TDAM: Top-Down Attention Module for Contextually Guided Feature Selection in CNNs

Shantanu Jaiswal, Basura Fernando, Cheston Tan

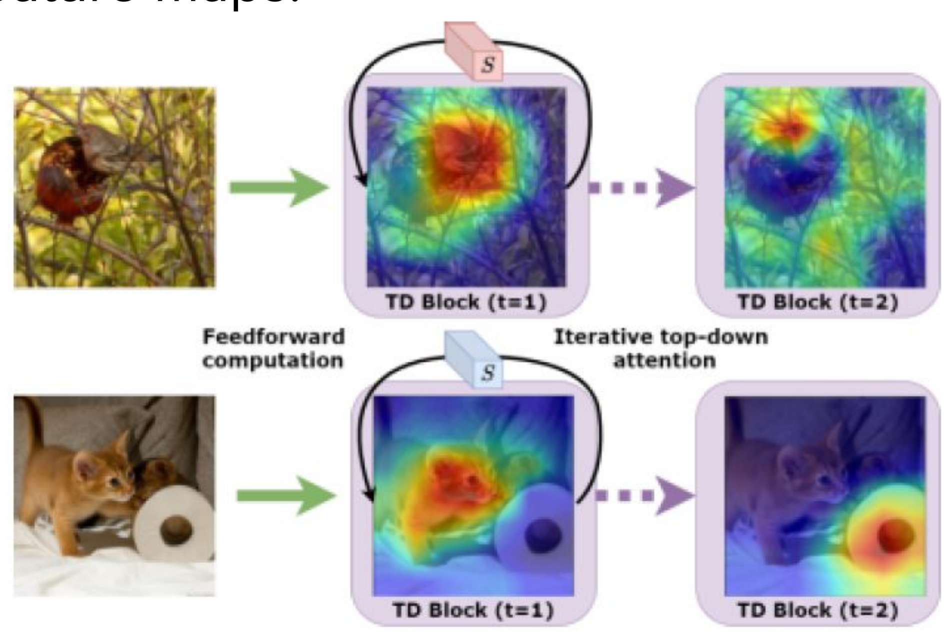
Agency for Science, Technology and Research SINGAPORE

ECCV EUROPEAN CONFERENCE ON COMPUTER VISION TEL AVIV 2022

Motivation

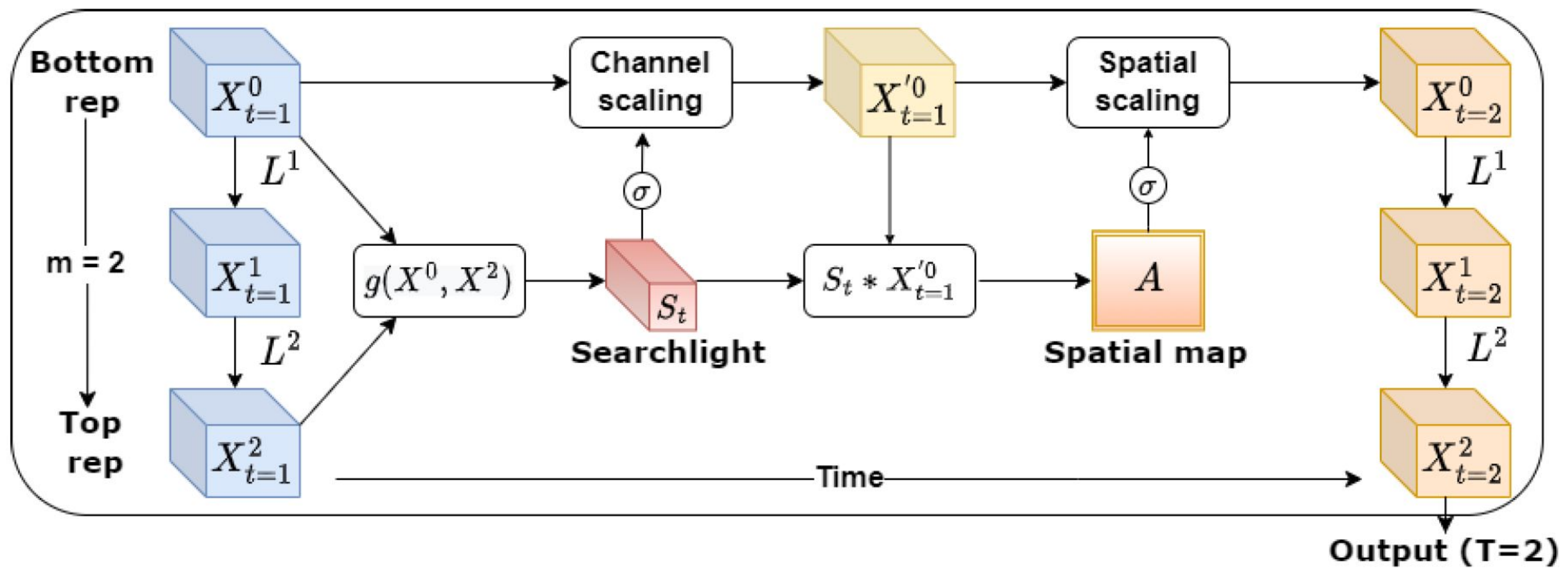
Prominent attention mechanisms for CNNs primarily operate in a feedforward bottom-up manner, thereby being constrained to local information of input feature maps.

Top-down information flow can enable higher-layers to provide semantically-richer contextual information and specify "what and where to look" in lower-level feature maps.



Aim: A lightweight top-down attention module that iteratively generates a "visual searchlight" to perform channel and spatial modulation of its inputs and outputs more contextually-relevant feature maps at each computation step.

Module design



$$X_t^N = L_N(L_{N-1}..(L_1(X_t^0)))$$

$$\mathcal{S}_t = \begin{cases} g(X_t^N, X_t^0) = W_s(\mathrm{ReLU}[W_t(X_{t,p}^N); W_b(X_{t,p}^0)] & \text{if joint attention} \\ g(X_t^N) = W_s[\mathrm{ReLU}(W_t\ (X_{t,p}^N)]) & \text{if top attention} \end{cases}$$

$$X_{t+1}^0 = \boxed{\mathrm{att}(X_t^0; \mathcal{S}_t)}$$

$$X_t'^0 = X_t^0 \otimes \sigma(\mathcal{S}_t)$$

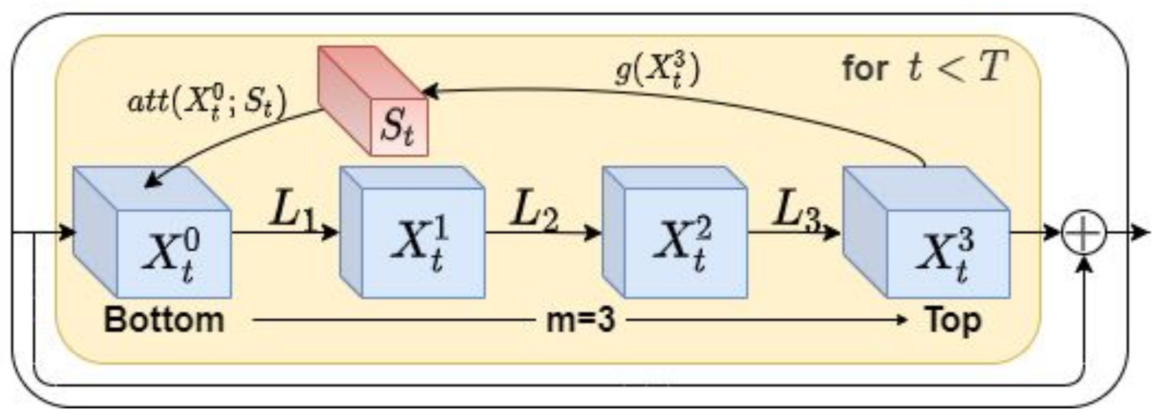
$$\mathcal{A} = \mathcal{S}_t * X_t'^0$$

$$X_{t+1}^0 = X_t'^0 \otimes \sigma(\mathcal{A})$$

Primary module specifications:

1. Number of computation steps ('T')
2. Feedback distance ('m')
3. Attention technique ('joint' or 'top')

Example integration in a ResNet BottleNeck block

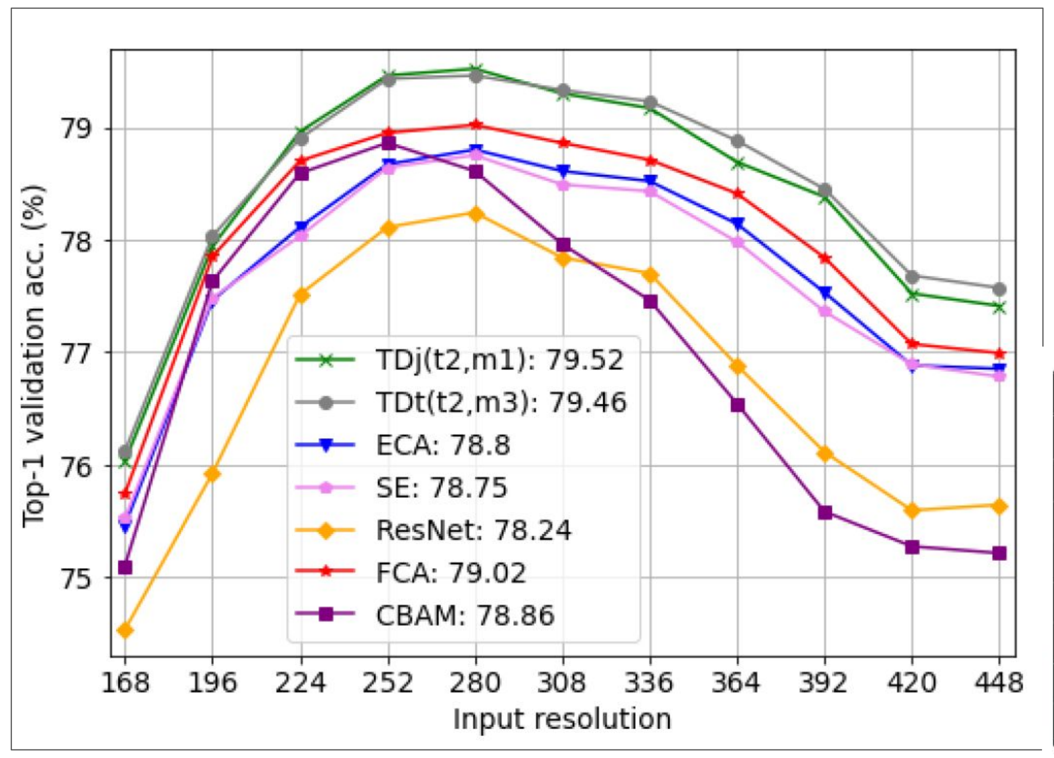


Experiments on ImageNet-1k and resolution robustness

Increases top-1 classification accuracy of a ResNet50 to 79.0% and 67.7% on Ver. 1 and Ver. 2 validation splits respectively.

Outperforms prior state-of-the-art attention modules while having lesser parameters and comparable FPS in most cases (with the exception of ECA).

Method	BB.	Param.	FLOPs	FPS	V2	V1
-	-	-	-	-	Top1	Top1
ResNet (CVPR16)	Resnet50	25.56 M	4.12 G	2143	66.39	77.51
SE (CVPR18)		28.07 M	4.13 G	1911	66.92	78.03
CBAM (ECCV18)		28.07 M	4.14 G	1442	67.28	78.59
ECA (CVPR20)		25.56 M	4.13 G	1989	66.72	78.11
FCA-TS (ICCV21)		28.07 M	4.13 G	1876	67.19	78.70
TDjoint (t=2, m=1)		27.65 M	4.59 G	1890	67.66	**78.96**
TDtop (t=2, m=1)		27.06 M	4.59 G	1905	67.21	78.82
TDtop (t=2, m=3)		27.66 M	5.98 G	1539	**67.70**	78.90
ResNet	Resnet101	44.55 M	7.85 G	1376	69.64	80.36
SE		49.29 M	7.86 G	1201	69.88	80.84
CBAM		49.29 M	7.88 G	862	70.03	81.20
FCA-TS		49.29 M	7.86 G	1164	70.12	81.15
TDjoint (t=2,m=1)		46.75 M	8.37 G	1237	**70.56**	**81.62**
TDjoint(t=2,m=1,L4)		45.94 M	8.01 G	1258	70.28	81.12

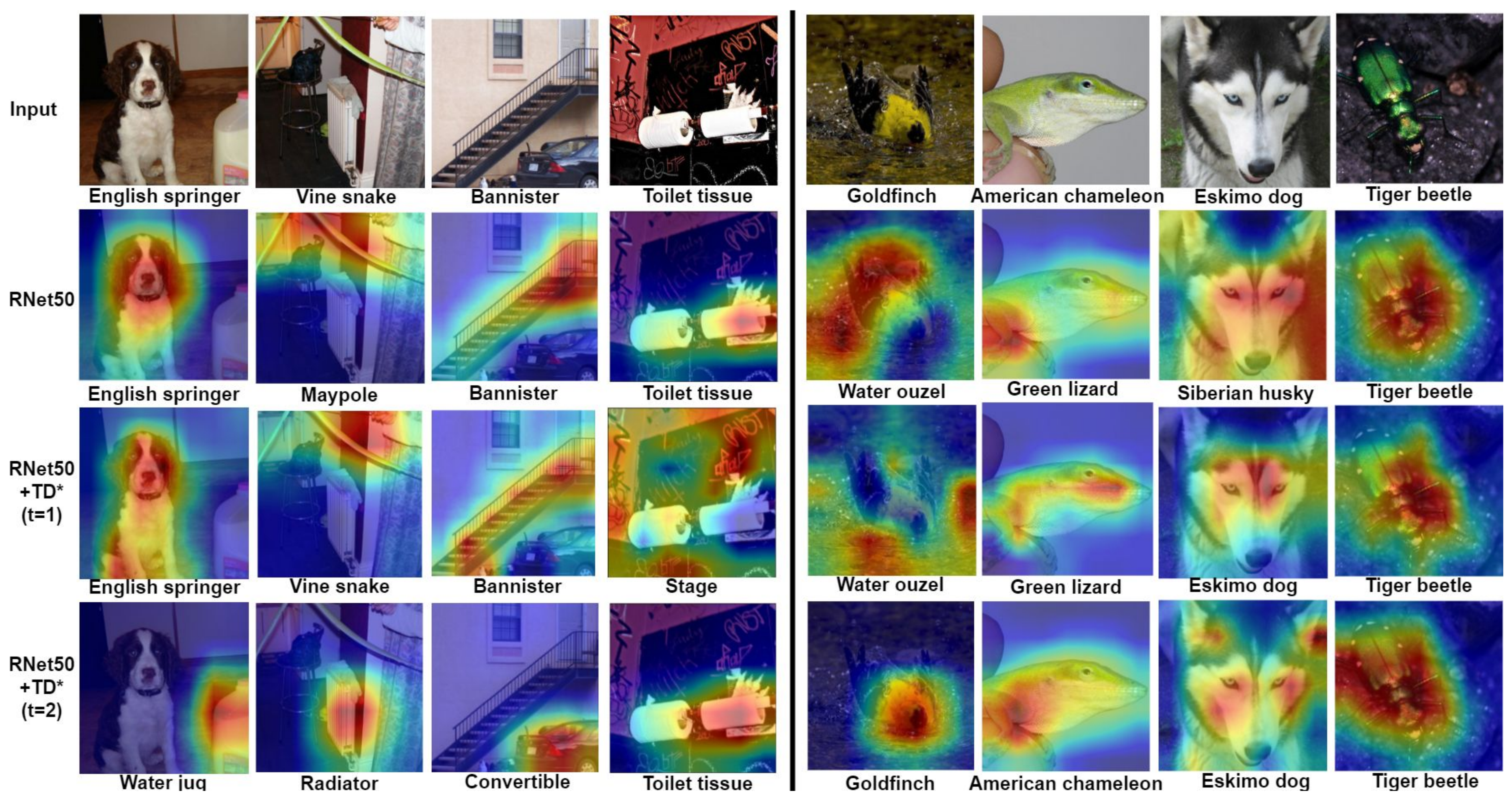


TDAM-models are relatively more robust to changes in input resolution during inference, thereby obtaining better results at higher resolutions.

Model (RNet50)	ImageNet-V1 Top1 Acc.			
-	Best	224^2	168^2	448^2
ResNet	78.24	77.51	74.53	75.64
SE	78.75	78.03	75.52	76.78
CBAM	78.86	78.59	75.10	75.21
ECA	78.80	78.11	75.46	76.85
FCA-TS	79.02	78.70	75.74	76.99
TDjoint(t2,m1)	**79.52**	**78.96**	76.03	77.41
TDtop(t2,m3)	79.46	78.90	**76.12**	**77.57**

Visualizing attention of TDAM-models over time (computation steps)

- TDAM-models learn to "shift attention" by localizing individual objects or features at each computation step without any explicit supervision.
- Visualization script provided in source code to try on arbitrary examples.



(Localizing individual objects) (Localizing relevant features)

Weakly-supervised object localization and other recognition tasks

Model	ImageNet(V1)	
-	Top1	Top5
RNet50	57.04	68.67
RNet101	58.54	69.86
RNet50 + SE	56.62	67.88
RNet50 + CBAM	58.91	70.54
RNet50 + ECA	56.94	68.38
RNet50 + FCA-TS	56.88	67.86
RNet50 + TDjoint(t=2,m=1)	61.55	72.10
RNet50 + TDtop(t=2,m=3)	**61.97**	**72.37**

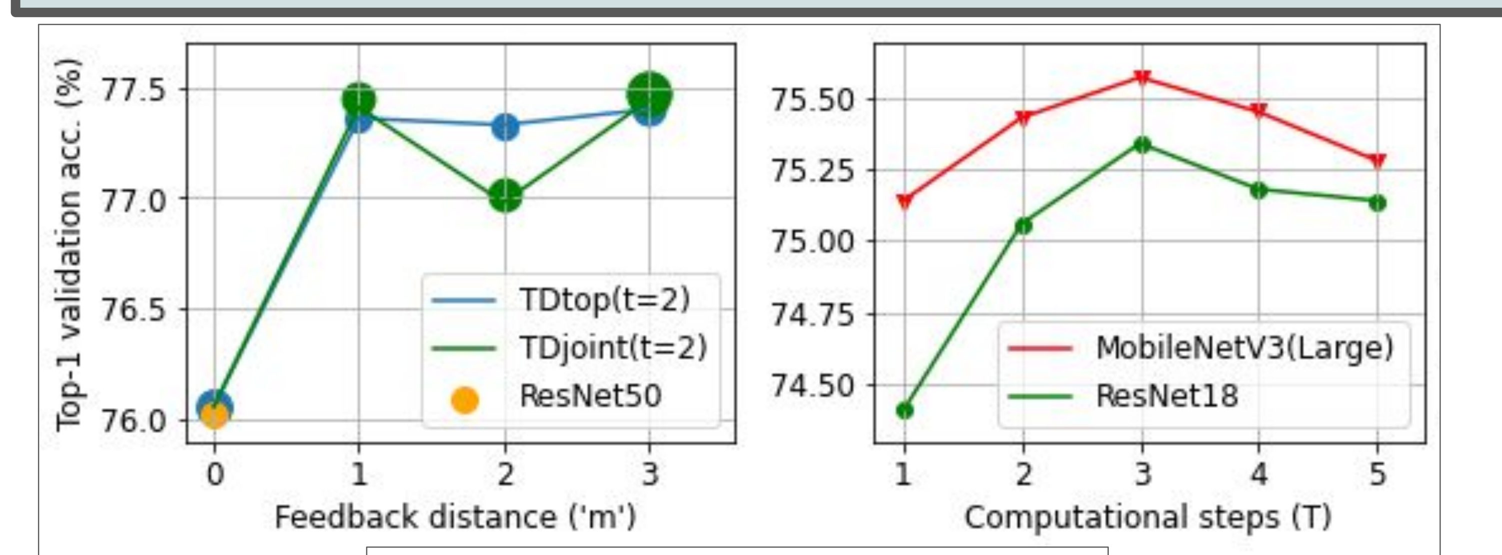
Improves weakly supervised localization acc. by 5% indicating more precise feature activation maps.

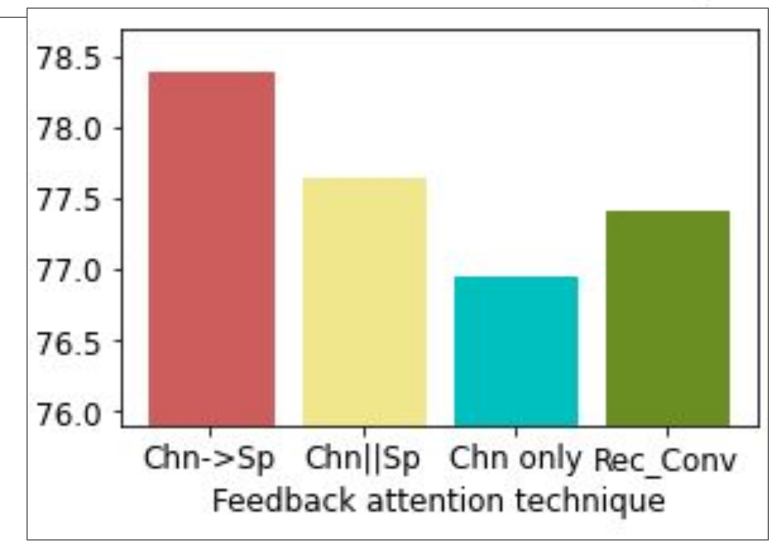
Improves state-of-the-arts for fine-grained (CUB, Stanford Dogs) and multi-label (MS-COCO) object recognition.

Model (ResNet50)	CUB	Dogs	MS-COCO	
-	Top1	Top1	mAP	F1-O
ResNet	88.26	85.97	77.58	75.45
SE	88.89	86.55	78.21	76.37
CBAM	89.37	86.98	79.17	77.15
FCA-TS	88.94	86.76	79.05	77.08
TDjoint(t=2,m=1)	89.61	87.08	**79.61**	**77.71**
TDtop(t=2,m=3)	**89.75**	**87.30**	79.56	77.62

- Improvements on multiple tasks and benchmarks suggest benefits of feedback-driven channel and spatial attention in enabling iterative task-specific refinement of constituent feature maps within the backbone.

Ablation study (Impact of feedback steps, distance and technique)





(Ablations performed on a hierarchically-reduced subset of ImageNet-1k comprising 200 classes; Further analyses in paper)

Source code and models publicly available at: https://github.com/shantanuj/TDAM_Top_down_attention_module