Beyond Softmax: Dual-Branch Sigmoid Architecture for Accurate Class Activation Maps

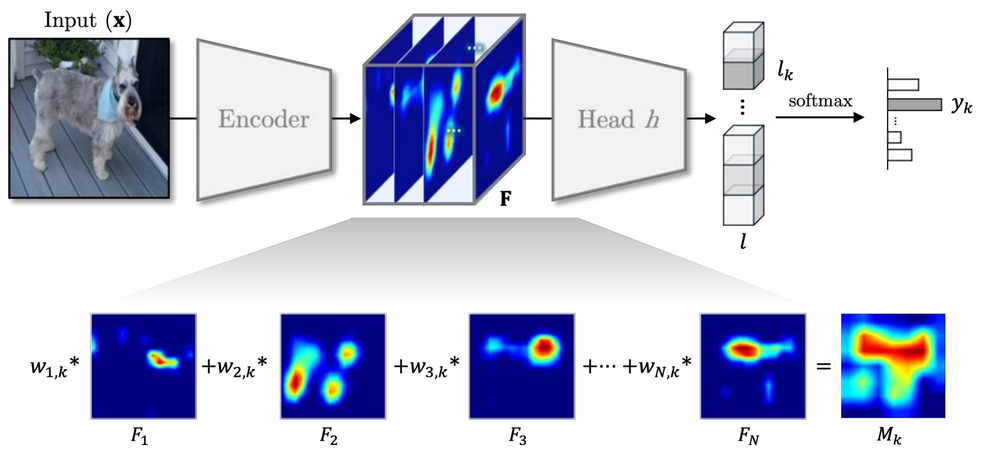
BMVC 2025

Yoojin Oh, Junhyug Noh
Ewha Womans University

Project Page

Motivation

CAM variants yields a heatmap that highlights discriminative regions for a given class by linearly combining feature maps with per-channel weights.
e.g., CAM, Grad-CAM, Grad-CAM++, etc.



Why Softmax fails for CAMs?

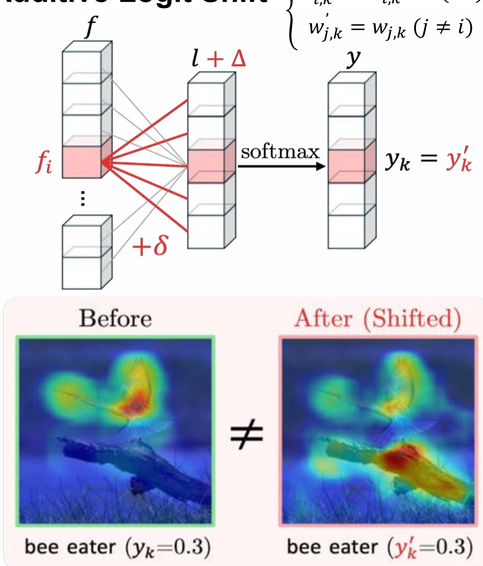
For semantically valid linear combination, the weight $w_{i,k}$ should satisfy:

1. Relative magnitude: how important channel i is for class k than others.
2. Sign semantics: $w_{i,k} > 0$ for evidence; $w_{i,k} < 0$ for inhibition.

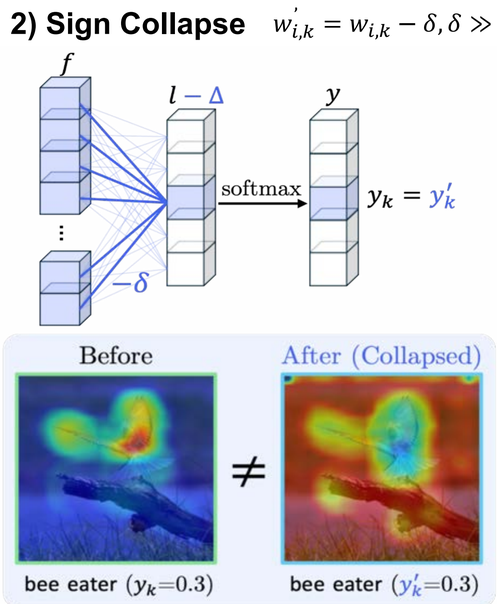
However, softmax's invariances violate these assumptions.

Softmax-Induced Distortions

1) Additive Logit Shift $\begin{cases} w'_{i,k} = w_{i,k} + \delta(\forall k) \\ w'_{j,k} = w_{j,k}\ (j \neq i) \end{cases}$



2) Sign Collapse $w'_{i,k} = w_{i,k} - \delta, \delta \gg 0$



- **Uniform shift** of f_i weights arbitrarily brightens/dims the feature map.
- **Flipping sign** of all weights inverts CAM contributions.

→ Identical classification outputs ($y_k' = \frac{e^{\ell_k + \Delta}}{\sum_{c=1}^{C} e^{\ell_c + \Delta}} = y_k$) produce drastically different localization maps.

Can we decouple localization from classification?

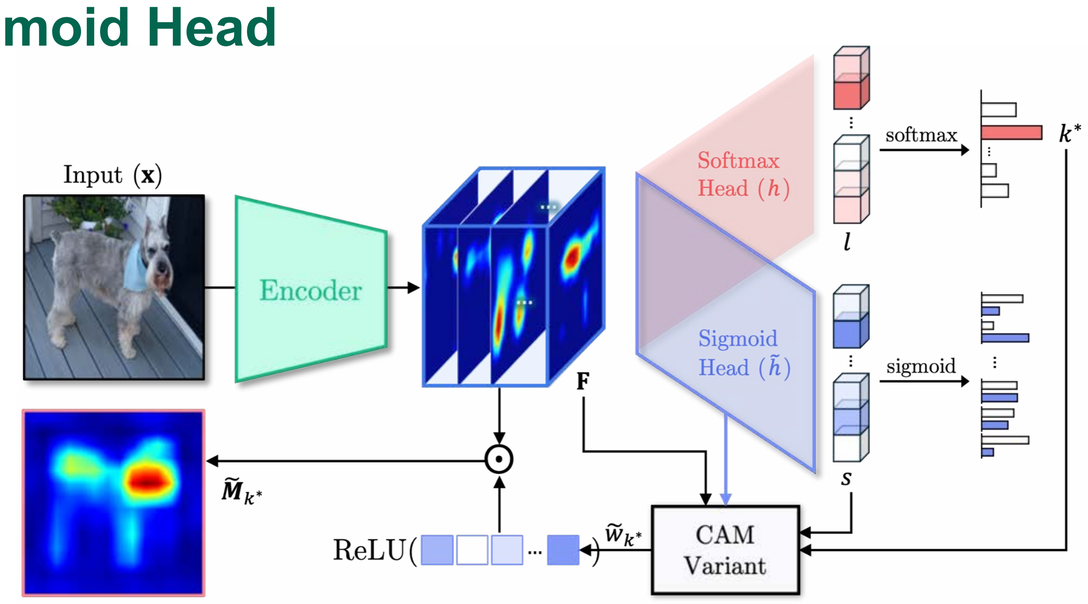
Solution: Dual-Branch Sigmoid Head

Training

- Replicate a pretrained classifier into new sigmoid branch with identical architecture and fresh parameters.
- Class-balanced BCE to correct strong imbalance.

$$\mathcal{L} = -\frac{1}{B}\sum_{n=1}^{B}\Big[\underbrace{\Big(1-\frac{1}{C}\Big)\log s_{y^{(n)}}}_{\text{positive}} + \underbrace{\frac{1}{C}\sum_{k\neq y^{(n)}}\log(1-s_k)}_{\text{negatives}}\Big]$$

- The original softmax head and backbone remain frozen.



Inference

- **Classification**: Predict class k^* via the original softmax branch.
- **Localization**: Compute per-channel weights for class k^* from the sigmoid logit s_{k^*}, not softmax logit ℓ_{k^*}.
- Clamp negative weights to zero by ReLU and combine with feature maps to obtain signed, distortion-free heatmaps.

All without any drop in classification accuracy!

Why it works? "Sigmoid Branch Semantics"

- **Relative importance** – BCE objective ensures within-class comparability; Larger $|\tilde{w}_{i,k}|$ denotes stronger influence on s_k for class k
- **Sign semantics (evidence-only mapping)** – Since σ(·) strictly increases, weight sign encodes evidence (+) vs. inhibition (-)

Quantitative Results

1. Results on Weakly-Supervised Object Localization (WSOL)

Backbone	Method	ImageNet-1K Top1 Cls	ImageNet-1K Top1 Loc	ImageNet-1K GT Loc	ImageNet-1K MBAv2	OpenImages-30K Top1 Cls	OpenImages-30K PxAP
VGG-16	CAM[1]	66.56	43.13	59.54	60.00	70.00	58.17
	+ Ours	–	43.14	59.57	60.25	–	60.11
	Δ	–	(+0.01)	(+0.03)	(+0.25)	–	(+1.94)
	Grad-CAM[2]	69.61	37.04	49.02	52.49	70.12	55.25
	+ Ours	–	44.68	59.43	60.17	–	56.53
	Δ	–	(+7.64)	(+10.41)	(+7.68)	–	(+1.28)
ResNet-50	CAM	73.89	47.69	60.80	63.53	74.02	59.36
	+ Ours	–	49.44	62.68	64.73	–	60.49
	Δ	–	(+1.75)	(+1.88)	(+1.20)	–	(+1.13)
	Grad-CAM	69.01	41.14	55.28	57.04	74.02	60.56
	+ Ours	–	40.34	53.61	57.98	–	60.24
	Δ	–	(-0.80)	(-1.67)	(+0.94)	–	(-0.32)
InceptionV3	CAM	69.72	42.16	57.69	63.51	56.22	62.25
	+ Ours	–	44.11	60.12	64.98	–	63.65
	Δ	–	(+1.95)	(+2.43)	(+1.47)	–	(+1.40)
	Grad-CAM	67.79	34.57	46.35	52.30	68.56	49.64
	+ Ours	–	33.72	45.18	53.06	–	48.16
	Δ	–	(-0.85)	(-1.17)	(+0.76)	–	(-1.48)

- Consistently enhanced Top-1 Localization gains, with only minor trade-offs in select Grad-CAM cases.

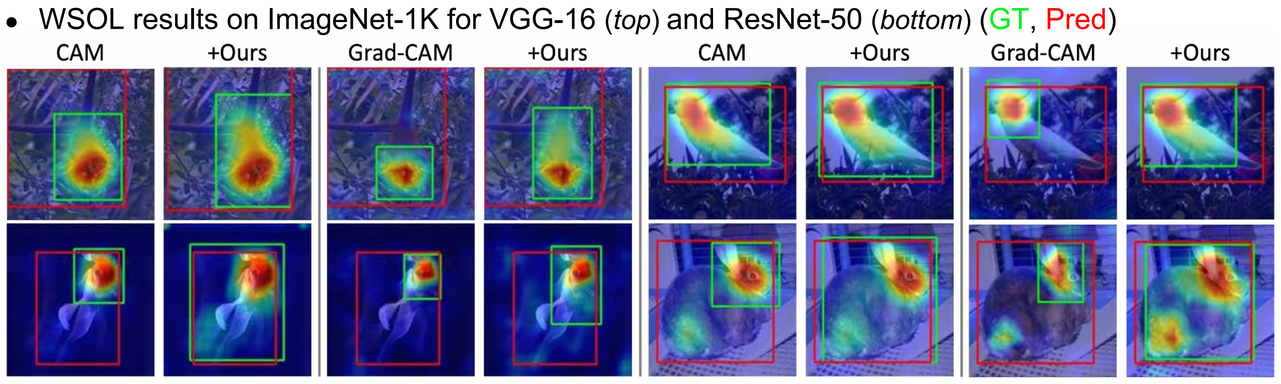
2. Results on Fine-grained Explanation Fidelity

Backbone	Method	CUB-200-2011 % Avg Drop (↓)	CUB-200-2011 % Inc Conf (↑)	Stanford Cars % Avg Drop (↓)	Stanford Cars % Inc Conf (↑)
VGG-16	CAM	45.36	7.08	31.45	9.68
	+ Ours	37.20 (-8.16)	18.86 (+11.78)	43.58 (+12.13)	10.74 (+1.06)
	Grad-CAM	38.88	11.36	56.10	8.11
	+ Ours	35.66 (-3.22)	22.06 (+10.70)	9.96 (-46.14)	19.46 (+11.35)
	Grad-CAM++[3]	31.11	14.46	39.63	11.71
	+ Ours	33.98 (+2.87)	21.75 (+7.29)	9.46 (-30.17)	18.46 (+6.75)
	XGrad-CAM[4]	38.56	11.10	50.59	9.14
	+ Ours	38.04 (-0.52)	20.99 (+9.89)	9.69 (-40.90)	20.53 (+11.39)
	Layer-CAM[5]	37.19	11.27	47.51	9.70
	+ Ours	41.58 (+4.39)	19.50 (+8.23)	10.51 (-37.00)	19.39 (+9.69)
ResNet-50	CAM	44.17	14.76	48.86	4.32
	+ Ours	13.86 (-30.31)	18.42 (+3.66)	3.77 (-45.09)	40.62 (+36.30)
	Grad-CAM	46.38	13.07	29.53	14.56
	+ Ours	19.07 (-27.31)	13.22 (+0.15)	3.66 (-25.87)	41.41 (+26.85)
	Grad-CAM++	43.15	14.50	24.68	17.31
	+ Ours	18.88 (-24.27)	13.72 (-0.78)	3.65 (-21.03)	41.62 (+24.31)
	XGrad-CAM	46.38	13.07	29.53	14.56
	+ Ours	19.07 (-27.31)	13.22 (+0.15)	3.66 (-25.87)	41.41 (+26.85)
	Layer-CAM	43.69	13.91	22.64	18.90
	+ Ours	19.07 (-24.62)	13.22 (-0.69)	3.66 (-18.98)	41.39 (+22.49)

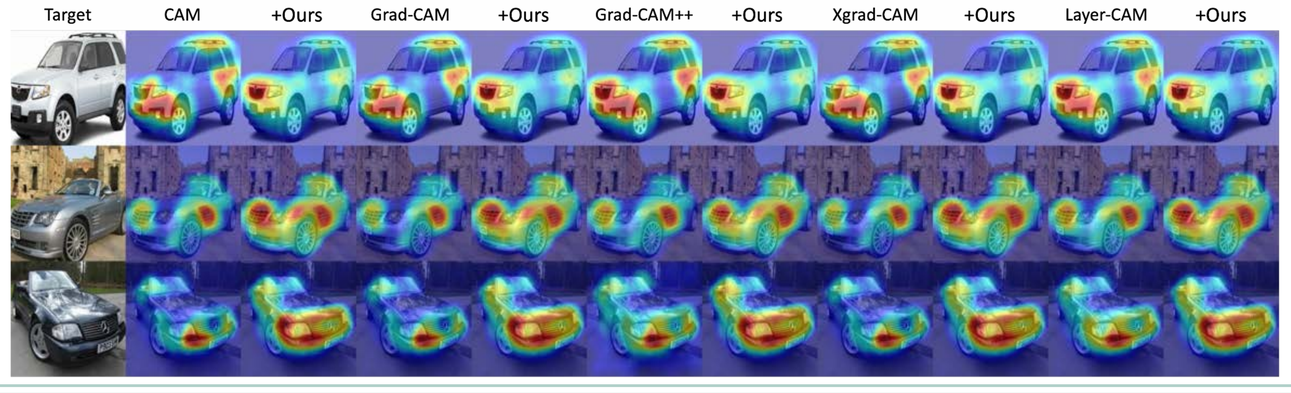
- Sharply reduced drop and elevated confidence, improving explanation fidelity.

Qualitative Results

- WSOL results on ImageNet-1K for VGG-16 (*top*) and ResNet-50 (*bottom*) (GT, Pred)



- Visualization of our model on fine-grained dataset



Ablation Study

- Computational overhead on ImageNet-1K

Backbone	Training Time (h) Baseline	Training Time (h) Ours	Training Time (h) Δ (%)	Inference Time (ms) Baseline	Inference Time (ms) Ours	Inference Time (ms) Δ (%)	# Params (M) Baseline	# Params (M) Ours	# Params (M) Δ (%)
VGG-16	17.67	20.87	+3.20 (18.11)	84.7	101.2	+16.5 (19.48)	20.46	21.49	+1.03 (5.03)
ResNet-50	58.19	70.69	+12.50 (21.48)	117.4	138.4	+21.0 (17.89)	25.56	27.61	+2.05 (8.02)
InceptionV3	50.42	52.84	+2.42 (4.80)	108.1	123.4	+15.3 (14.15)	26.51	27.54	+1.03 (3.89)

**Inference time is measured per image and includes the cost of CAM generation.*

References

[1] Zhou et al., *Learning Deep Features for Discriminative Localization*, CVPR 2016
[2] Selvaraju et al., *Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization*, ICCV 2017
[3] Chattopadhyay et al., *Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks,* WACV 2018
[4] Fu et al., *Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs*, BMVC 2020
[5] Jiang et al., *LayerCAM: Exploring Hierarchical Class Activation Maps for Localization,* TIP 2021