Context-ware Attentional Pooling (CAP) for Fine-grained Visual Classification

The AAAI Conference on Artificial Intelligence

Ardhendu Behera, Zachary Wharton, Pradeep Hewage and Asish Bera
Department of Computer Science, Edge Hill University, Lancashire, UK
{beheraa, whartonz, hewagep, beraa}@edgehill.ac.uk

Edge Hill University

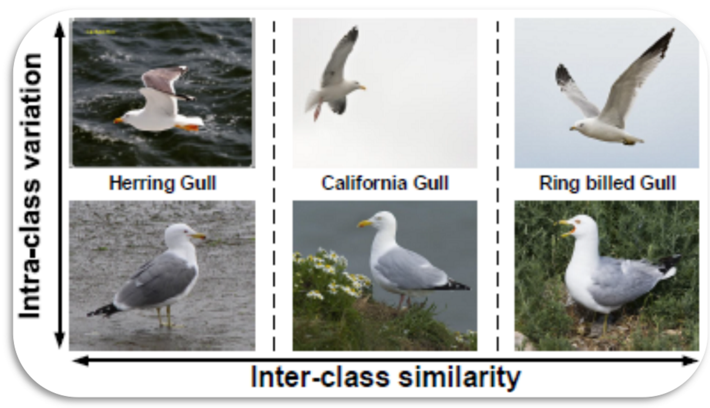
Problem Definition and Motivation

Goal: Distinguishing subordinate categories in fine-grained visual classification (FGVC).

Generic Visual Classification

Fine-Grained Visual Classification (FGVC)

Motivation: Deep CNNs for Generic Visual Recognition learn discriminative features based on changes in global shape and appearance.

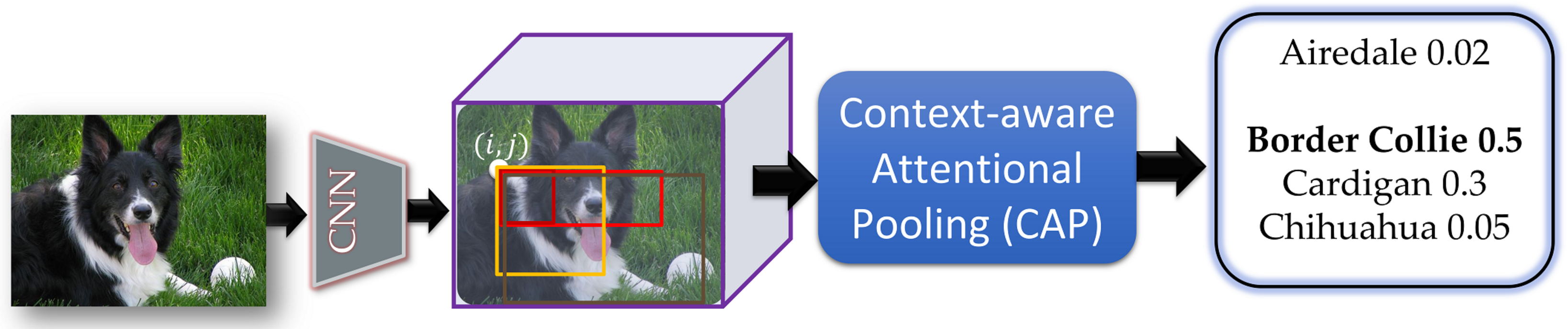


Hsu et al. AAAI 2020

This is inappropriate for distinguishing subordinate categories due to:

- Large **inter-class** similarities
- Large **intra-class** variations

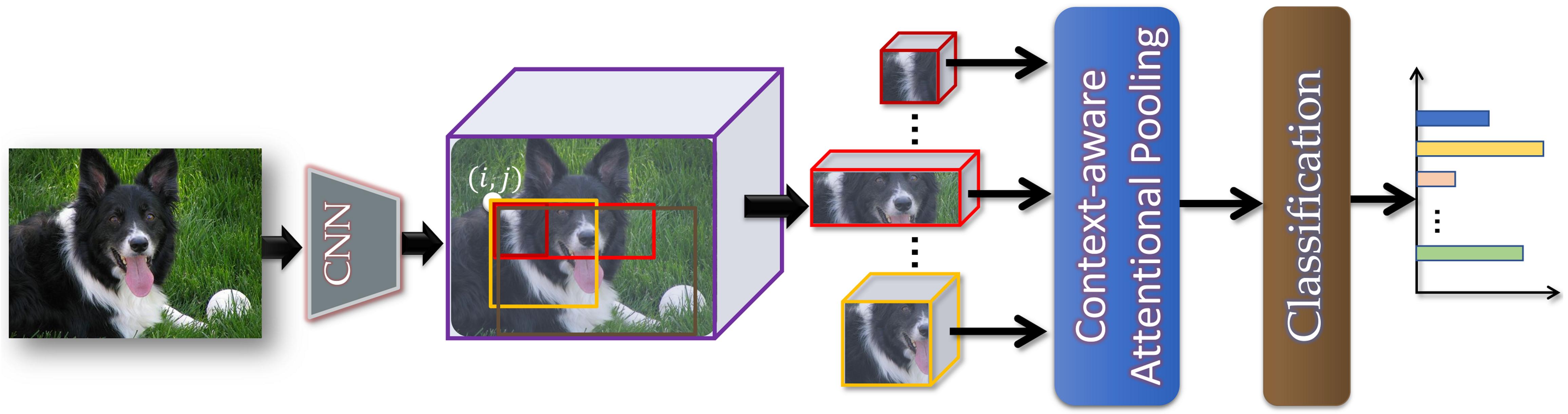
Main Idea: Context-aware Attentional Pooling (CAP) to consider the intrinsic consistency between the informativeness of integral regions and their spatial structures to capture the semantic correlation among them.



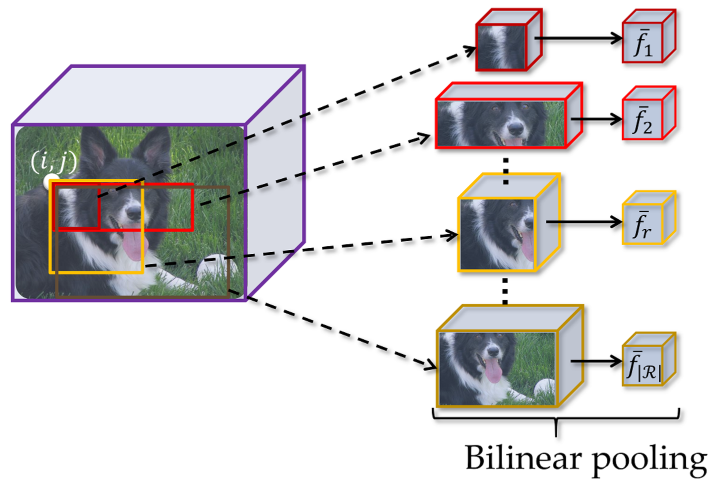
Key Contributions

1. An easy-to-use extension to existing CNNs by incorporating CAP to achieve a considerable improvement in FGVC.
2. Context-aware attention guided rich representation to discriminate the subtle changes in an object/scene.
3. A learnable pooling to automatically select the hidden states of a recurrent network to encode spatial arrangement and appearance features.
4. Extensive evaluation of eight FGVC datasets, obtaining state-of-the-art results.

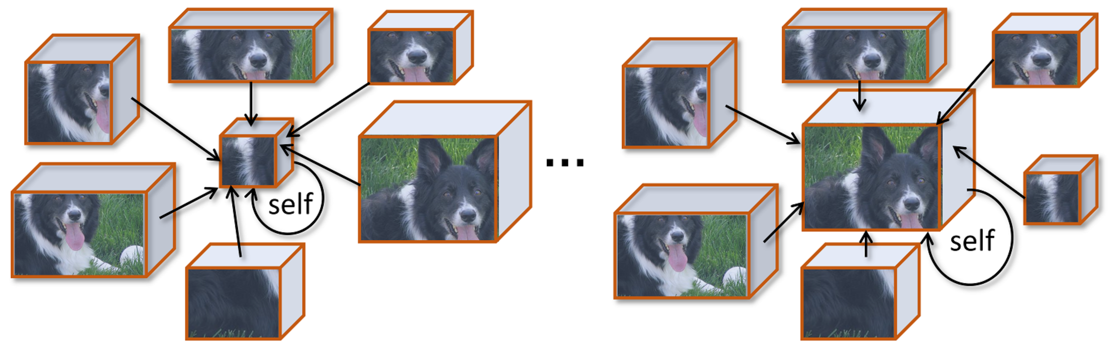
Proposed Context-aware Attentional Pooling



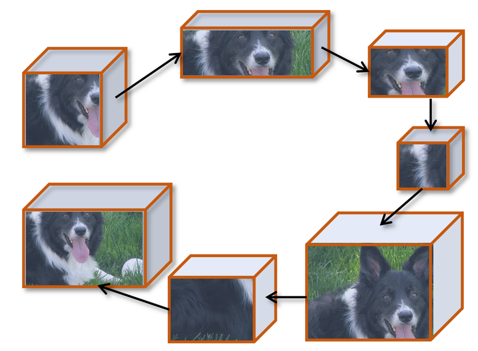
Model Architecture



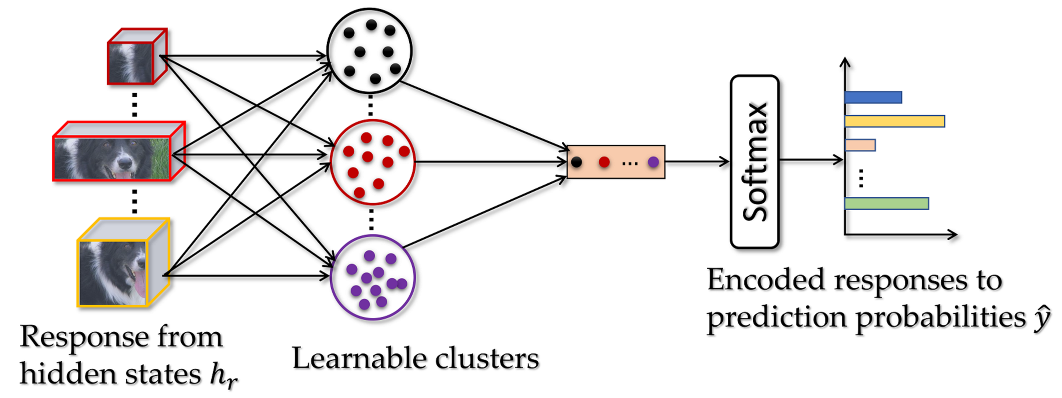
Bilinear pooling from integral regions



Attention-focused contextual information from integral regions $\mathcal{R}$ i.e. $\mathbf{c}_r = \sum_{r,r'=1}^{|\mathcal{R}|} \alpha_{r,r'} \bar{f}_{r'}$



Learning structures
$h_r = \mathcal{F}_h(h_{r-1}, f_r; \theta_h)$
where $f_r = GAP(\mathbf{c}_r)$

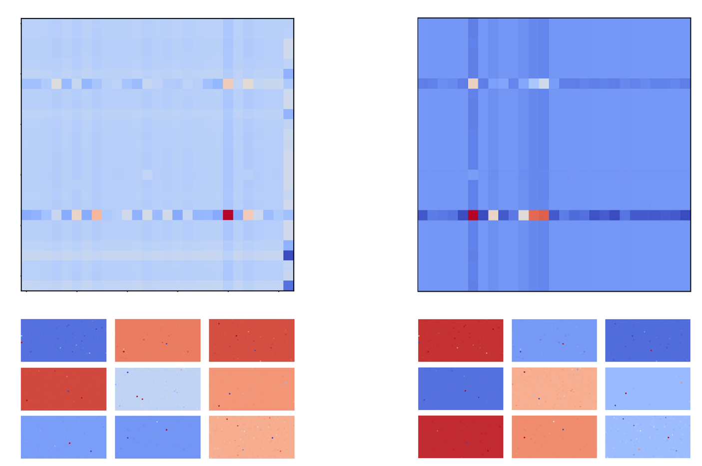


Classification by adapting NetVLAD i.e.
$N_v(o,k) = \sum_{r=1}^{|\mathcal{R}|} \gamma_k(h_r) h_r(o)$ and
$\hat{y} = Softmax(W_N N_v)$

Experimental Evaluation using Eight Benchmarked Datasets.

Dataset	#Train / #Test	#Classes	Our	Past Best (primary)	Past Best (primary + secondary)
Aircraft	6,667 / 3,333	100	**94.9**	93.0 (Chen et al. CVPR 2019)	92.9 (Yu et al. CVPR 2018)
Food-101	75,750 / 25,250	101	**98.6**	93.0 (Huang et al. NIPS 2019)	90.4 (Cui et al. CVPR 2018)
Stanford Cars	8,144 / 8,041	196	**95.7**	94.6 (Huang et al. NIPS 2019)	94.8 (Cubuk et al. CVPR 2019)
Stanford Dogs	12,000 / 8,580	120	96.1	93.9 (Ge et al. CVPR 2019)	**97.1** (Ge et al. CVPR 2019)
CUB-200	5,994 / 5,794	200	**91.8**	90.3 (Ge et al. CVPR 2019)	90.4 (Ge et al. CVPR 2019)
Oxford Flower	2,040 / 6,149	102	97.7	96.4 (Xie et al. CVPR 2016)	**97.7** (Chang et al. TIP 2020)
Oxford Pets	3,680 / 3,669	37	**97.3**	95.9 (Huang et al. NIPS 2019)	93.8 (Peng et al. TIP 2018)
NABirds	23,929 / 24,633	555	**91.0**	86.4 (Luo et al. ICCV 2019)	87.9 (Cui et al. CVPR 2018)

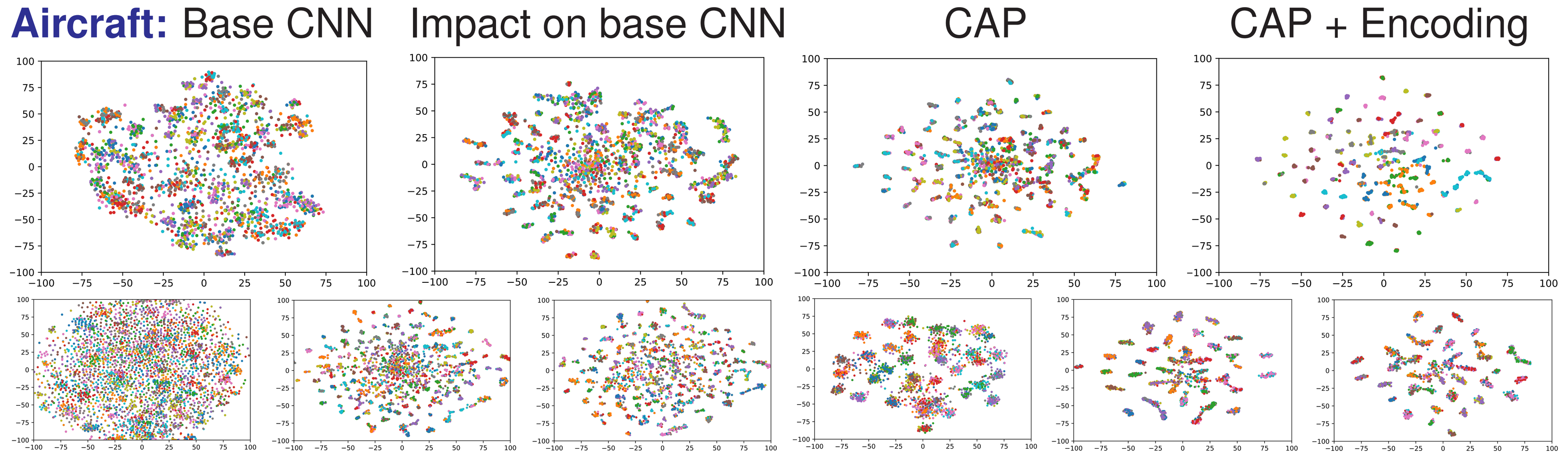
Table 1: Dataset statistics and performance evaluation. FGVC accuracy (%) of our model and the previous best using only the primary dataset. The last column involves the transfer/joint learning strategy consisting of more than one dataset.



CAP's attention-aware response $\alpha_{r,r'}$ for class 1 and class 2 (top row). Class-specific $\mathbf{c}_r$ for 9 classes (3×3) from region 1 and 20 (row 2). Blue to red represents *less* to *more* attention.

Visualizing Discriminability using t-SNE

Aircraft: Base CNN　Impact on base CNN　CAP　CAP + Encoding



Cars: Base CNN, CAP & CAP+Encoding　**Pets:** Base CNN, CAP & CAP+Encoding

Qualitative analysis to monitor class separability and compactness. Visualization of **Aircraft, Stanford Cars** and **Oxford-IIIT Pets** test images.

Misclassification Examples

Left to right: Boeing 747-200 vs Boeing 747-100, Audi TTS Coupe 2012 vs Audi TT RS Coupe 2012, Birman vs Ragdoll, and Staffordshire Bull Terrier vs American Pitbull.

Acknowledgments

This research is supported by the UKIERI-DST grant CHARM (DST UKIERI-2018-19-10) and Edge Hill University Research Investment Fund (RIF). The GPU used in this research is generously donated by the NVIDIA Corporation.