

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

Ardhendu Behera, Zachary Wharton, Pradeep Hewage, Asish Bera

Department of Computer Science, Edge Hill University

St Helen Road, Lancashire

United Kingdom, L39 4QP

beheraa@edgehill.ac.uk, zachary.wharton@google.edgehill.ac.uk, pradeep.hewage@edgehill.ac.uk, beraa@edgehill.ac.uk

Abstract

Deep convolutional neural networks (CNNs) have shown a strong ability in mining discriminative object pose and parts information for image recognition. For fine-grained recognition, context-aware rich feature representation of object/scene plays a key role since it exhibits a significant variance in the same subcategory and subtle variance among different sub-categories. Finding the subtle variance that fully characterizes the object/scene is not straightforward. To address this, we propose a novel context-aware attentional pooling (CAP) that effectively captures subtle changes via sub-pixel gradients, and learns to attend informative integral regions and their importance in discriminating different subcategories without requiring the bounding-box and/or distinguishable part annotations. We also introduce a novel feature encoding by considering the intrinsic consistency between the informativeness of the integral regions and their spatial structures to capture the semantic correlation among them. Our approach is simple yet extremely effective and can be easily applied on top of a standard classification backbone network. We evaluate our approach using six state-of-the-art (SotA) backbone networks and eight benchmark datasets. Our method significantly outperforms the SotA approaches on six datasets and is very competitive with the remaining two.

Introduction

Over recent years, there has been significant progress in the landscape of computer vision due to the adaptation and enhancement of a fast, scalable and end-to-end learning framework, the CNN (LeCun et al. 1998). This is not a recent invention, but we now see a profusion of CNN-based models achieving SotA results in visual recognition (He et al. 2016; Huang et al. 2017; Zoph et al. 2018; Sandler et al. 2018). The performance gain primarily comes from the model’s ability to reason about image content by disentangling discriminative object pose and part information from texture and shape. Most discriminative features are often based on changes in global shape and appearance. They are often ill-suited to distinguish subordinate categories, involving subtle visual differences within various natural objects such as bird species (Wah et al. 2011; Van Horn et al. 2015), flower categories (Nilsback and Zisserman 2008), dog breeds (Khosla et al.

2011), pets (Parkhi et al. 2012) and man-made objects like aircraft types (Maji et al. 2013), car models (Krause et al. 2013), etc. To address this, a global descriptor is essential which ensembles features from multiple local parts and their hierarchy so that the subtle changes can be discriminated as a misalignment of local parts or pattern. The descriptor should also be able to emphasize the importance of a part.

There have been some excellent works on fine-grained visual recognition (FGVC) using weakly-supervised complementary parts (Ge, Lin, and Yu 2019), part attention (Liu et al. 2016), object-part attention (Peng, He, and Zhao 2018), multi-agent cooperative learning (Yang et al. 2018), recurrent attention (Fu, Zheng, and Mei 2017), and destruction and construction learning (Chen et al. 2019). All these approaches avoid part-level annotations and automatically discriminate local parts in an unsupervised/weakly-supervised manner. Many of them use a pre-trained object/parts detector and lack rich representation of regions/parts to capture the object-parts relationships better. To truly describe an image, we need to consider the image generation process from pixels to object to the scene in a more fine-grained way, not only to regulate the object/parts and their spatial arrangements but also defining their appearances using multiple partial descriptions as well as their importance in discriminating subtle changes. These partial descriptions should be rich and complementary to each other to provide a complete description of the object/image. In this work, we propose a simple yet compelling approach that embraces the above properties systematically to address the challenges associated with the FGVC. Thus, it can benefit to a wide variety of applications such as image captioning (Herdade et al. 2019; Huang et al. 2019a; Li et al. 2019), expert-level image recognition (Valan et al. 2019; Krause et al. 2016), and so on.

Our work: To describe objects in a conventional way as in CNNs as well as maintaining their visual appearance, we design a context-aware attentional pooling (CAP) to encode spatial arrangements and visual appearance of the parts effectively. The module takes the input as a convolutional feature from a base CNN and then *learns to emphasize* the latent representation of multiple integral regions (varying coarseness) to describe hierarchies within objects and parts. Each region has an anchor in the feature map, and thus many regions have the same anchor due to the integral characteristics. These integral regions are then fed into a recurrent net-

work (e.g. LSTM) to capture their spatial arrangements, and is inspired by the visual recognition literature, which suggests that humans do not focus their attention on an entire scene at once. Instead, they focus sequentially by attending different parts to extract relevant information (Zoran et al. 2020). A vital characteristic of our CAP is that it generates a new feature map by focusing on a given region conditioned on all other regions and itself. Moreover, it efficiently captures subtle variations in each region by the sub-pixel gradients via bilinear pooling. The recurrent networks are mainly designed for sequence analysis/recognition. We aim to capture the subtle changes between integral regions and their spatial arrangements. Thus, we introduce a learnable pooling to emphasize the most-informative hidden states of the recurrent network, automatically. It learns to encode the spatial arrangement of the latent representation of integral regions and uses it to infer the fine-grained subcategories.

Our contributions: Our main contributions can be summarized as: 1) an easy-to-use extension to SotA base CNNs by incorporating context-aware attention to achieve a considerable improvement in FGVC; 2) to discriminate the subtle changes in an object/scene, context-aware attention-guided rich representation of integral regions is proposed; 3) a learnable pooling is also introduced to automatically select the hidden states of a recurrent network to encode spatial arrangement and appearance features; 4) extensive analysis of the proposed model on eight FGVC datasets, obtaining SotA results; and 5) analysis of various base networks for the wider applicability of our CAP.

Related Work

Unsupervised/weakly-supervised parts/regions based approaches: Such methods learn a diverse collection of discriminative parts/regions to represent the complete description of an image. In (Chen et al. 2019), the global structure of an image is substantially changed by a random patch-shuffling mechanism to select informative regions. An adversarial loss is used to learn essential patches. In (Ge, Lin, and Yu 2019), Mask R-CNN and conditional random field are used for object detection and segmentation. A bidirectional LSTM is used to encode rich complementary information from selected part proposals for classification. A hierarchical bilinear pooling framework is presented in (Yu et al. 2018a) to learn the inter-layer part feature interaction from intermediate convolution layers. This pooling scheme enables inter-layer feature interaction and discriminative part feature learning in a mutually reinforced manner. In (Cai, Zuo, and Zhang 2017), a higher-order integration of hierarchical convolutional features is described for representing parts semantic at different scales. A polynomial kernel-based predictor is defined for modelling part interaction using higher-order statistics of convolutional activations. A general pooling scheme is demonstrated in (Cui et al. 2017) to represent higher-order and nonlinear feature interactions via compact and explicit feature mapping using kernels. Our approach is complementary to these approaches by exploring integral regions and learns to attend these regions using a bilinear pooling that encodes partial information from multiple integral regions to a

comprehensive feature vector for subordinate classification.

Object and/or part-level attention-based approaches: Recently, there has been significant progress to include attention mechanisms (Zhao, Jia, and Koltun 2020; Leng, Liu, and Chen 2019; Bello et al. 2019; Parmar et al. 2019) to boost image recognition accuracy. It is also explored in FGVC (Zheng et al. 2019; Ji et al. 2018; Sun et al. 2018). In (Zheng et al. 2020), a part proposal network produces several local attention maps, and a part rectification network learns rich part hierarchies. Recurrent attention in (Fu, Zheng, and Mei 2017) learns crucial regions at multiple scales. The attended regions are cropped and scaled up with a higher resolution to compute rich features. Object-part attention model (OPAM) in (Peng, He, and Zhao 2018) incorporates object-level attention for object localization, and part-level attention for the vital parts selection. Both jointly learn multi-view and multi-scale features to improve performance. In (Liu et al. 2019), a bidirectional attention-recognition model (BARM) is proposed to optimize the region proposals via a feedback path from the recognition module to the part localization module. Similarly, in attention pyramid hierarchy (Ding et al. 2020), top-down and bottom-up attentions are integrated to learn both high-level semantic and low-level detailed feature representations. In (Rodríguez et al. 2020), a modular feed-forward attention mechanism consisting of attention modules and attention gates is applied to learn low-level feature activations. Our novel paradigm is a step forward and takes inspiration from these approaches. It is advantageous over the existing methods as it uses a single network and the proposed attention mechanism learns to attend both appearance and shape information from a single-scale image in a hierarchical fashion by exploring integral regions. We further extend it by innovating the classification layer, where the subtle changes in integral regions are learned by focusing on the most informative hidden states of an LSTM.

Proposed Approach

The overall pipeline of our model is shown in Fig. 1a. It takes an input image and provides output as a subordinate class label. To solve this, we are given N images $I = \{I_n | n = 1, \dots, N\}$ and their respective fine-grained labels. The aim is to find a mapping function \mathcal{F} that predicts $\hat{y}_n = \mathcal{F}(I_n)$, which matches the true label y_n . The ultimate goal is to learn \mathcal{F} by minimizing a loss $L(y_n, \hat{y}_n)$ between the true and the predicted label. Our model consists of three elements (Fig. 1a): 1) a base CNN $\mathcal{F}_b(\cdot; \theta_b)$, and our novel 2) CAP $\mathcal{F}_c(\cdot; \theta_c)$ and 3) classification $\mathcal{F}_d(\cdot; \theta_d)$ modules. We aim to learn the model's parameters $\theta = \{\theta_b, \theta_c, \theta_d\}$ via end-to-end training. We use the SotA CNN architecture as a base CNN $\mathcal{F}_b(\cdot; \theta_b)$ and thus, we emphasize on the design and implementation of the rest two modules $\mathcal{F}_c(\cdot; \theta_c)$ and $\mathcal{F}_d(\cdot; \theta_d)$.

Context-aware attentional pooling (CAP)

It takes the output of a base CNN as an input. Let us consider $\mathbf{x} = \mathcal{F}_b(I_n; \theta_b)$ to be the convolutional feature map as the output of the base network F_b for input image I_n . The proposed CAP considers contextual information from

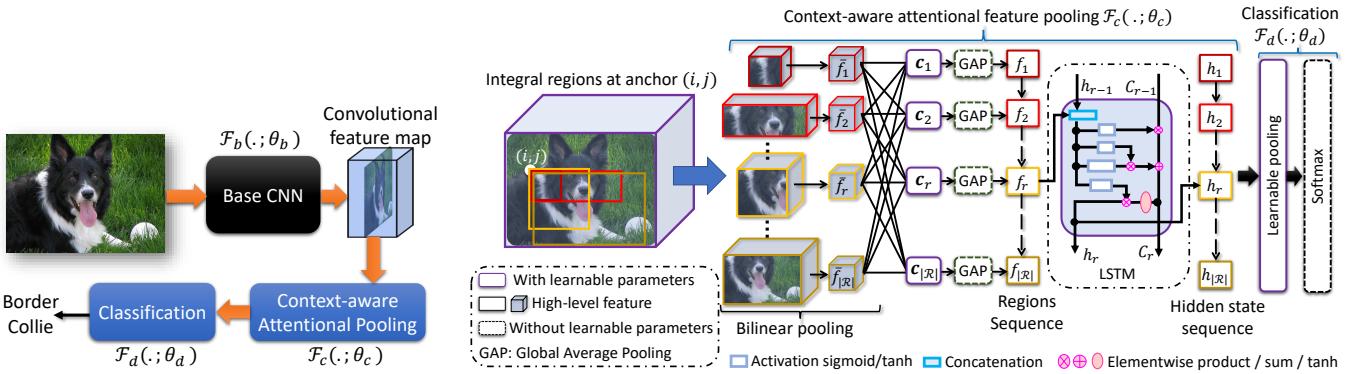


Figure 1: a) High-level illustration of our model (left). b) The detailed architecture of our novel CAP (right).

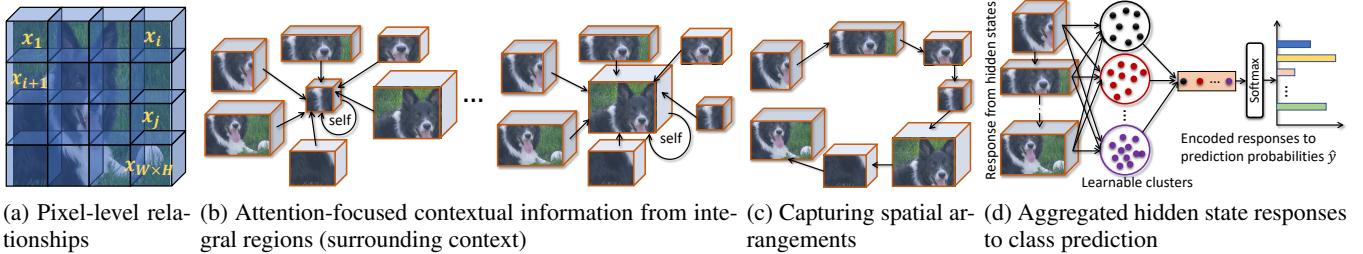


Figure 2: a) Learning pixel-level relationships from the convolutional feature map of size $W \times H \times C$. b) CAP using integral regions to capture both self and neighborhood contextual information. c) Encapsulating spatial structure of the integral regions using an LSTM. d) Classification by learnable aggregation of hidden states of the LSTM.

pixel-level to small patches to large patches to image-level. The pixel refers to a spatial location in the convolutional feature map \mathbf{x} of width W , height H and channels C . The aim is to capture contextual information hierarchically to better model the subtle changes observed in FGVC tasks. Our attention mechanism learns to emphasize pixels, as well as regions of different sizes located in various parts of the image I_n . At pixel-level, we explicitly learn the relationships between pixels, i.e. $p(\mathbf{x}_i|\mathbf{x}_j; \theta_p)$, $\forall i \neq j$ and $1 \leq i, j \leq W \times H$, even they are located far apart in \mathbf{x} . It signifies how much the model should attend the i^{th} location when synthesizing the j^{th} position in \mathbf{x} (Fig. 2a). To achieve this, we compute the attention map θ_p by revisiting the self-attention concept (Zhang et al. 2018) where $key k(\mathbf{x}) = \mathbf{W}_k \mathbf{x}$, $query q(\mathbf{x}) = \mathbf{W}_q \mathbf{x}$ and $value v(\mathbf{x}) = \mathbf{W}_v \mathbf{x}$ in \mathbf{x} are computed using separate 1×1 convolutions. The attentional output feature map \mathbf{o} is a dot-product of attention map θ_p and \mathbf{x} . $\theta_p = \{\mathbf{W}_k, \mathbf{W}_q, \mathbf{W}_v\} \in \theta_c$ is learned.

Proposing integral regions: To learn contextual information efficiently, we propose multiple integral regions with varying level of coarseness on the feature map \mathbf{o} . The level of coarseness is captured by different size of a rectangular region. Let us consider the smallest region $r(i, j, \Delta_x, \Delta_y)$ of width Δ_x , height Δ_y and is located (top-left corner) at the i^{th} column and j^{th} row of \mathbf{o} . Using $r(i, j, \Delta_x, \Delta_y)$, we derive a set of regions by varying their widths and heights i.e. $R = \{r(i, j, m\Delta_x, n\Delta_y)\}; m, n = 1, 2, 3, \dots$ and

$i < i + m\Delta_x \leq W, j < j + n\Delta_y \leq H$. This is illustrated in Fig. 1b (left) for the given spatial location of (i, j) . The goal is to generate the similar set of regions R at various spatial locations ($0 < i < W, 0 < j < H$) in \mathbf{o} . In this way, we generate a final set of regions $\mathcal{R} = \{R\}$ located at different places with different sizes and aspect ratios, as shown in Fig. 1b. The approach is a comprehensive context-aware representation to capture the rich contextual information characterizing subtle changes in images hierarchically. **Bilinear pooling:** There are $|\mathcal{R}|$ regions with size varies from a minimum of $\Delta_x \times \Delta_y \times C$ to a maximum of $W \times H \times C$. The goal is to represent these variable size regions ($X \times Y \times C \Rightarrow (w \times h \times C)$) with a fixed size feature vector. Thus, we use bilinear pooling, typically bilinear interpolation to implement differentiable image transformations, which requires indexing operation. Let $T_\psi(\mathbf{y})$ be the coordinate transformation with parameters ψ and $\mathbf{y} = (i, j) \in \mathbb{R}^2$ denotes a region coordinates at which the feature value is $\mathbf{R}(\mathbf{y}) \in \mathbb{R}^C$. The transformed image $\tilde{\mathbf{R}}$ at the target coordinate $\tilde{\mathbf{y}}$ is:

$$\tilde{\mathbf{R}}(\tilde{\mathbf{y}}) = \sum_{\mathbf{y}} \mathbf{R}(T_\psi(\mathbf{y})) K(\tilde{\mathbf{y}}, T_\psi(\mathbf{y})), \quad (1)$$

where $\mathbf{R}(T_\psi(\mathbf{y}))$ is the image indexing operation and is non-differentiable; thus, the way gradients propagate through the network depends on the kernel $K(., .)$. In bilinear interpolation, the kernel $K(\mathbf{y}_1, \mathbf{y}_2) = 0$ when \mathbf{y}_1 and \mathbf{y}_2 are not

direct neighbors. Therefore, the sub-pixel gradients (i.e. the feature value difference between neighboring locations in the original region) only flow through during propagation (Jiang et al. 2019). This is an inherent flaw in bilinear interpolation since the sub-pixel gradients will not associate to the large-scale changes which cannot be captured by the immediate neighborhood of a point. To overcome this, several variants (Jiang et al. 2019; Lin and Lucey 2017) have been proposed. However, for our work, we exploit this flaw to capture the subtle changes in all regions via sub-pixel gradients. Note that the bilinear interpolation, although is not differentiable at all points due to the *floor* and *ceiling* functions, can backpropagate the error and is differentiable in most inputs as mentioned in the seminal work of Spatial Transform Networks (Jaderberg et al. 2015). We use bilinear kernel $K(.,.)$ in (1) to pool fixed size features \bar{f}_r ($w \times h \times C$) from all $r \in \mathcal{R}$.

Context-aware attention: In this step, we capture the contextual information using our novel attention mechanism, which transforms \bar{f}_r to a weighted version of itself and conditioned on the rest of the feature maps $\bar{f}_{r'}$ ($r, r' \in \mathcal{R}$). It enables our model to selectively focus on more relevant integral regions to generate holistic context information. The proposed context-aware attention takes a *query* $\mathbf{q}(\bar{f}_r)$ and maps against a set of *keys* $\mathbf{k}(\bar{f}_{r'})$ associated with the integral regions r' in a given image, and then returns the output as a context vector \mathbf{c}_r and is computed as:

$$\begin{aligned} \mathbf{c}_r &= \sum_{r'=1}^{|\mathcal{R}|} \alpha_{r,r'} \bar{f}_{r'}, \quad \alpha_{r,r'} = \text{softmax}(W_\alpha \beta_{r,r'} + b_\alpha) \\ \beta_{r,r'} &= \tanh(\mathbf{q}(\bar{f}_r) + \mathbf{k}(\bar{f}_{r'}) + b_\beta) \\ \mathbf{q}(\bar{f}_r) &= W_\beta \bar{f}_r \text{ and } \mathbf{k}(\bar{f}_{r'}) = W_{\beta'} \bar{f}_{r'}, \end{aligned} \quad (2)$$

where weight matrices W_β and $W_{\beta'}$ are for estimating the *query* and *key* from the respective feature maps \bar{f}_r and $\bar{f}_{r'}$; W_α is their nonlinear combination; b_α and b_β are the biases. These matrices and biases ($\{W_\beta, W_{\beta'}, W_\alpha, b_\alpha, b_\beta\} \in \theta_c$) are learnable parameters. The context-aware attention element $\alpha_{r,r'}$ captures the similarity between the feature maps \bar{f}_r and $\bar{f}_{r'}$ of regions r and r' , respectively. The attention-focused context vector \mathbf{c}_r determines the *strength* of \bar{f}_r in focus *conditioned on itself and its neighborhood context*. This applies to all integral regions r (refer Fig. 2b).

Spatial structure encoding: The context vectors $\mathbf{c} = \{\mathbf{c}_r | r = 1 \dots |\mathcal{R}|\}$ characterize the attention and saliency. To include the structural information involving the spatial arrangements of regions (see Fig. 1b and 2b), we represent \mathbf{c} as a sequence of regions (Fig. 2c) and adapt a recurrent network to capture the structural knowledge using its internal states, which is modeled via hidden units $h_r \in \mathbb{R}^n$. Thus, the internal state representing the region r is updated as: $h_r = \mathcal{F}_h(h_{r-1}, f_r; \theta_h)$, where \mathcal{F}_h is a nonlinear function with learnable parameter θ_h . We use a fully-gated LSTM as \mathcal{F}_h (Hochreiter and Schmidhuber 1997) which is capable of learning long-term dependencies. The parameter $\theta_h \in \theta_c$ consists of weight matrices and biases linking input, forget and output gates, and cell states of \mathcal{F}_h . For simplicity, we omitted equations to compute these parameters

and refer interested readers to (Hochreiter and Schmidhuber 1997) for further details. To improve the generalizability and lower the computational complexity of our CAP, the context feature f_r is extracted from the context vector \mathbf{c}_r via global average pooling (GAP). This results in the reduction of feature map size from $(w \times h \times C)$ to $(1 \times C)$. The sequence of hidden states $h = (h_1, h_2, \dots, h_r, \dots, h_{|\mathcal{R}|})$ corresponding to the input sequence of context feature $f = (f_1, f_2, \dots, f_r, \dots, f_{|\mathcal{R}|})$ (see Fig. 1b) is used by our classification module $\mathcal{F}_d(\cdot; \theta_d)$.

Classification

To further guide our model to discriminate the subtle changes, we propose a learnable pooling approach (Fig. 2c), which aggregates information by grouping similar responses from the hidden states h_r . It is inspired by the existing feature encoding approach, such as NetVLAD (Arandjelovic et al. 2016). We adapt this differentiable clustering approach for the soft assignment of the responses from hidden states h_r to cluster κ and their contribution to the VLAD encoding.

$$\begin{aligned} \gamma_\kappa(h_r) &= \frac{e^{W_\kappa^T h_r + b_\kappa}}{\sum_{i=1}^K e^{W_i^T h_r + b_i}} \\ N_v(o, \kappa) &= \sum_{r=1}^{|\mathcal{R}|} \gamma_\kappa(h_r) h_r(o), \hat{y} = \text{softmax}(W_N N_v) \end{aligned} \quad (3)$$

where W_i and b_i are learnable clusters' weights and biases. T signifies transpose. The term $\gamma_\kappa(h_r)$ refers to the soft assignment of h_r to cluster κ , and N_v is the encoded responses of hidden states from all the regions $r \in \mathcal{R}$. In the original implementation of VLAD, the weighted sum of the residuals is used i.e. $\sum_{r=1}^{|\mathcal{R}|} \gamma_\kappa(h_r) (h_r(o) - \hat{c}_\kappa(o))$ in which \hat{c}_κ is the κ^{th} cluster center and $o \in h_r$ is one of the elements in the hidden state response. We adapt the simplified version that averages the actual responses instead of residuals (Miech, Laptev, and Sivic 2017), which requires fewer parameters and computing operations. The encoded response is mapped into prediction probability \hat{y} by using a learnable weight W_N and softmax. The learnable parameter for the classification module \mathcal{F}_d is $\theta_d = \{W_i, b_i, W_N\}$.

Experiments and Discussion

We comprehensively evaluate our model on widely used eight benchmark FGVC datasets: Aircraft (Maji et al. 2013), Food-101 (Bossard, Guillaumin, and Gool 2014), Stanford Cars (Krause et al. 2013), Stanford Dogs (Khosla et al. 2011), Caltech Birds (CUB-200) (Wah et al. 2011), Oxford Flower (Nilsback and Zisserman 2008), Oxford-IIIT Pets (Parkhi et al. 2012), and NABirds (Van Horn et al. 2015). We do not use any bounding box/part annotation. Thus, we do not compare with methods which rely on these. Statistics of datasets and their train/test splits are shown in Table 1. We use the top-1 accuracy (%) for evaluation.

Experimental settings: In all our experiments, we resize images to size 256×256 , apply data augmentation techniques of random rotation (± 15 degrees), random scaling (1 ± 0.15) and then random cropping to select the final size

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. 2019)	92.9 (Yu et al. 2018b)
Food-101	75,750 / 25,250	101	98.6	93.0 (Huang et al. 2019b)	90.4 (Cui et al. 2018)
Stanford Cars	8,144 / 8,041	196	95.7	94.6 (Huang et al. 2019b)	94.8 (Cubuk et al. 2019)
Stanford Dogs	12,000 / 8,580	120	96.1	93.9 (Ge, Lin, and Yu 2019)	97.1 (Ge, Lin, and Yu 2019)
CUB-200	5,994 / 5,794	200	91.8	90.3 (Ge, Lin, and Yu 2019)	90.4 (Ge, Lin, and Yu 2019)
Oxford Flower	2,040 / 6,149	102	97.7	96.4 (Xie et al. 2016)	97.7 (Chang et al. 2020)
Oxford Pets	3,680 / 3,669	37	97.3	95.9 (Huang et al. 2019b)	93.8 (Peng, He, and Zhao 2018)
NABirds	23,929 / 24,633	555	91.0	86.4 (Luo et al. 2019)	87.9 (Cui et al. 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.

Aircraft		Food-101		Stanford Cars	
Method	ACC	Method	ACC	Method	ACC
DFL (Wang et al. 2018)	92.0	WISER (Martinel et al., 2018)	90.3	BARM (Liu et al. 2019)	94.3
BARM (Liu et al. 2019)	92.5	DSTL* (Cui et al. 2018)	90.4	MC [*] _{Loss} (Chang et al. 2020)	94.4
GPipe (Huang et al. 2019b)	92.7	MSMVA (Jiang et al. 2020)	90.6	DCL(Chen et al. 2019)	94.5
MC [*] _{Loss} (Chang et al. 2020)	92.9	JDNet* (Zhao et al. 2020)	91.2	GPipe (Huang et al. 2019b)	94.6
DCL (Chen et al. 2019)	93.0	GPipe (Huang et al. 2019b)	93.0	AutoAug* (Cubuk et al. 2019)	94.8
Proposed	94.9	Proposed	98.6	Proposed	95.7
CUB-200		Oxford-IIIT Pets		NABirds	
iSQRT (Li et al. 2018)	88.7	NAC (Simon and Rodner 2015)	91.6	T-Loss (Taha et al. 2020)	79.6
DSTL* (Cui et al. 2018)	89.3	TL-Attn* (Xiao et al. 2015)	92.5	PC-CNN (Dubey et al. 2018a)	82.8
DAN (Hu et al. 2019)	89.4	InterAct (Xie et al. 2016)	93.5	MaxEnt* (Dubey et al. 2018b)	83.0
BARM (Liu et al. 2019)	89.5	OPAM* (Peng, He, and Zhao 2018)	93.8	Cross-X (Luo et al. 2019)	86.4
CPM* (Ge, Lin, and Yu 2019)	90.4	GPipe (Huang et al. 2019b)	95.9	DSTL* (Cui et al. 2018)	87.9
Proposed	91.8	Proposed	97.3	Proposed	91.0

Table 2: Accuracy (%) comparison with the recent top-five SotA approaches. Methods marked with * involve transfer/joint learning strategy for objects/patches/regions consisting more than one dataset (primary and secondary). Please refer to the supplementary page in the end for the results of Stanford Dogs and Oxford Flowers.

of 224×224 from 256×256 . The last Conv layer of the base CNN (e.g. 7×7 pixels) is increased to 42×42 by using an upsampling layer (as in GAN) and then fed into our CAP (Fig. 1a) to pool features from multiple integral regions \mathcal{R} . We fix bilinear pooling size of $w = h = 7$ for each region with minimum width $\Delta_x = 7$ and height $\Delta_y = 7$. We use spatial location gap of 7 pixels between consecutive anchors to generate $|\mathcal{R}| = 27$ integral regions. This is decided experimentally by considering the trade-off between accuracy and computational complexity. We set the cluster size to 32 in our learnable pooling approach. We apply Stochastic Gradient Descent (SGD) optimizer to optimize the categorical cross-entropy loss function. The SGD is initialized with a momentum of 0.99, and initial learning rate 1e-4, which is multiplied by 0.1 after every 50 epochs. The model is trained for 150 epochs using an NVIDIA Titan V GPU (12 GB). We use Keras+Tensorflow to implement our algorithm.

Quantitative results and comparison to the SotA approaches: Overall, our model outperforms the SotA approaches by a clear margin on all datasets except the Stanford Dogs (Khosla et al. 2011) and Oxford Flowers (Nilsback and Zisserman 2008) (Table 1). In Table 1, we compare our performances with the two previous best (last two columns). One uses only the target dataset (primary) for

training and evaluation (past best) and is the case in our model. The other (last column) uses primary and additional secondary (e.g. ImageNet, COCO, iNat, etc.) datasets for joint/transfer learning of objects/patches/regions during training. It is worth mentioning that we use only the primary datasets and our performance in most datasets is significantly better than those uses additional datasets. This demonstrates the benefit of the proposed approach for discriminating fine-grained changes in recognizing subordinate categories. Moreover, we use only one network for end-to-end training, and our novel CAP and classification layers are added on top of a base CNN. Therefore, the major computations are associated with the base CNNs.

Using our model, the two highest gains are 5.6% and 3.1% in the respective Food-101 (Bossard, Guillaumin, and Gool 2014) and NABirds (Van Horn et al. 2015) datasets. In Dogs, our method (96.1%) is significantly better than the best SotA approach (93.9%) (Ge, Lin, and Yu 2019) using only primary data. However, their accuracy increases to 97.1% when joint fine-tuning with selected ImageNet images are used. Similarly, in Flowers, our accuracy (97.7%) is the same as in (Chang et al. 2020) which uses both primary and secondary datasets, and we achieve an improvement of 1.3% compared to the best SotA approach in (Xie et al. 2016)

Base CNN	Plane	Cars	Dogs	CUB	Flowers	Pets
ResNet-50	94.9	94.9	95.8	90.9	97.5	96.7
Incep. V3	94.8	94.8	95.7	91.4	97.6	96.2
Xception	94.1	95.7	96.1	91.8	97.7	97.0
DenseNet	94.6	93.6	95.5	91.6	97.6	96.9
NASNet-M	93.8	93.7	96.0	89.7	97.7	97.3
Mob-NetV2	94.4	94.0	95.9	89.2	97.4	96.4

Table 3: Our model’s accuracy (%) with different SotA base CNN architectures. Previous best accuracies for these results are; Aircraft: 93.0 (Chen et al. 2019), Cars: 94.6 (Huang et al. 2019b), Dogs: 93.9 (Ge, Lin, and Yu 2019), CUB: 90.3 (Ge, Lin, and Yu 2019), Flowers: 96.4 (Xie et al. 2016), and Pets: 95.9 (Huang et al. 2019b). The result of the Birds dataset is included in the supplementary document in the end.

using only primary data. We also compare our model’s accuracy with the top-five SotA approaches on each dataset in Table 2. Our accuracy is significantly higher than SotA methods using primary data in all six datasets in Table 2 and two in supplementary (provided in the end). Furthermore, it is also considerably higher than SotA methods, which use both primary and secondary data in six datasets (Aircraft, Food-101, Cars, CUB-200, Pets and NABirds). This clearly proves our model’s powerful ability to discriminate subtle changes in recognizing subordinate categories without requiring additional datasets and/or subnetworks and thus, has an advantage of easy implementation and a little computational overhead in solving FGVC.

Ablation study: We compare the performance of our approach using the benchmarked base CNN architectures such as ResNet-50 (He et al. 2016), Inception-V3 (Szegedy et al. 2016), Xception (Chollet 2017) and DenseNet121 (Huang et al. 2017), as well as SotA lightweight architectures such as NASNetMobile (Zoph et al. 2018) and MobileNetV2 (Sandler et al. 2018). The performance is shown in Table 3. In all datasets, both standard and lightweight architectures have performed exceptionally well when our proposed CAP and classification modules are incorporated. Even our model outperforms the previous best (primary data) for both standard and lightweight base CNNs except in Cars and CUB-200 datasets in which our model with standard base CNNs exceed the previous best. Our results in Table 1 & 2 are the best accuracy among these backbones. Nevertheless, the accuracy of our model using any standard backbones (ResNet50 / Inception V3 / Xception; Table 3) is better than the SotA. In Flowers and Pets datasets, the lightweight NASNetMobile is the best performer, and the MobileNetV2 is not far behind (Table 3). This could be linked to the dataset size since these two are of smallest in comparison to the rest (Table 1). However, in other datasets (e.g. Aircraft, Cars and Dogs), there is a little gap in performance between standard and lightweight CNNs. These lightweight CNNs involve significantly less computational costs, and by adding our modules, the performance can be as competitive as the standard CNNs. This proves the importance of our modules in enhancing performance and its broader applicability.

We have also evaluated the above base CNNs (B), and the influence of our novel CAP (+C) and the classification module (+E) in the recognition accuracy on Aircraft, Cars and Pets datasets (more in the supplementary in the end). The results are shown in Table 4. It is evident that the accuracy improves as we add our modules to the base networks, i.e., (B+C+E) > (B+C) > (B+E) > B, resulting in the largest gain contributed by our novel CAP (B+C). This signifies the impact of our CAP. In B+C, the minimum gain is 7.2%, 5.7% and 5.1% on the respective Aircraft, Cars and Pets datasets for the Inception-V3 as a base CNN. Similarly, the highest gain is 12.5% and 11.3% in Aircraft and Cars, respectively. These two datasets are relatively larger than the Pets (Table 1) in which the highest gain (7.9%) is achieved by using ResNet-50 as a base CNN. We also observe that there is a significant gap in baseline accuracy between lightweight and standard base CNNs in larger (Aircraft and Cars) datasets. These gaps are considerably reduced when our CAP is added. There is a further increase in accuracy when we add the classification module (B+C+E). This justifies the inclusion of our novel encoding by grouping hidden responses using residual-less NetVLAD and then infer class probability using learnable pooling from these encoded responses. For base CNNs, we use the standard *transfer learning* by fine-tuning it on the target dataset using the same data augmentation and hyper-parameters. For our models, we use pre-trained weights for faster convergence. We experimentally found that the random initialization takes nearly double iterations to converge (similar accuracy) than the pre-trained weights. A similar observation is shown in (He, Girshick, and Dollár 2019).

Our model’s accuracy is also compared using different numbers of regions $|\mathcal{R}|$. It is a hyper-parameter and is computed from Δ_x and Δ_y . The results are shown in Table 5 (best $|\mathcal{R}| = 27$). We have also provided results for top-N accuracy in the supplementary document provided in the end. The top-2 accuracy is around 99% and is independent of the CNN types.

Model complexity: It is represented as a number of trainable parameters in millions and per-image inference time in millisecond (Table 4). It also depends on the base CNNs types (e.g. standard vs lightweight). Given the number of trainable parameters (9.7M) and inference time (3.5ms), the performance of the lightweight NASNetMobile is very competitive in comparison to the rest. The role of secondary data has improved accuracy in (Chang et al. 2020; Cubuk et al. 2019; Ge, Lin, and Yu 2019; Ge and Yu 2017). However, such models involve multiple steps and resource-intensive, resulting in difficulty in implementing. For example, 3 steps in (Ge, Lin, and Yu 2019): 1) object detection and instance segmentation (Mask R-CNN and CRF), 2) complementary part mining (512 ROIs) and 3) classification using context gating. The model is trained using 4 GPUs. In contrast, our model can be trained on a single GPU (12 GB). The per-image inference time is 4.1ms. In (Ge, Lin, and Yu 2019), it is 27ms for step 3 and additional 227ms in step 2. FCANs (Liu et al. 2016) reported its inference time as 150ms. Using 27 integral regions and ResNet50 as a base, the training time for the Aircraft is \sim 4.75 hrs for 150 epochs (12 batch size).

Base CNN	Aircraft/Planes				Stanford Cars				Oxford-IIIT Pets				Param (M)	Time ms
	Base	B+C	B+E	B+C+E	Base	B+C	B+E	B+C+E	Base	B+C	B+E	B+C+E		
ResNet-50	79.7	88.8	81.1	94.9	84.7	91.5	85.7	94.9	86.8	94.7	86.3	96.7	36.9	4.1
Incep. V3	82.4	89.6	83.3	94.8	85.7	91.4	85.7	94.8	90.2	95.3	92.4	96.2	35.1	3.8
Xception	79.5	89.5	89.3	94.1	84.8	91.6	89.1	95.7	91.0	96.2	96.0	97.0	34.2	4.2
NASNet-M	77.1	89.6	80.4	93.8	80.4	91.7	82.7	93.7	89.9	95.6	94.9	97.3	9.7	3.5

Table 4: Performance (accuracy in %) of our model with the addition of our novel CAP (+C) and classification (+E) module to various SotA base (B) CNNs. The observed accuracy trend is (B+C+E) > (B+C) > (B+E) > B for all base CNNs. Final model's (B+C+E) trainable parameters (Param) are given in million (M) and the respective per-frame inference time in millisecond (ms).

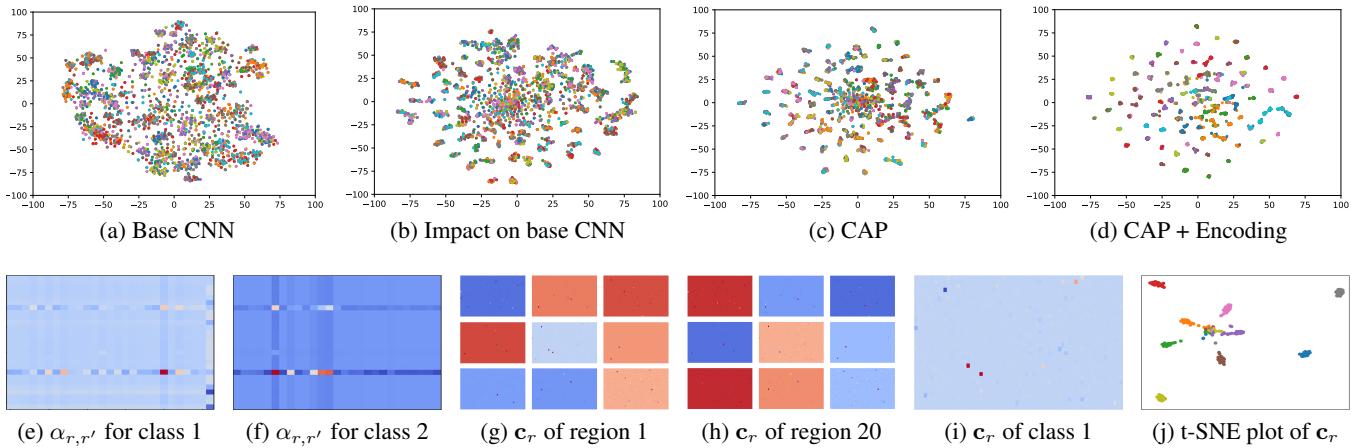


Figure 3: Discriminability using t-SNE to visualize class separability and compactness (a-d). Aircraft test images using Xception: a) base CNN's output, b) our CAP's impact on the base CNN's output, c) our CAP's output, and d) our model's final output. Our CAP's class-specific attention-aware response for class 1 (e) and class 2 (f) to capture the similarity between 27 integral regions (27×27). Class-specific \mathbf{c}_r in (2) for 9 classes (3×3) from region 1 (g) and 20 (h). Blue to red represents class-specific less to more attention towards that region. Class-specific individual feature response within \mathbf{c}_r of the region 1 and class 4 (i). t-SNE plot of \mathbf{c}_r representing images from the above 9 classes (j).

Base CNN	Aircraft			Cars		
	#9	#27	#36	#9	#27	#36
ResNet-50	85.9	94.9	91.2	92.9	94.9	91.9
Xception	87.8	94.1	90.0	93.9	95.7	92.6
NASNet-M	92.7	93.8	90.3	92.4	93.7	90.9

Table 5: Accuracy (%) of our model with a varying number of integral regions. More results in the supplementary in the end.

It is ~ 5.7 hrs for Cars and ~ 8.5 hrs for Dogs.

Qualitative analysis: To understand the discriminability of our model, we use t-SNE (Van Der Maaten 2014) to visualize the class separability and compactness in the features extracted from a base CNN, and our novel CAP and classification modules. We also analyze the impact of our CAP in enhancing the discriminability of a base CNN. We use test images in Aircraft and Xception as a base CNN. In Fig. 3(a-d), it is evident that when we include our CAP + encoding modules, the clusters are farther apart and compact, resulting in a clear distinction of various clusters representing dif-

ferent subcategories. Moreover, the discriminability of the base CNN is significantly improved (Fig. 3b) in comparison to without our modules shown in Fig. 3a. More results are shown in the supplementary material, added in the end. We have also looked the inside of our CAP by visualizing its class-specific attention-aware response using $\alpha_{r,r'}$ and context vector \mathbf{c}_r in (2). Aircraft images (randomly selected 9 classes) are used in Fig. 3(e-j). Such results clearly show our model's power in capturing the context information for discriminating subtle changes in FGVC problems. We have also included some examples, which are incorrectly classified by our model with an explanation in the supplementary information in the end.

Conclusion

We have proposed a novel approach for recognizing subcategories by introducing a simple formulation of context-aware attention via learning where to look when pooling features across an image. Our attention allows for explicit integration of bottom-up saliency by taking advantages of integral regions and their importance, without requiring the bounding box/part annotations. We have also proposed a feature

encoding by considering the semantic correlation among the regions and their spatial layouts to encode complementary partial information. Finally, our model's SotA results on eight benchmarked datasets, quantitative/qualitative results and ablation study justify the efficiency of our approach. Code is available at <https://ardhendubehera.github.io/cap/>.

Acknowledgments

This research is supported by the UKIERI-DST grant CHARM (DST UKIERI-2018-19-10). The GPU used in this research is generously donated by the NVIDIA Corporation.

References

- Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. NetVLAD: CNN architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5297–5307.
- Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; and Le, Q. V. 2019. Attention augmented convolutional networks. In *IEEE International Conference on Computer Vision*, 3286–3295.
- Bossard, L.; Guillaumin, M.; and Gool, L. V. 2014. Food-101 - Mining Discriminative Components with Random Forests. In *Proc. 13th Eur. Conf., Part VI*, volume 8694, 446–461.
- Cai, S.; Zuo, W.; and Zhang, L. 2017. Higher-order integration of hierarchical convolutional activations for fine-grained visual categorization. In *Proceedings of the IEEE International Conference on Computer Vision*, 511–520.
- Chang, D.; Ding, Y.; Xie, J.; Bhunia, A. K.; Li, X.; Ma, Z.; Wu, M.; Guo, J.; and Song, Y.-Z. 2020. The devil is in the channels: Mutual-channel loss for fine-grained image classification. *IEEE Trans. on Image Processing* 29: 4683–4695.
- Chen, Y.; Bai, Y.; Zhang, W.; and Mei, T. 2019. Destruction and construction learning for fine-grained image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5157–5166.
- Chollet, F. 2017. Xception: Deep learning with depthwise separable convolutions. In *IEEE conference on computer vision and pattern recognition*, 1251–1258.
- Cubuk, E. D.; Zoph, B.; Mane, D.; Vasudevan, V.; and Le, Q. V. 2019. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE CVPR*, 113–123.
- Cui, Y.; Song, Y.; Sun, C.; Howard, A.; and Belongie, S. 2018. Large scale fine-grained categorization and domain-specific transfer learning. In *IEEE CVPR*, 4109–4118.
- Cui, Y.; Zhou, F.; Wang, J.; Liu, X.; Lin, Y.; and Belongie, S. 2017. Kernel pooling for convolutional neural networks. In *IEEE conference on computer vision and pattern recognition*, 2921–2930.
- Ding, Y.; Wen, S.; Xie, J.; Chang, D.; Ma, Z.; Si, Z.; and Ling, H. 2020. Weakly Supervised Attention Pyramid Convolutional Neural Network for Fine-Grained Visual Classification. *arXiv preprint arXiv:2002.03353*.
- Dubey, A.; Gupta, O.; Guo, P.; Raskar, R.; Farrell, R.; and Naik, N. 2018a. Pairwise confusion for fine-grained visual classification. In *Euro. Conf. on Computer Vision*, 70–86.
- Dubey, A.; Gupta, O.; Raskar, R.; and Naik, N. 2018b. Maximum-entropy fine grained classification. In *Advances in Neural Information Processing Systems*, 637–647.
- Fu, J.; Zheng, H.; and Mei, T. 2017. Look closer to see better: Recurrent attention convolutional neural network for fine-grained image recognition. In *IEEE conference on computer vision and pattern recognition*, 4438–4446.
- Ge, W.; Lin, X.; and Yu, Y. 2019. Weakly supervised complementary parts models for fine-grained image classification from the bottom up. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3034–3043.
- Ge, W.; and Yu, Y. 2017. Borrowing treasures from the wealthy: Deep transfer learning through selective joint fine-tuning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1086–1095.
- He, K.; Girshick, R.; and Dollár, P. 2019. Rethinking imagenet pre-training. In *Proceedings of the IEEE international conference on computer vision*, 4918–4927.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proc. IEEE conf. Comp. Vis. Patt. Recog. (CVPR)*, 770–778.
- Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2019. Image Captioning: Transforming Objects into Words. In *Advances in Neural Info. Processing Systems*, 11135–11145.
- Hochreiter, S.; and Schmidhuber, J. 1997. Long short-term memory. *Neural computation* 9(8): 1735–1780.
- Hu, T.; Qi, H.; Huang, Q.; and Lu, Y. 2019. See better before looking closer: Weakly supervised data augmentation network for fine-grained visual classification. *arXiv preprint arXiv:1901.09891*.
- Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.
- Huang, L.; Wang, W.; Chen, J.; and Wei, X.-Y. 2019a. Attention on attention for image captioning. In *IEEE International Conference on Computer Vision*, 4634–4643.
- Huang, Y.; Cheng, Y.; Bapna, A.; Firat, O.; Chen, D.; Chen, M.; Lee, H.; Ngiam, J.; Le, Q. V.; Wu, Y.; et al. 2019b. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Advances in Neural Information Processing Systems*, 103–112.
- Jaderberg, M.; Simonyan, K.; Zisserman, A.; et al. 2015. Spatial transformer networks. In *Advances in neural information processing systems*, 2017–2025.
- Ji, Z.; Fu, Y.; Guo, J.; Pang, Y.; Zhang, Z. M.; et al. 2018. Stacked semantics-guided attention model for fine-grained zero-shot learning. In *Advances in Neural Information Processing Systems*, 5995–6004.
- Jiang, S.; Min, W.; Liu, L.; and Luo, Z. 2020. Multi-Scale Multi-View Deep Feature Aggregation for Food Recognition. *IEEE Transaction on Image Processing* 29: 265–276.
- Jiang, W.; Sun, W.; Tagliasacchi, A.; Trulls, E.; and Yi, K. M. 2019. Linearized Multi-Sampling for Differentiable Image Transformation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2988–2997.
- Khosla, A.; Jayadevaprakash, N.; Yao, B.; and Li, F.-F. 2011. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, volume 2.

- Krause, J.; Sapp, B.; Howard, A.; Zhou, H.; Toshev, A.; Duerig, T.; Philbin, J.; and Fei-Fei, L. 2016. The unreasonable effectiveness of noisy data for fine-grained recognition. In *European Conference on Computer Vision*, 301–320.
- Krause, J.; Stark, M.; Deng, J.; and Fei-Fei, L. 2013. 3d object representations for fine-grained categorization. In *IEEE international conf. on computer vision workshops*, 554–561.
- LeCun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 86(11): 2278–2324.
- Leng, J.; Liu, Y.; and Chen, S. 2019. Context-aware attention network for image recognition. *Neural Computing and Applications* 31(12): 9295–9305.
- Li, G.; Zhu, L.; Liu, P.; and Yang, Y. 2019. Entangled Transformer for Image Captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, 8928–8937.
- Li, P.; Xie, J.; Wang, Q.; and Gao, Z. 2018. Towards faster training of global covariance pooling networks by iterative matrix square root normalization. In *IEEE Conference on Computer Vision and Pattern Recognition*, 947–955.
- Lin, C.-H.; and Lucey, S. 2017. Inverse compositional spatial transformer networks. In *IEEE CVPR*, 2568–2576.
- Liu, C.; Xie, H.; Zha, Z.-J.; Yu, L.; Chen, Z.; and Zhang, Y. 2019. Bidirectional Attention-Recognition Model for Fine-grained Object Classification. *IEEE Trans. on Multimedia* .
- Liu, X.; Xia, T.; Wang, J.; Yang, Y.; Zhou, F.; and Lin, Y. 2016. Fully convolutional attention networks for fine-grained recognition. *arXiv preprint arXiv:1603.06765*.
- Luo, W.; Yang, X.; Mo, X.; Lu, Y.; Davis, L. S.; Li, J.; Yang, J.; and Lim, S.-N. 2019. Cross-X Learning for Fine-Grained Visual Categorization. In *IEEE ICCV*, 8242–8251.
- Maji, S.; Rahtu, E.; Kannala, J.; Blaschko, M.; and Vedaldi, A. 2013. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151* .
- Miech, A.; Laptev, I.; and Sivic, J. 2017. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905* .
- Nilsback, M.-E.; and Zisserman, A. 2008. Automated flower classification over a large number of classes. In *Indian Conf. on Comp. Vision, Graphics & Image Processing*, 722–729.
- Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *IEEE CVPR*, 3498–3505.
- Parmar, N.; Ramachandran, P.; Vaswani, A.; Bello, I.; Levskaya, A.; and Shlens, J. 2019. Stand-alone self-attention in vision models. In *Proceedings of NeurIPS*, 68–80.
- Peng, Y.; He, X.; and Zhao, J. 2018. Object-part attention model for fine-grained image classification. *IEEE Transactions on Image Processing* 27(3): 1487–1500.
- Rodríguez, P.; Velazquez, D.; Cucurull, G.; Gonfaus, J. M.; Roca, F. X.; and González, J. 2020. Pay attention to the activations: a modular attention mechanism for fine-grained image recognition. *IEEE Trans. on Multimedia* 22(2): 502–514.
- Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE CVPR*, 4510–4520.
- Simon, M.; and Rodner, E. 2015. Neural activation constellations: Unsupervised part model discovery with convolutional networks. In *IEEE Intl. Conf. on Comp. Vision*, 1143–1151.
- Sun, M.; Yuan, Y.; Zhou, F.; and Ding, E. 2018. Multi-attention multi-class constraint for fine-grained image recognition. In *European Conf. on Computer Vision*, 805–821.
- Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *IEEE CVPR*, 2818–2826.
- Taha, A.; Chen, Y.-T.; Misu, T.; Shrivastava, A.; and Davis, L. 2020. Boosting Standard Classification Architectures Through a Ranking Regularizer. In *The IEEE Winter Conference on Applications of Computer Vision*, 758–766.
- Valan, M.; Makonyi, K.; Maki, A.; Vondráček, D.; and Ronquist, F. 2019. Automated taxonomic identification of insects with expert-level accuracy using effective feature transfer from convolutional networks. *Systematic biology* 68(6): 876–895.
- Van Der Maaten, L. 2014. Accelerating t-SNE using tree-based algorithms. *The Journal of Machine Learning Research* 15(1): 3221–3245.
- Van Horn, G.; Branson, S.; Farrell, R.; Haber, S.; Barry, J.; Ipeirotis, P.; Perona, P.; and Belongie, S. 2015. Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 595–604.
- Wah, C.; Branson, S.; Welinder, P.; Perona, P.; and Belongie, S. 2011. The caltech-ucsd birds-200-2011 dataset .
- Xiao, T.; Xu, Y.; Yang, K.; Zhang, J.; Peng, Y.; and Zhang, Z. 2015. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE CVPR*, 842–850.
- Xie, L.; Zheng, L.; Wang, J.; Yuille, A. L.; and Tian, Q. 2016. Interactive: Inter-layer activeness propagation. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 270–279.
- Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. In *European Conference on Computer Vision (ECCV)*, 420–435.
- Yu, C.; Zhao, X.; Zheng, Q.; Zhang, P.; and You, X. 2018a. Hierarchical bilinear pooling for fine-grained visual recognition. In *European Conference on Computer Vision*, 574–589.
- Yu, F.; Wang, D.; Shelhamer, E.; and Darrell, T. 2018b. Deep layer aggregation. In *IEEE CVPR*, 2403–2412.
- Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318* .
- Zhao, H.; Jia, J.; and Koltun, V. 2020. Exploring self-attention for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 10076–10085.
- Zhao, H.; Yap, K.-H.; Kot, A. C.; and Duan, L. 2020. JDNet: A Joint-learning Distilled Network for Mobile Visual Food Recognition. *IEEE Journal of Selected Topics in Signal Processing* .
- Zheng, H.; Fu, J.; Zha, Z.-J.; and Luo, J. 2019. Looking for the devil in the details: Learning trilinear attention sampling network for fine-grained image recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 5012–5021.
- Zheng, H.; Fu, J.; Zha, Z.-J.; Luo, J.; and Mei, T. 2020. Learning rich part hierarchies with progressive attention networks for fine-grained image recognition. *IEEE Transactions on Image Processing* 29: 476–488.

Zoph, B.; Vasudevan, V.; Shlens, J.; and Le, Q. V. 2018. Learning transferable architectures for scalable image recognition. In *IEEE CVPR*, 8697–8710.

Zoran, D.; Chrzanowski, M.; Huang, P.-S.; Gowal, S.; Mott, A.; and Kohli, P. 2020. Towards robust image classification using sequential attention models. In *IEEE Conference on Computer Vision and Pattern Recognition*, 9483–9492.

Supplementary Document

In this document, we have included the remaining quantitative and qualitative results, which we could not include in the main paper.

Remaining results of Table 2: The performance comparison (accuracy in %) using the remaining two datasets (Stanford Dogs and Oxford Flowers) for Table 2 in the main paper. It is presented in Table 6 below.

Table 6: Performance comparison with the recent top-five SotA approaches on each dataset. Methods marked with * involve transfer/joint learning strategy for objects/patches/regions consisting more than one dataset (primary and secondary)

Stanford Dogs		Oxford Flowers	
Method	Accuracy (%)	Method	Accuracy (%)
FCANs (Liu et al. 2016)	89.0	InterAct (Xie et al. 2016)	96.4
SJFT* (Ge and Yu 2017)	90.3	SJFT* (Ge and Yu 2017)	97.0
DAN (Hu et al. 2019)	92.2	OPAM* (Peng, He, and Zhao 2018)	97.1
WARN (Rodríguez et al. 2020)	92.9	DSTL* (Cui et al. 2018)	97.6
CPM* (Ge, Lin, and Yu 2019)	97.1	MC _{Loss} * (Chang et al. 2020)	97.7
Proposed	96.1	Proposed	97.7

Remaining results of Table 3: The accuracy of the proposed method is evaluated on the **NABirds** dataset using six different SotA base CNNs for Table 3 in the main paper. It is presented in Table 7 below.

Table 7: Our model’s accuracy (%) on the **NABirds** dataset with different SotA base CNN architectures. Previous best accuracy is 86.4% (Luo et al. 2019) for primary only and 87.9% (Cui et al. 2018) for combined primary and secondary datasets.

Base CNN	Accuracy(%)
ResNet-50	88.8
Inception V3	89.1
Xception	91.0
DenseNet-121	88.3
NASNet-Mobile	88.7
MobileNet V2	89.1

Remaining results of Table 4: In ablation study (Table 4 of the main paper), we have presented the performance of the proposed model (with the addition of our novel context-aware attentional pooling (+C) and classification (+E) module) on the Aircraft, Stanford Cars and Oxford-IIIT Pets datasets. The same evaluation procedure is performed on the Stanford Dogs, Oxford Flowers and Caltech Birds (CUB-200) datasets and the recognition accuracy (%) is presented in Table 8. Like in Table 4, a similar trend is observed in the improvement of accuracy when our context-aware attentional pooling (+C) and classification (+E) modules are added to various SotA base CNN architectures (B).

Table 8: Accuracy (%) of the proposed model with the addition of our novel context-aware attentional pooling (+C) and classification (+E) module to various SotA base (B) CNN architectures. It presents the remaining evaluation of Table 4.

Base CNN	Stanford Dogs			Oxford Flowers			Caltech Birds: CUB-200		
	B	B+C	B+C+E	B	B+C	B+C+E	B	B+C	B+C+E
Inception-V3	78.7	94.2	95.7	92.3	94.9	97.6	76.0	87.1	91.4
Xception	82.7	94.8	96.1	91.9	94.9	97.7	75.6	87.4	91.8
DenseNet-121	79.5	94.5	95.5	94.4	95.1	97.6	79.1	87.2	91.6
NASNet-Mobile	79.5	94.7	96.0	90.7	95.0	97.7	73.0	86.8	89.7
MobileNetV2	76.5	94.3	95.9	92.3	95.0	97.4	74.5	87.0	89.2
Previous Best	(Ge et al. 2019)		93.9	(Xie et al. 2016)		96.4	(Ge et al. 2019)		90.3

Remaining results of Table 5: The performance is evaluated using a different number of integral regions on the Aircraft and Stanford Cars datasets (Table 5). The same experiment is also carried out on the Stanford Dogs dataset, and the results are given in Table 9 below.

Table 9: Accuracy (%) of our model with numbers of 9, 27, and 36 integral regions on **Stanford Dogs** dataset.

Base CNN	#9	#27	#36
ResNet-50	90.5	95.8	92.1
Xception	95.3	96.1	95.2
NASNet-M	91.7	96.0	93.3

Top-N Accuracy (%): We have also evaluated the proposed approach using top-N accuracy metric on Oxford-IIIT Pets, Stanford Cars and Aircraft datasets. The performance of our modules on top of various base architectures is presented in Table 10 below. On all three datasets, the top-2 accuracy is around 99% and is independent of the type of base CNN architecture used. Moreover, the top-5 accuracy is nearly 100%. This justifies the significance of our novel attentional pooling and encoding modules in enhancing performance and their wider applicability.

Table 10: Top-N accuracy (in %) of the proposed model using different base architectures on Oxford-IIIT Pets, Stanford Cars and Aircraft datasets. The top-2 accuracy is around 99% and is independent of the type of base CNN architecture used. The top-5 accuracy is nearly 100%. This shows the significance of the proposed attentional pooling and encoding modules.

Dataset	Base CNN architecture	Top 1	Top 2	Top 3	Top 5
Oxford-IIIT Pets	Inception-V3	96.2	99.0	99.5	99.9
	Xception	97.0	99.7	99.9	99.9
	DenseNet121	96.9	99.2	99.6	99.7
	NASNetMobile	97.3	99.4	99.8	99.9
	MobileNetV2	96.4	98.9	99.5	99.6
Stanford Cars	Inception-V3	94.8	99.4	99.7	99.8
	Xception	95.7	99.3	99.7	99.8
	DenseNet121	93.6	98.7	99.5	99.9
	NASNetMobile	93.7	99.1	99.7	99.8
	MobileNetV2	94.0	99.3	99.8	99.9
Aircraft	Inception-V3	94.8	99.1	99.7	99.8
	Xception	94.1	98.9	99.2	99.5
	DenseNet121	94.6	98.8	99.3	99.4
	NASNetMobile	93.8	99.4	99.8	99.8
	MobileNetV2	94.4	99.1	99.7	99.8

Additional Qualitative Analysis:

We have provided the additional qualitative analysis of our model’s performance by selecting a few example images, which are wrongly classified against the label they are mistaken for (selected from the mistaken subcategories). This is presented in Figure 4. It is evident that the mistaken labels come from classes with extremely similar features, often being from the same manufacturer (Boeing 747, Audi, etc.). We have also noticed that subcategories can have very specific defining features that are not clearly visible in every image due to poor angles or lighting conditions (e.g. The chin of a Ragdoll and legs of a Birman cat shown in Fig. 4g).



(a) 747-200 vs 747-100

(b) 747-300 vs 747-400

(c) C-47 vs DC-3



(d) Audi TTS Coupe 2012 vs Audi TT RS Coupe 2012

(e) Bentley Continental GT Coupe 2012 vs Bentley Continental GT Coupe 2007

(f) Chevrolet Express Cargo Van 2007 vs Chevrolet Express Van 2007



(g) Birman vs Ragdoll

(h) American Pitbull Terrier vs American Bulldog

(i) Staffordshire Bull Terrier vs American Pitbull Terrier



(j) Spotted Catbird vs Gray Catbird

(k) Red Winged Blackbird vs Brewer Blackbird

(l) Laysan Albatross vs Sooty Albatross



(m) English Marigold vs Dandelion

(n) Sweet Pea vs Lenten Rose

(o) Clematis vs Hibiscus

Figure 4: Some of the example images, which are incorrectly classified by our model (left) against the label they are mistaken for (right - selected from the mistaken subcategories): Aircraft (a-c), Stanford Cars (d-f), Oxford-IIIT Pets (g-i), Caltech-UCSD Birds - CUB-200 (j-l), and Oxford Flowers (m-o). It can be seen that the mistaken labelling comes from classes with extremely similar appearance features and/or perspective changes, often being from the same manufacturer (Boeing 747, Audi, etc.). We have also noticed that subcategories can have very specific defining features that are not clearly visible in every image due to poor angles or lighting conditions (e.g. The chin of a Ragdoll and legs of a Birman cat).

We have also included an additional qualitative analysis of discriminating ability (Figure 5 to Figure 9) of our model using t-SNE to visualize class separability and compactness on the different datasets as well as various backbone CNNs.

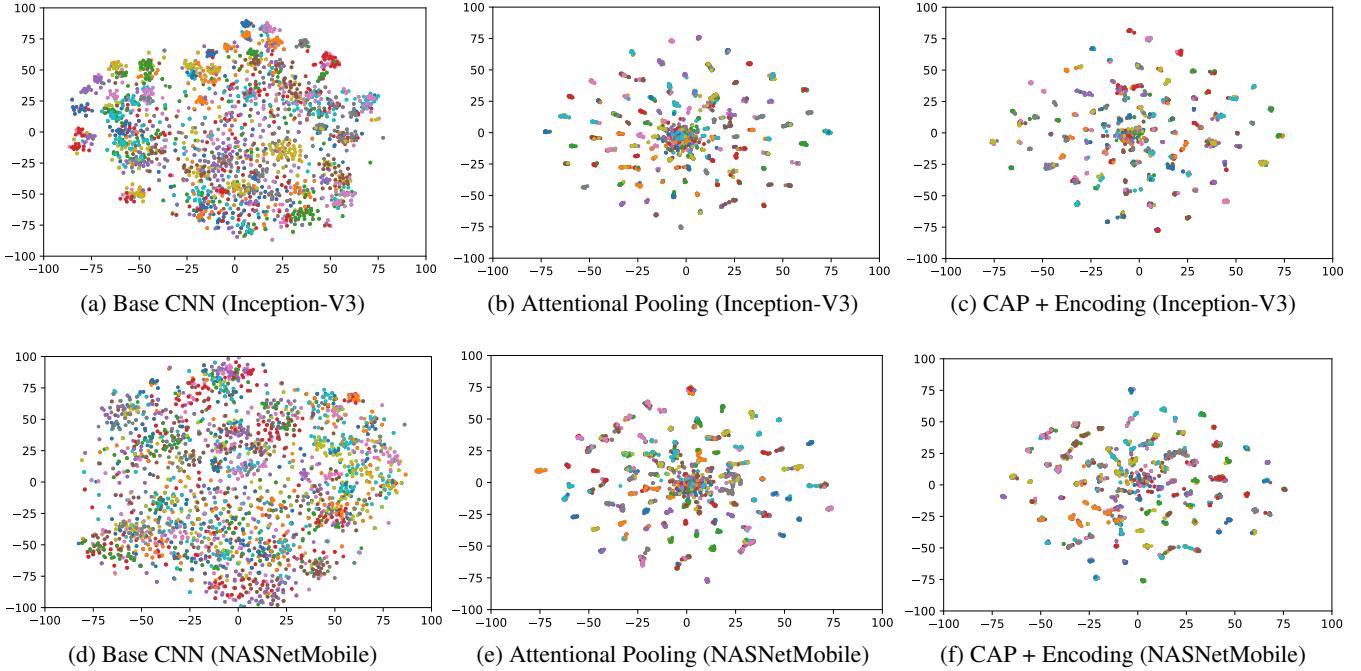


Figure 5: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of **Aircraft** test images using Inception-V3 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model's final feature maps (CAP+Encoding). Best view in color.

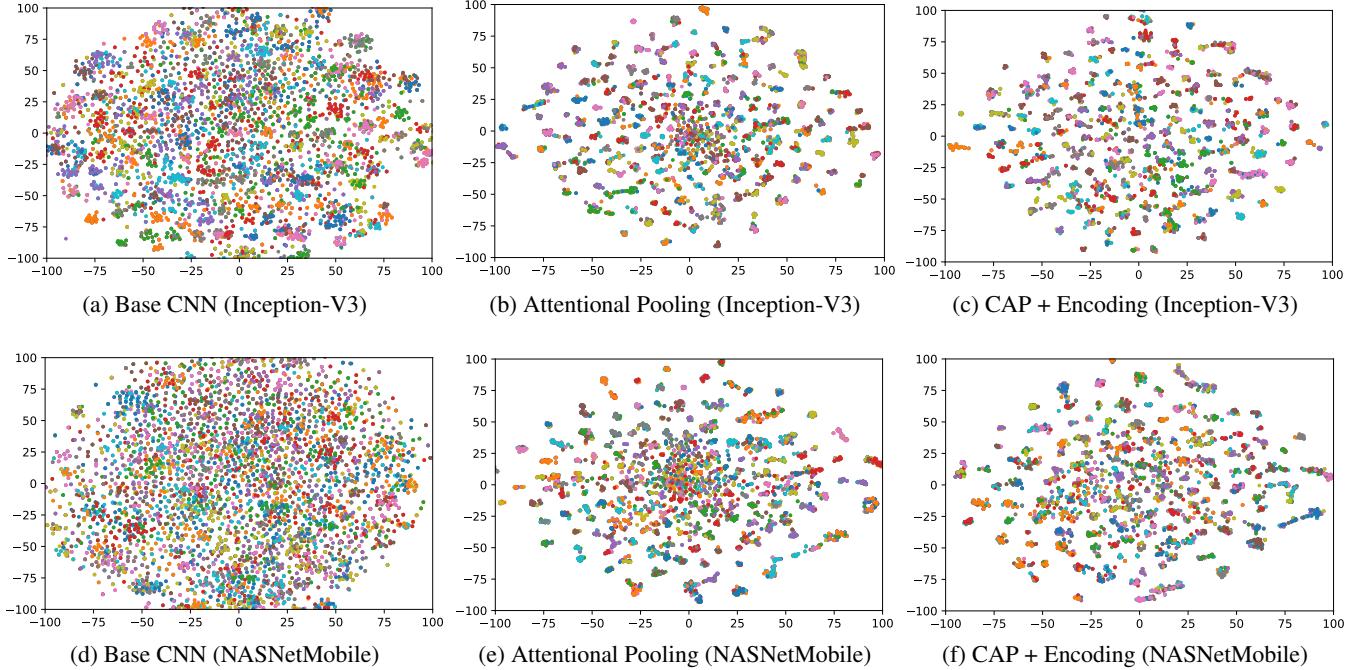


Figure 6: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of **Stanford Cars** test images using Inception-V3 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model’s final feature maps (CAP+Encoding). Best view in color.

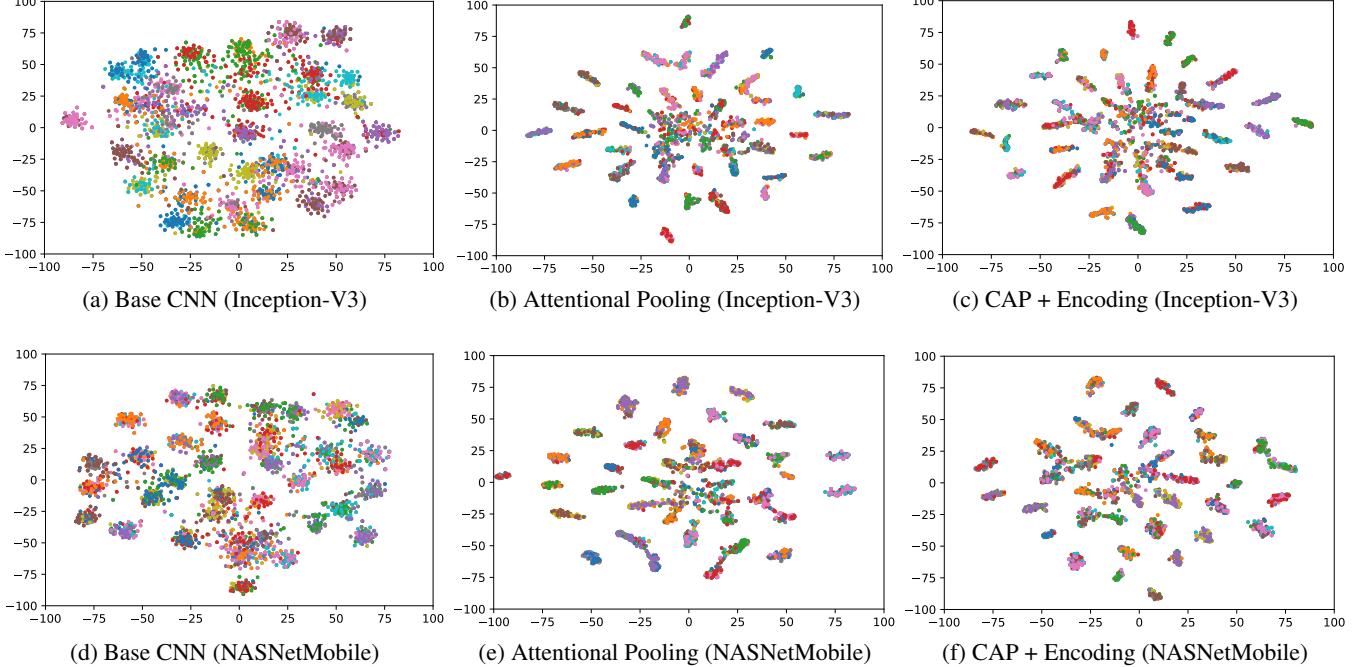


Figure 7: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of **Oxford-IIIT Pets** test images using Inception-V3 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model’s final feature maps (CAP+Encoding). Best view in color.

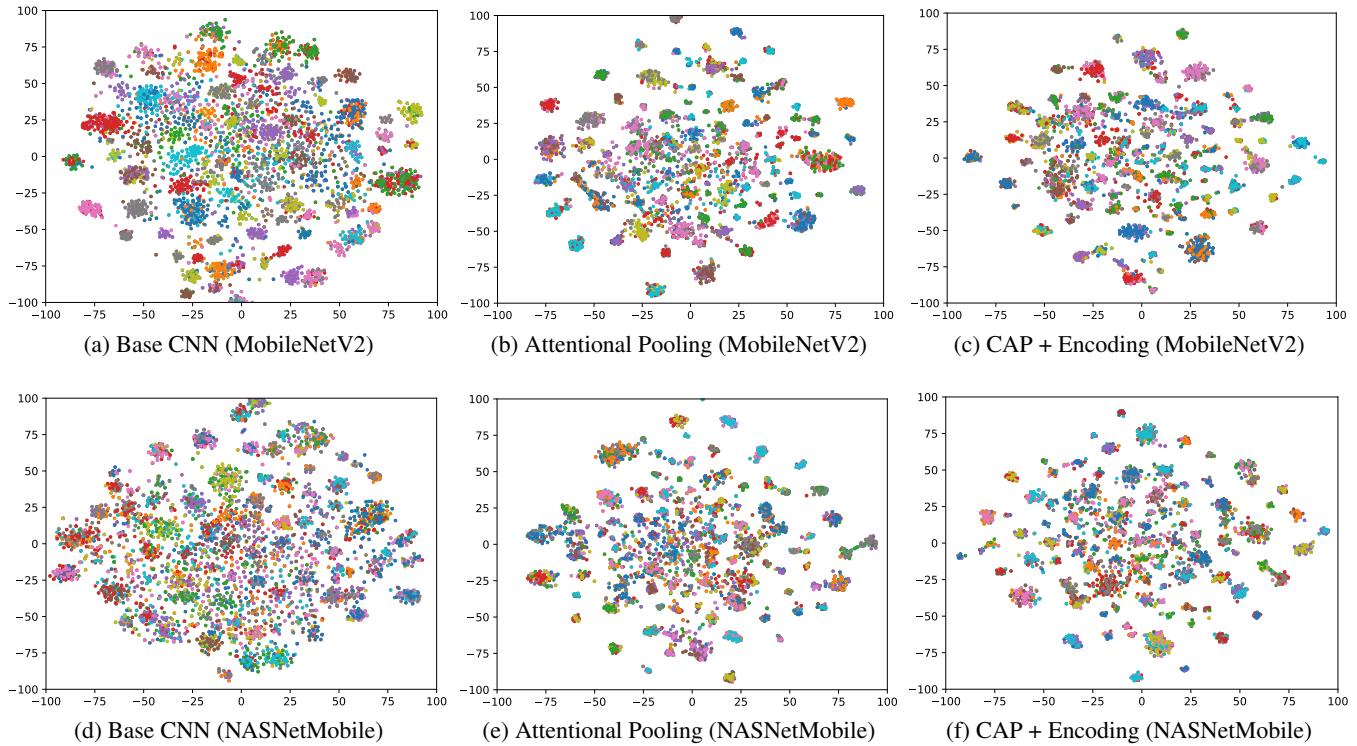


Figure 8: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of **Oxford Flowers** test images using MobileNetV2 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model's final feature maps (CAP+Encoding). Best view in color.

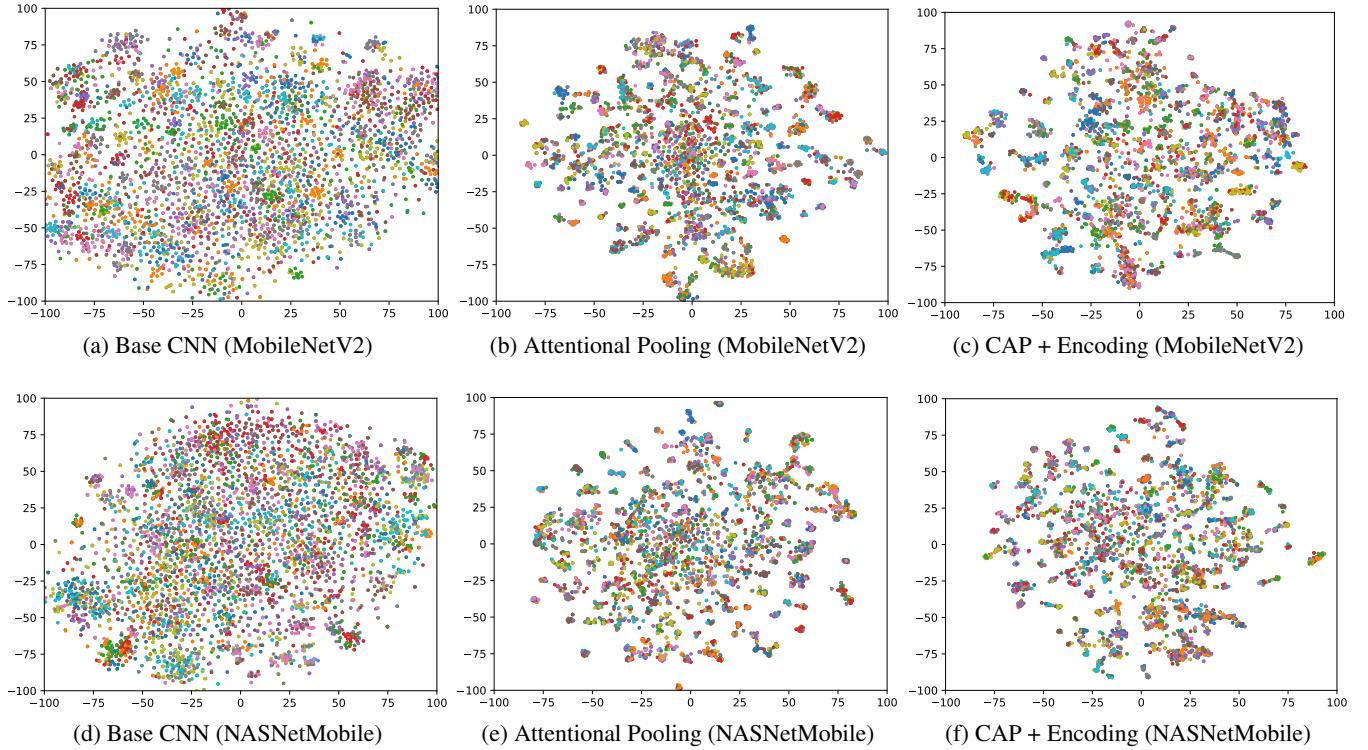


Figure 9: Qualitative analysis of discriminating ability using t-SNE to monitor class separability and compactness. Visualization of the **Caltech-UCSD Birds (CUB-200)** test images using MobileNetV2 and NASNetMobile as a base CNN: (a & d) output of the base CNN, (b & e) feature maps from our attentional pooling (CAP), and (c & f) our model’s final feature maps (CAP+Encoding). Best view in color.