.3**	–	–	–
GLAM-GeM-R101-ArcFace baseline	GLDv2-clean	512	**91.9**	**94.5**	72.8	**86.7**	**58.1**	**78.2**	84.2	95.9	**63.9**	93.3	49.9	62.1	**31.6**	**49.7**	69.7	88.4	**37.7**	73.7
+local	GLDv2-clean	512	**91.2**	**95.4**	73.7	**86.2**	**60.5**	**77.4**	**86.5**	95.6	**68.0**	93.9	**52.6**	**65.3**	**36.1**	**55.6**	**73.7**	89.3	**44.7**	79.1
+global	GLDv2-clean	512	**92.3**	**95.3**	**77.2**	**87.0**	**63.8**	**79.3**	**86.7**	95.4	67.8	93.7	**57.4**	**69.6**	**38.7**	**57.9**	**75.0**	89.4	**45.0**	77.0
+global+local	GLDv2-clean	512	**94.2**	**95.6**	**78.6**	**88.2**	**68.0**	**82.4**	**88.5**	97.0	**73.5**	94.9	**60.2**	**72.9**	**43.5**	**62.1**	**76.8**	**93.4**	**53.1**	**84.0**

V16: VGG16; R101: ResNet101. [O]: Off-the-shelf (pre-trained on ImageNet). Red: best results. Black bold: best previous methods. Blue: GLAM higher than previous methods.

Paper: https://arxiv.org/abs/2107.08000 Contact: hyejoo@oddconcepts.kr