COLORING DEEP CNN LAYERS WITH ACTIVATION HUE LOSS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes a novel hue-like angular parameter to model the structure of deep convolutional neural network (CNN) activation space, referred to as the *activation hue*, for the purpose of regularizing models for more effective learning. The activation hue generalizes the notion of color hue angle in standard 3-channel RGB intensity space to N-channel activation space. A series of observations based on nearest neighbor indexing of activation vectors with pre-trained networks indicate that class-informative activations are concentrated about an angle θ in both the (x, y) image plane and in multi-channel activation space. A regularization term in the form of hue-like angular θ labels is proposed to complement standard one-hot loss. Training from scratch using combined one-hot + activation hue loss improves classification performance modestly for a wide variety of classification tasks, including ImageNet.

1 INTRODUCTION

The success of deep convolutional neural networks (CNNs) is largely due to trainable multi-channel filter banks and the informative activation spaces they produce Krizhevsky et al. (2012). A significant body of research thus focuses on understanding the activation space for the purpose of improving or regularizing models for more effective learning. Examples include geometrical regularization based on hyperspheres Mettes et al. (2019); Shen et al. (2021), enforcing constant radial distance from the feature space origin Zheng et al. (2018) or angular loss between prototypes Wang et al. (2017). Similar works try to leverage this activation space after training by extracting specific features with various optimized methods Kornblith et al. (2019); Azizpour et al. (2015); Cimpoi et al. (2016).

Our work is inspired by the well-known fact that in standard three-channel (red, green, blue) intensity space, human color perception is largely determined by the angular hue parameter. Might an analogous hue-like parameter play a similarly important role in image classification from multi-channel activation space? Our contribution is to propose a novel angular parameter θ, which we refer to as the *activation hue*, that can be used to model and regularize activation space. The activation hue may be viewed as a generalization of the standard hue parameter from 3D red-green-blue (RGB) color space to general multi-dimensional activation space, as shown in Figure 1. The RGB space may be viewed as a 3D cube, where colorless pixels lie along a medial greyscale axis signifying a maximum entropy or uniform distribution. The hue angle θ is measured in the plane perpendicular to the greyscale axis and thus encodes the bias of a lower entropy non-uniform distribution towards a dominant color. The multi-channel activation space of a CNN layer may be thought of analogously, i.e. uniform or uninformative vectors lie along a medial axis in activation space, similar to the RGB greyscale line, and a class-informative activation hue angle θ may be defined in the plane perpendicular to the uniform axis.

Observations and experiments demonstrate the role of the activation hue in both pre-trained networks and model training from scratch. Initial observations investigate classification via memory-based indexing of activations Cover & Hart (1967) using generic networks pre-trained on the ImageNet dataset Deng et al. (2009) with standard one-hot loss, focusing on transfer learning from activations in bottleneck layers similarly to Zeiler & Fergus (2014); Lenc & Vedaldi (2015). Distributions of correct indexing solutions exhibit consistent class-specific bias towards an angular direction θ in both the image plane and activation space. Training from scratch using a combined one-hot + activation

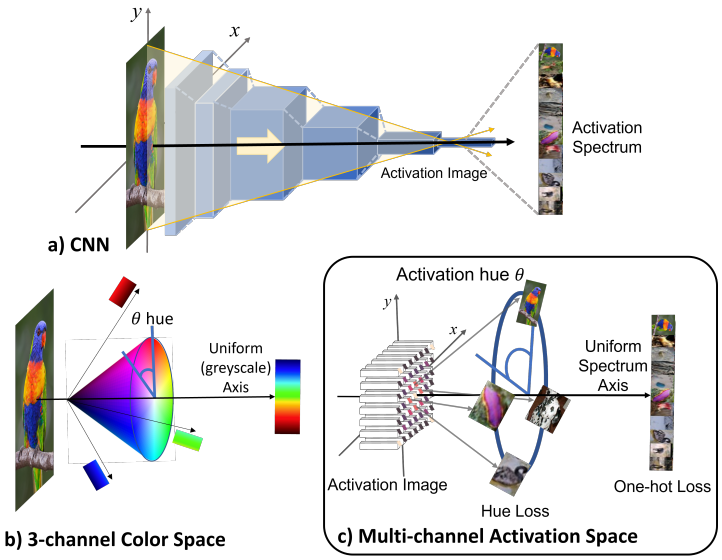

Figure 1: Illustrating Activation Hue in deep CNN layers. a) Shows a deep neural network (CNN-based encoder), where information is increasingly concentrated into a class-informative activation distribution via standard loss from one-hot labels. b) Shows the standard RGB space with color characterized by a hue angle θ. c) Our proposed model of activation hue angle θ with respect to the (x, y) image plane and activation space $I(x, y) \in R^N$. Deep CNN activations exhibit consistent angular bias or activation hue θ according to the class (e.g., the rainbow lorikeet), and a novel activation hue loss improves classification for a variety of networks and classification tasks.

hue loss function leads to consistently higher accuracy across a variety of CNN architectures and classification tasks of varying specificity. In order to draw the broadest possible conclusions, we use a variety of architectures (ResNet He et al. (2016), DenseNet Huang et al. (2016), Inception Szegedy et al. (2016), VGG Simonyan & Zisserman (2014)) and various in and out-of-distribution datasets specifically not used in training, including textures (DTD) Cimpoi et al. (2014), specific object datasets: sex and family classification from brain MRI (HCP) Van Essen et al. (2013), sex from face images (UTKFaces) Zhang et al. (2017), dog species (Stanford Dogs) Khosla et al. (2011), and general object categories including Caltech 101 Fei-Fei et al. (2006) and Imagenet Deng et al. (2009).

2 Related Work

We propose to generalize the notion of hue from standard RGB color space to the space of deep CNN activations, which to our knowledge is novel in the computer vision and machine learning literature. In standard notion in tri-chromatic color analysis Fairchild (2013), hue is closely related to human color perception. It is represented by an angle $\theta \in [0, 2\pi]$ in the 2D plane perpendicular to the greyscale axis or uniform color spectrum. Whereas standard color hue has been modeled using deep networks Flachot et al. (2022); Avi-Aharon et al. (2019), we propose activation hue as an analogy to hue in general multi-channel CNN activation space.

In our work, we consider a generic CNN activation layer an vector-valued image $I_{t,\bar{x}}$ or $I(t, \bar{x})$, where $I \in R^N$ is an N-channel activation image, $\bar{x} = (x, y) \in R^2$ are 2D pixels coordinates centered upon $(x, y) = (0, 0)$ and $t \in R^1$ represents the CNN layer. Activation hue is a single unitary $U(1)$ variable defined by (x, y) coordinates constrained to the unit circle $x^2 + y^2 = 1$ or equivalently an angle $\theta = atan2(y, x)$. Unitary variables are well known in mathematical analysis and increasingly used in machine learning formulations Kiani et al. (2022); Tang et al. (2021), however our model of a hue-like angle in activation space is unique in the literature.

Our preliminary observations of activation hue in pre-trained networks are based on nearest neighbor (NN) indexing and classification Cover & Hart (1967), specifically using spatially-localized activation vectors. This follows the transfer learning approach, where networks pre-trained on large generic datasets such as ImageNet Deng et al. (2009) are used as general feature extractors for new tasks Kornblith et al. (2019); Azizpour et al. (2015); Cimpoi et al. (2016). Deep bottleneck activations tend to outperform specialized shallower networks and meta-learning methods Chen et al. (2019), particularly in the case of few training data and a large domain shift between training and testing data Guo et al. (2020). Various approaches seek to adapt ImageNet models to fine-grained tasks by encoding activations at bottleneck layers, such as via descriptor information (e.g. extracted off-the-shelf features Sharif Razavian et al. (2014), VLAD Arandjelović et al. (2016)), global average or max pooling Razavian et al. (2016), generalized mean (GeM) Radenović et al. (2018), regional max pooling (R-MAC) Tolias et al. (2015) in intermediate layers or modulated by attention operators Noh et al. (2017). Additional training may consider joint loss between classification and instance retrieval terms Berman et al. (2019). The mechanism of spatially localized activations (as opposed to global descriptors) is closely linked to the attention mechanisms Huang et al. (2019), including non-local networks Wang et al. (2018), squeeze-and-excitation networks Hu et al. (2018), transformer architectures Vaswani et al. (2017); Carion et al. (2020); Han et al. (2020) including hierarchically shifted windows Liu et al. (2021), thin bottleneck layers Sandler et al. (2018), self-attention mechanisms considering locations and channels Woo et al. (2018), intra-kernel correlations Haase & Amthor (2020), multi-layer perceptrons incorporating Euler's angle Tang et al. (2021), correspondence-based transformers Jiang et al. (2021) and detectors Sun et al. (2021), and geometrical embedding of spatial information via graphs Kipf & Welling (2016); Henaff et al. (2015). Whereas these works typically seek end-to-end learning solutions fitting within GPU memory constraints Gordo et al. (2016), we first demonstrate the hue-regularized model in basic memory lookup observations, then propose a novel loss function based on activation hue.

Our final results training from scratch using one-hot + hue loss are similar in spirit to work seeking to regularize the label and/or the activation space, including using real-valued rather than strictly one-hot training labels Rodríguez et al. (2018), learning-based classifiers Wang et al. (2019); Wen et al. (2016), prototypical networks for few-shot learning Nguyen et al. (2020); Snell et al. (2017), deep k-nearest neighbors Papernot & McDaniel (2018), geometrical regularization based on hyperspheres Mettes et al. (2019); Shen et al. (2021), enforcing constant radial distance from the feature space origin Zheng et al. (2018) or angular loss between prototypes Wang et al. (2017). We seek to present our theory in the general context, we deliberately eschew architectural modifications that might limit the generality of our analysis. We demonstrate our model using a wide variety of pre-trained off-the-shelf CNN architectures including DenseNet Huang et al. (2016), Inception Szegedy et al. (2016), ResNet He et al. (2016), VGG Simonyan & Zisserman (2014) directly imported from TensorFlow Abadi et al. (2015). We consider a variety of testing datasets both used and not used in ImageNet Deng et al. (2009) training (datasets in and out of distributions), including general categories (e.g. Caltech 101 Fei-Fei et al. (2006)), and specific instances (e.g. birds Welinder et al. (2010), human brain MRIs of family members Van Essen et al. (2013), faces Zhang et al. (2017)).

3 Preliminary Observations from Pre-trained Networks

We begin by presenting several novel observations regarding the structure of activation information in generic deep CNN layers, which are not widely known in the computer vision community, and motivate our model of activation hue in the following section. Our observations are based on rudimentary nearest neighbor classification Cover & Hart (1967), specifically using spatially-localized *activation vectors* $I(\bar{x}) \in R^N$ of bottleneck layers of Imagenet pre-trained networks, and stored in a memory along with their $\bar{x} = (x, y)$ positions. Using pixel-level activations rather than spatially pooled or flattened features leads to improved classification and allows observation of the fine structure of activation information with respect to the geometry of image space. In order to make generally pertinent observations, we consider a variety of generic CNN architectures trained on the ImageNet dataset Deng et al. (2009), and tested in basic memory-based retrieval settings using various in and out-of-distribution datasets specifically not used in training, including general objects (Caltech 101 Fei-Fei et al. (2006)), textures (DTD) Cimpoi et al. (2014), and specific object datasets: sex and family classification from brain MRI (HCP) Van Essen et al. (2013), sex from face images (UTKFaces) Zhang et al. (2017), and dog species (Stanford Dogs) Khosla et al. (2011).

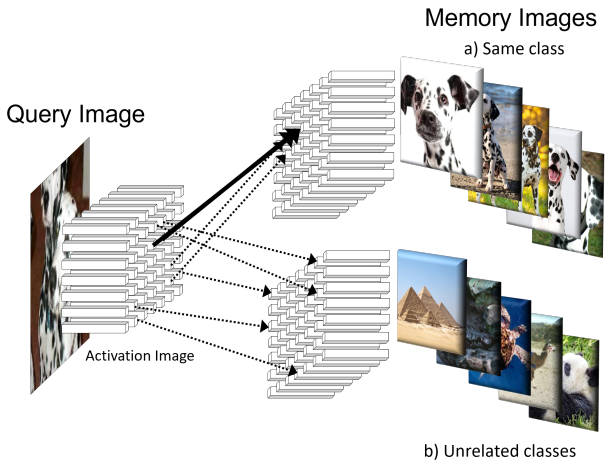

Figure 2: Our proposed retrieval architecture where classification is achieved from K=10 nearest neighbor matches between pixel-wise activation vectors (white bars) derived from a query image and labeled images stored in memory. Matches between images of the same class (a) generally exhibit lower spatial variability than (b) matches to unrelated class images, focusing towards a specific angular θ position relative to the image center $(x, y) = (0, 0)$.

Nearest Neighbor Retrieval: Figure 2 illustrates our proposed retrieval architecture, where individual vectors $I_{t,\bar{x}}$ in a query image are matched with vectors in memory extracted from images of similar classes. Individual vectors may match to similar vectors at any image locations in order to achieve classification via the procedure described below.

Classification is achieved by maximising the likelihood function $p(I|C, \{I'\})$ of class C associated with input image I from set of image examples $\{I'\}$ stored in memory:

$$C^* = \underset{C}{\operatorname{argmax}}\, p(I|C, \{I'\}), \tag{1}$$

where C^* is a maximum likelihood estimate of the image class. The likelihood function is defined as follows. An activation image I of resolution $W \times H$ is represented as a set $I = \{I_1, \ldots, I_i, \ldots, I_{W\times H}\}$ of N-channel activation vectors I_i located at pixel index i. Similarly, $\{I'\} = \{I'_1, \ldots, I'_j, \ldots, I'_{M\times W\times H}\}$ represents a set of pixel-wise activation vectors I'_j from M activation images stored in memory. For each input pixel vector I_i, a set of K nearest neighbors NN_i is defined as $NN_i : \{j : \|I_i - I'_j\| \leq \|I_i - I'_k\|\}$, where I'_k is the k^{th} nearest neighbor of I_i in memory. The likelihood may be expressed as a kernel density as a sum over input pixels i and nearest neighbors NN_i

$$p(I|C, \{I'\}) \propto \frac{\sum_i^{W\times H} \sum_{j\in NN_i} f(I_i, I'_j)[C = C'_j]}{\sum_j [C = C'_j]}, \tag{2}$$

where in Equation equation 2, $f(I_i, I'_j)$ is a kernel function, $[C = C'_j]$ is the Iverson bracket evaluating to 1 upon equality and 0 otherwise and the denominator normalizes for class frequency across the entire memory set $\sum_j [C = C'_j]$. The kernel function $f(I_i, I'_j)$ is based on activation vector (dis)similarity and is defined as:

$$f(I_i, I'_j) = exp - \left\{ \frac{\|I_i - I'_j\|^2}{\alpha_i^2 + \epsilon} \right\}, \tag{3}$$

where in Equation equation 3, $\alpha_i = \min_j \|I_i - I'_j\|$ is an adaptive kernel bandwidth parameter defined as the distance to nearest activation vector in memory $I'_j \in \{I'\}$, and ϵ is a small positive constant ensuring a non-zero denominator. Note all activation vectors are normalized to unit length $\|I_i\| = \|I'_j\| = 1$.

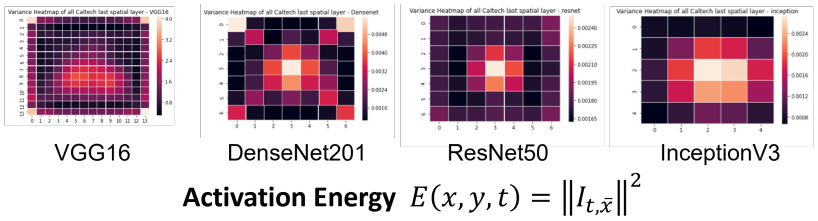

Figure 3: Activation energy maps $E_{t,\bar{x}} = \|I_{t,\bar{x}}\|^2$ computed from spatial bottleneck layers of various ImageNet pre-trained networks and one-hot loss on the Caltech 101 Fei-Fei et al. (2006) dataset from the pixel-level K-nearest neighbor indexing setup.

Observation 1: Activation energy is concentrated symmetrically about the (x, y) image center. Figure 3 shows the scalar energy $E_{t,\bar{x}} = \|I_{t,\bar{x}}\|^2 \in R^+$ of bottleneck activation layers $I_{t,\bar{x}}$ from a variety of generic architectures pre-trained on the ImageNet dataset Deng et al. (2009) in response to input images not used in training but with a similar distribution, here classes of the Caltech 101 Fei-Fei et al. (2006). Note how the activation energy maps trained via standard cross-entropy loss and one-hot labeling are generally symmetric and concentrated in the center of the image plane center. However, successful classification requires that the activation distribution exhibit an asymmetric bias towards an angle θ shared across individuals of a class, effectively breaking this symmetry, and this motivated us to investigate ways to leverage this hue-like observed behaviour in from-scratch training of models. We note that centrally concentrated activation energy may be in part due to the object-centered nature of datasets such as ImageNet or Caltech 101, however our experiments training from scratch extend also to non-object centered datasets such as DTD textures Cimpoi et al. (2014).

Observation 2: Pixel-wise activations lead to the highest classification accuracy. Figure 4 establishes baseline classification results across CNN architectures and descriptors, showing that localized pixel vector matching leads to the highest accuracy amongst alternatives, particularly for DenseNet Huang et al. (2016). The high accuracy of pixel vector matching is notably due to the trade-off between memory and computation (e.g. $7x7 = 49$ pixels vs. 1 global descriptor in the case of DenseNet). Note that our goal is to observe the asymmetry via accurate spatially localized activations, and efficiency is not an immediate concern in our work here. Nevertheless relatively efficient retrieval is achieved using the Approximate Nearest Neighbor library (Annoy) Bernhardsson (2015) indexing method with a rapid tree-based algorithm of $O(log\ N)$ query complexity for N elements in memory, and further efficiency could be achieved via compression (e.g PCA F.R.S. (1901)) or specialized architectures Sun et al. (2021); Jiang et al. (2021).

Observation 3: Class-specific activation information is concentrated according to an angle θ about the (x, y) image center. Figures 5 shows example distributions of NN activation matches, based on 1920-dimensional activation vectors following ReLu $I_x \in R^{1920+}$ from the $7 \times 7 = 49$-pixel bottleneck layer of an ImageNet-pre-trained DenseNet-201 Huang et al. (2016) architecture and test images from the Caltech 101 dataset Fei-Fei et al. (2006) not used in training, but in-distribution for the trained network. Note how distributions of NN pixel vector matches between images of the same class (Figures 5 a) are consistently biased towards a similar angle θ relative to the image center, for both specific classes and individuals, validating class-specific angular bias. Matches to unrelated classes tend to be scattered about the image periphery (Figures 5 b).

Observation 4: Activations are highly informative regarding (x, y) pixel location. Figure 6 shows distributions of nearest neighbor match locations conditioned on query pixel locations in order to understand the variability with respect to correct (same class a) vs. incorrect (unrelated class b) matches. Matches in both cases generally tend to be concentrated about the query pixel location, indicating that each pixel occupies a center in activation space, and that neighboring pixels in image space map to neighboring centers in activation space. Activation matching between instances of the same class a) exhibit much less variability than those between unrelated classes b). Our proposed activation hue loss in the following section seeks to minimize this source of variability.

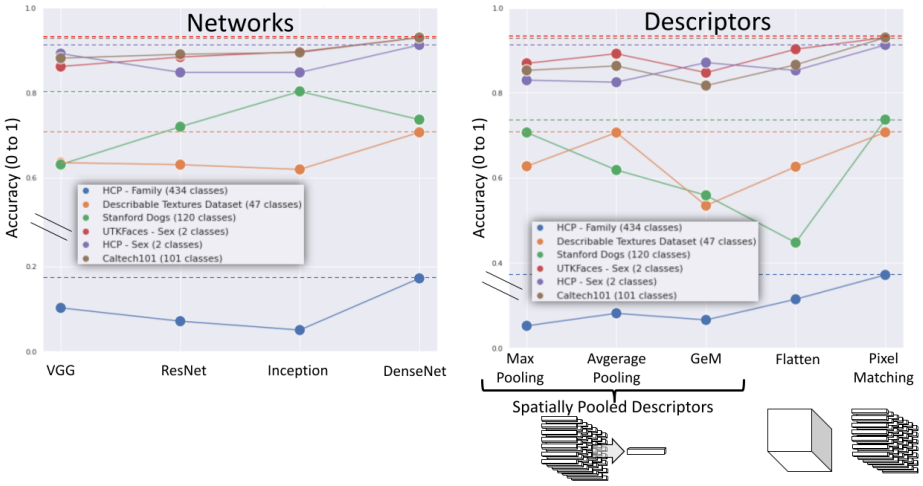

Figure 4: Baseline classification performance of architectures and descriptors on six memory-based classification tasks with $(K = 10)$ nearest neighbors. Dashed lines indicate the superior performance of the DenseNet201 Huang et al. (2016) architecture (max in 5 of 6 tasks) (left) and pixel vector descriptors (max for all) (right).

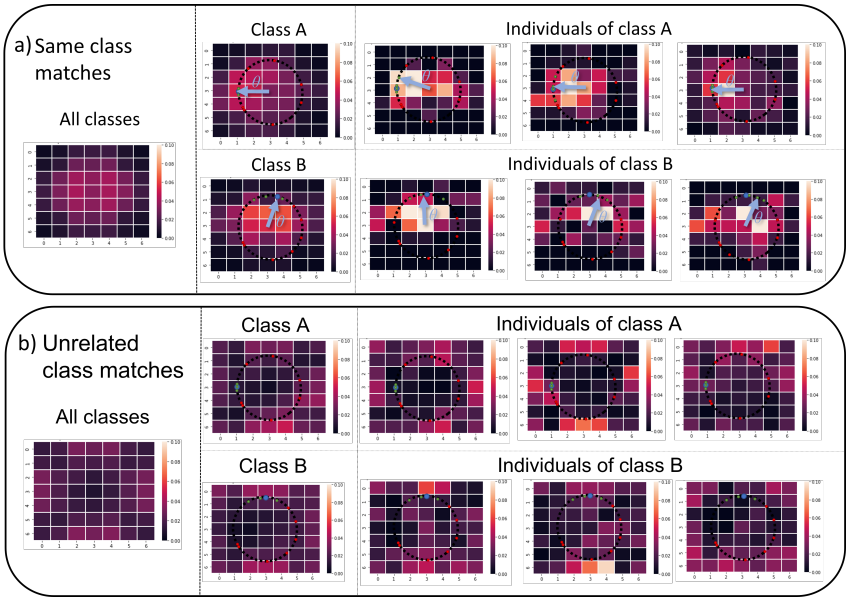

Figure 5: Distributions $p(x, y)$ of matching NN pixel locations a) Matches between activations from the same class are concentrated near the image center $(0, 0)$ and consistently biased towards a similar angle θ for specific classes and individuals of the same class (e.g. class A and B, blue arrows). b) Matches between images of unrelated classes are scattered about the periphery.

4 METHOD

Observations from the previous section revealed that while CNN activation energy or magnitude in deep layers is generally concentrated symmetrically about the center of the (x, y) image plane (Observation 1), activation information for specific classes tends to be concentrated asymmetrically and according to angle θ with respect to the image center as in Figure 5 (Observation 3). Further-

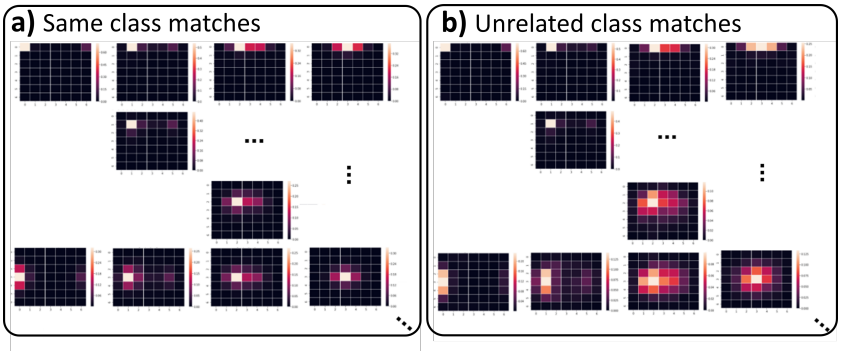

Figure 6: Conditional distributions $p(x, y|\bar{x}_i)$ of matching NN pixel locations given query location $\bar{x}_i$. Distributions are shown for a selection of individual query pixel locations $\bar{x}_i$. Note tight concentration around the original query pixel locations x_i (brightest pixels), and lower variance for matches to instances of the same class (top) vs. unrelated class (bottom). Variations for same class a) are generally stronger in tangential (as opposed to radial) directions, indicating hue-like angular θ deviations about the center.

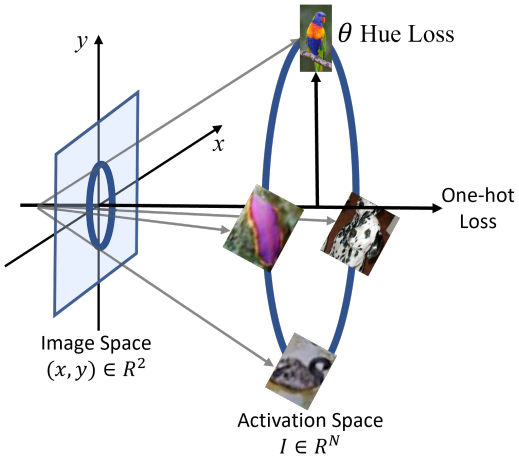

Figure 7: Illustration of our hue loss complementing the one-hot loss through an angular regularization term in order to model the activation hue θ of the image signal in the activation space, similar to the hue angle used to represent colors in the RGB space.

more, activation vectors are highly informative regarding spatial image location (x, y), i.e. nearest neighbor pixel vector activations are tightly concentrated about the original pixel location as in Figure 6 (Observation 4). Together, these observations lead us to hypothesize that in deep CNN layers, class-informative activations tend to follow a hue-like angle in both the image plane and activation space.

Figure 7 illustrates the geometry of activation hue in a deep CNN layer, including the $(x, y) \in R^2$ image space and the multi-channel activation space $I \in R^N$. Multi-channel activation information is non-negative following rectification (ReLu) Nair & Hinton (2010), and thus located in the positive orthant of activation space, i.e. within the larger circle of Figure 7. Standard CNN training from one-hot loss labels provides no information regarding the arrangement of activation information in the (x, y) image plane and may be considered to operate orthogonally to the image plane. Nevertheless, preliminary observations revealed that class-specific activations tend to be distributed according to a hue-like angular variable θ in both the image space and activation space. We refer to this variable as activation hue and note that it characterizes dominant activation channels analogously to how hue characterizes color in RGB space.

We thus propose a novel loss function leveraging activation hue, that may be used as a training signal in the (x, y) image plane of arbitrary CNN layers, most notably bottleneck layers with minimal spatial extent. We first note that the ubiquitous one-hot loss $L_{one-hot}$ used imposes no constraint regarding the geometry of activations in (x, y), and may thus be considered as operating symmetrically across the CNN bottleneck layer. We thus propose augmenting standard loss functions such as cross-entropy-based one-hot labels $L_{one-hot}$ with an additional loss term L_{hue} based on angular class label $\theta_c = \{\theta_{cx}, \theta_{cy}\} \in U(1)$ as follows:

$$L = L_{one-hot} + L_{hue} = -\sum_{c=1}^{M}(\log \frac{\exp(x_c)}{\sum_{i=1}^{M}\exp(x_i)} + \log \frac{\exp(-d\theta_c)}{\sum_{i=1}^{M}\exp(-d\theta_i)})y_c, \tag{4}$$

where M is the number of classes, y_c is a binary indicator equal to 1 for a correct class and 0 otherwise and x_c is the one-hot prediction vector for class c. In our proposed loss L_{hue}, $d\theta_c$ is the angular difference computed as the Euclidean distance between predicted θ'_c and assigned θ_c angular labels:

$$d\theta_c = \frac{\sqrt{(\theta_{cx} - \theta'_{cx})^2 + (\theta_{cy} - \theta'_{cy})^2}}{2 \times M} \tag{5}$$

Note that similarly to how classes are assigned an arbitrary one-hot label index, in training experiments here an angular label is assigned to each class c from a set of M angles equally distributed over the range $\theta_c = [0, 2\pi]$.

5 EXPERIMENTS

Our work hypothesizes that information in deep neural network activation layers may be characterized analogously to color hue in RGB space, and motivates our model of a hue-like angle θ linking (x, y) image space and activation space. This hypothesis was based on initial observations from nearest neighbor indexing trials, where activations in generic pre-trained CNNs exhibited noticeable class-specific angular bias. Here, we show how training with an additional hue-inspired component to the loss generally improves classification accuracy for a variety of tasks.

In the previous sections, we observed consistent class-related angular bias in activation layers of pre-trained networks, which was was surprizing as standard cross-entropy loss with one-hot class labels provides no explicit training mechanism for achieving this. We thus hypothesized that an additional L_{hue} loss term based angular class labels θ_c as in Equation equation 4 might improve classification. We performed experiments comparing standard cross-entropy with one-hot labels alone with a combined loss function regularized by hue loss and an additional angular training label θ_c. As in Equation equation 4, hue loss uses a single angular training label θ_c assigned randomly for each class as a pair $\theta_c = (\theta_{cx}, \theta_{cy})$ of point coordinates constrained to the unit circle $\theta_{cx}^2 + \theta_{cy}^2 = 1$, and predicted via fully-connected layers immediately following the network bottleneck. The hue loss L_{hue} is estimated as the L2 distance between angular labels θ_c and predicted (x, y) parameters as in Equation equation 4 and mixed in equal weighting with the one-hot difference to generate a combined cross-entropy loss for the backward step. We trained over a fixed 100 epochs, with original train, validation, and test sets and 5-fold cross-validation on a single Titan RTX GPU. We used the Adam optimizer Kingma & Ba (2014) and CosineAnnealingLR scheduler Loshchilov & Hutter (2016) with default parameters from PyTorch Paszke et al. (2019) in all experiments, along with basic data augmentation in the form of horizontal flips and random crops. Training and classification were evaluated in diverse few-shot learning tasks of varying degrees of granularity, including the Describable Textures Dataset (DTD) Cimpoi et al. (2014), Caltech-UCSD Birds Welinder et al. (2010), Stanford Dogs Khosla et al. (2011), Flowers Nilsback & Zisserman (2008), Pets Parkhi et al. (2012), Indoor67 Quattoni & Torralba (2009), FGVC-Aircraft Maji et al. (2013), Cars Krause et al. (2013), and ImageNet Deng et al. (2009).

Training with one-hot + hue loss improved classification for all tested networks compared to one-hot alone. Table 1 reports results for the network architectures leading to the highest overall accuracy (EfficientNet-B0 Tan & Le (2019)) and the most improved (Resnet-18 He et al. (2016)). Similar improvements were observed with a variety of different network architectures including VGG, Inception v3, DenseNet (as shown with the activation energy in 3).

Table 1: Classification results training from scratch, comparing our proposed one-hot + hue loss with conventional loss based on one-hot encoding with ResNet and EfficientNet baselines. Combined one-hot + hue loss resulted in improved classification in all cases tested (bold).

Model	DTD Cimpoi et al. (2014)	UCSD Birds Welinder et al. (2010)	Stanford Dogs Khosla et al. (2011)	Flowers Nilsback & Zisserman (2008)	Pets Parkhi et al. (2012)	Indoor67 Quattoni & Torralba (2009)	FGVC-Aircraft Maji et al. (2013)	Cars Krause et al. (2013)	ImageNet Deng et al. (2009)
ResNet-18 He et al. (2016)									
One-hot	0.4153	0.3875	0.4015	0.4863	0.4090	0.4866	0.8703	0.3575	0.6350
One-hot + hue	**0.5230**	**0.5313**	**0.5754**	**0.6487**	**0.5919**	**0.6224**	**0.8920**	**0.8011**	**0.6621**
EfficientNet-B0 Tan & Le (2019)									
One-hot	0.5310	0.4907	0.5655	0.7252	0.6689	0.5458	0.8685	0.7882	0.7144
One-hot + hue	**0.5368**	**0.5569**	**0.5764**	**0.7438**	**0.6797**	**0.5477**	**0.8768**	**0.8079**	**0.7171**

6 DISCUSSION

Our paper proposes a novel angular parameter entitled activation hue in order to characterize and regularize deep CNN activation space. The activation hue represents a high-dimensional generalization of the standard hue angle, which is closely linked to human color perception in RGB intensity space, and thus represents an intuitively appealing mechanism for modeling activation space for general classification tasks.

We first motivate the activation hue through a number of preliminary observations in the context of kNN activation vector retrieval, which provide a number of novel insights regarding the structure of information in deep activation layers. Notably, activation vectors tend to be highly informative regarding pixel location in the image plane in general, and class-informative vectors tend to cluster according to an angle about the (x, y) image center. These observations motivate the hypothesis of a class-informative hue-like angle, defined both in 2D image space and in multi-channel activation space.

We then describe a novel activation hue loss function that makes use of angular θ_c class label information, and thereby complements standard loss functions such as cross-entropy from one-hot labels that provide no explicit information regarding the spatial distribution of activations in image space. In experiments training from scratch, combined one-hot + activation hue loss improves classification modestly but consistently in comparison to standard one-hot loss alone on a diverse variety of classification tasks, including Imagenet.

The mechanism behind the activation hue may be understood by considering an activation vector, including an RGB intensity vector, as representing a measurement distribution or spectrum. A uniform activation spectrum is generally uninformative regarding class and lies along an uninformative medial axis in activation space, similarly to how a pixel lying along the greyscale axis in RGB space is uninformative regarding color. An angular activation hue parameter in the plane perpendicular to the medial axis may thus be used to characterize bias towards a non-uniform, class-informative activation distribution, similarly to how standard hue characterizes a dominant color in RGB space.

Several practical aspects of our work are worth noting. Experiments made use of multiple widely-known, generic neural network architectures, diverse datasets with a wide range of task granularity including ImageNet, and basic NN classification methods Cover & Hart (1967), in order to demonstrate our hue-based model in the broadest possible context. We note that this angular loss behaviour is not limited to object-centric datasets like ImageNet or animal pictures. Our observations were confirmed on non-centric objects datasets, including the Describable Textures Dataset (DTD) Cimpoi et al. (2014).

Further investigation into the training scheme would be interesting, such as label modification during training for better convergence instead of forced assigned labels. The pixel vector matching method was more effective than other widely used bottleneck encodings, including pooling and flattened representations. To our knowledge, this is a novel result that may be useful if the computational requirement may be optimized for applications with tight memory or timing constraints. Finally, we believe activation hue may prove insightful to researchers investigating optics, information propagation, and next-generation computer vision systems, based on deep CNN models and activations.

REFERENCES

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition, 2016.

Mor Avi-Aharon, Assaf Arbelle, and Tammy Riklin Raviv. Hue-net: Intensity-based image-to-image translation with differentiable histogram loss functions. *arXiv preprint arXiv:1912.06044*, 2019.

Hossein Azizpour, Ali Sharif Razavian, Josephine Sullivan, Atsuto Maki, and Stefan Carlsson. Factors of transferability for a generic convnet representation, 2015.

Maxim Berman, Hervé Jégou, Andrea Vedaldi, Iasonas Kokkinos, and Matthijs Douze. Multigrain: a unified image embedding for classes and instances. *arXiv preprint arXiv:1902.05509*, 2019.

E. Bernhardsson. Annoy on github. `https://github.com/spotify/annoy`, 2015.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision*, pp. 213–229. Springer, 2020.

Wei-Yu Chen, Yen-Cheng Liu, Zsolt Kira, Yu-Chiang Frank Wang, and Jia-Bin Huang. A closer look at few-shot classification. *arXiv preprint arXiv:1904.04232*, 2019.

M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, , and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2014.

Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, and Andrea Vedaldi. Deep filter banks for texture recognition, description, and segmentation. *International Journal of Computer Vision*, 118(1):65–94, 2016.

Thomas Cover and Peter Hart. Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1):21–27, 1967.

J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Mark D Fairchild. *Color appearance models*. John Wiley & Sons, 2013.

Li Fei-Fei, Rob Fergus, and Pietro Perona. One-shot learning of object categories. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(4):594–611, 2006.

Alban Flachot, Arash Akbarinia, Heiko H Schütt, Roland W Fleming, Felix A Wichmann, and Karl R Gegenfurtner. Deep neural models for color classification and color constancy. *Journal of Vision*, 22(4):17–17, 2022.

Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720.

Albert Gordo, Jon Almazán, Jerome Revaud, and Diane Larlus. Deep image retrieval: Learning global representations for image search. In *European conference on computer vision*, pp. 241–257. Springer, 2016.

Yunhui Guo, Noel C Codella, Leonid Karlinsky, James V Codella, John R Smith, Kate Saenko, Tajana Rosing, and Rogerio Feris. A broader study of cross-domain few-shot learning. In *European Conference on Computer Vision*, pp. 124–141. Springer, 2020.

Daniel Haase and Manuel Amthor. Rethinking depthwise separable convolutions: How intra-kernel correlations lead to improved mobilenets. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14600–14609, 2020.

Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on visual transformer. *arXiv preprint arXiv:2012.12556*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Mikael Henaff, Joan Bruna, and Yann LeCun. Deep convolutional networks on graph-structured data. *arXiv preprint arXiv:1506.05163*, 2015.

Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7132–7141, 2018.

Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. Densely connected convolutional networks. *CoRR*, abs/1608.06993, 2016. URL http://arxiv.org/abs/1608.06993.

Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 603–612, 2019.

Wei Jiang, Eduard Trulls, Jan Hosang, Andrea Tagliasacchi, and Kwang Moo Yi. Cotr: Correspondence transformer for matching across images. *arXiv preprint arXiv:2103.14167*, 2021.

Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, June 2011.

Bobak Kiani, Randall Balestriero, Yann Lecun, and Seth Lloyd. projunn: efficient method for training deep networks with unitary matrices. *arXiv preprint arXiv:2203.05483*, 2022.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014.

Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better imagenet models transfer better?, 2019.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13)*, Sydney, Australia, 2013.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 991–999, 2015.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with restarts. *CoRR*, abs/1608.03983, 2016. URL http://arxiv.org/abs/1608.03983.

Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew B. Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *CoRR*, abs/1306.5151, 2013. URL http://arxiv.org/abs/1306.5151.

Pascal Mettes, Elise van der Pol, and Cees Snoek. Hyperspherical prototype networks. *Advances in neural information processing systems*, 32, 2019.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, pp. 807–814, Madison, WI, USA, 2010. Omnipress. ISBN 9781605589077.

Van Nhan Nguyen, Sigurd Løkse, Kristoffer Wickstrøm, Michael Kampffmeyer, Davide Roverso, and Robert Jenssen. Sen: A novel feature normalization dissimilarity measure for prototypical few-shot learning networks. In *European Conference on Computer Vision*, pp. 118–134. Springer, 2020.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

Hyeonwoo Noh, Andre Araujo, Jack Sim, Tobias Weyand, and Bohyung Han. Large-scale image retrieval with attentive deep local features. In *Proceedings of the IEEE international conference on computer vision*, pp. 3456–3465, 2017.

Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Ariadna Quattoni and Antonio Torralba. Recognizing indoor scenes. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 413–420, 2009. doi: 10.1109/CVPR.2009.5206537.

Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE transactions on pattern analysis and machine intelligence*, 41(7):1655–1668, 2018.

Ali S Razavian, Josephine Sullivan, Stefan Carlsson, and Atsuto Maki. Visual instance retrieval with deep convolutional networks. *ITE Transactions on Media Technology and Applications*, 4 (3):251–258, 2016.

Pau Rodríguez, Miguel A Bautista, Jordi Gonzàlez, and Sergio Escalera. Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing*, 75:21–31, 2018.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 806–813, 2014.

Jiayi Shen, Zehao Xiao, Xiantong Zhen, and Lei Zhang. Spherical zero-shot learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):634–645, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30, 2017.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *arXiv preprint arXiv:2104.00680*, 2021.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. URL http://arxiv.org/abs/1905.11946.

Yehui Tang, Kai Han, Jianyuan Guo, Chang Xu, Yanxi Li, Chao Xu, and Yunhe Wang. An image patch is a wave: Phase-aware vision mlp. *arXiv preprint arXiv:2111.12294*, 2021.

Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015.

David C. Van Essen, Stephen M. Smith, Deanna M. Barch, Timothy E.J. Behrens, Essa Yacoub, and Kamil Ugurbil. The wu-minn human connectome project: An overview. *NeuroImage*, 80:62–79, 2013. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.05.041. Mapping the Connectome.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Jian Wang, Feng Zhou, Shilei Wen, Xiao Liu, and Yuanqing Lin. Deep metric learning with angular loss. In *Proceedings of the IEEE international conference on computer vision*, pp. 2593–2601, 2017.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7794–7803, 2018.

Yan Wang, Wei-Lun Chao, Kilian Q. Weinberger, and Laurens van der Maaten. Simpleshot: Revisiting nearest-neighbor classification for few-shot learning. *CoRR*, abs/1911.04623, 2019. URL http://arxiv.org/abs/1911.04623.

Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-ucsd birds 200. Technical Report CNS-TR-201, Caltech, 2010. URL /se3/wp-content/uploads/2014/09/WelinderEtal10_CUB-200.pdf,http://www.vision.caltech.edu/visipedia/CUB-200.html.

Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*, pp. 499–515. Springer, 2016.

Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 3–19, 2018.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.

Zhifei Zhang, Yang Song, and Hairong Qi. Age progression/regression by conditional adversarial autoencoder. *CoRR*, abs/1702.08423, 2017. URL http://arxiv.org/abs/1702.08423.

Yutong Zheng, Dipan K Pal, and Marios Savvides. Ring loss: Convex feature normalization for face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5089–5097, 2018.