DISTRIBUTED CONVOLUTIONAL NEURAL NETWORKS: LEARNING DISENTANGLED FEATURE EMBEDDINGS

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper proposes a loss function for training the distributed convolutional neural network (DisCNN), which embeds positive-class features into a compact ball and drives the embeddings of dissimilar negative-class features toward the origin. We theoretically validate that the convolutional layers of DisCNN exclusively extract positive features while completely ignoring any dissimilar features, thereby fully disentangling positive features from dissimilar features. Furthermore,We verify a positive monotonic association between the presence probability of positive features and the ℓ_2-norms of feature embeddings. The model achieves favorable generalization performance on test data, even on training-unseen samples bearing partial positive features. Moreover, a preliminary experiment demonstrates DisCNN's potential for object detection.

1 INTRODUCTION

Training the convolutional neural network (CNN) with cross-entropy loss has long been the dominant paradigm for image classification LeCun et al. (2015; 1999; 2002; 2010). Its underlying principle is that the stacked convolutional layers of the CNN first map input images with diverse visual features into hierarchical two-dimensional feature maps, which capture representations from simple low-level features to abstract high-level ones across layers. Afterwards, the abstract feature maps output by the final convolutional layer are flattened and projected via fully-connected (FC) layers into multiple sets of compact embeddings corresponding to various object classes. Nevertheless, convolutional layers collectively yield abstract feature maps for all categories, making it infeasible to explicitly and uniquely assign individual feature maps to specific object classes. Existing feed-forward visualization-based analytical methods can only perform post-hoc inspection of feature maps after model training Yosinski et al. (2015); Zeiler & Fergus (2014); Simonyan et al. (2013), and cannot fundamentally resolve feature entanglement.

Meanwhile, cognitive neuroscience studies indicate that object recognition in the human brain originates from two visual pathways extending from the retina to the visual cortex: the ventral pathway and the dorsal pathway Gazzaniga (2009). The ventral pathway is responsible for object recognition, whereas the dorsal pathway processes motion and spatial information. To achieve robust visual perception, the ventral pathway exhibits a hierarchical organization: the primary visual cortex extracts elementary visual features, while higher-level visual cortices assemble these low-level cues into abstract object representations to facilitate object recognition and category discrimination. This hierarchical feature learning mechanism bears close resemblance to the working principle of CNNs. Beyond this hierarchical processing, higher-level visual cortices within the ventral pathway exhibit a distributed organization, whereby specialized cortical regions process information for different object categories. For instance, one region specializes in face processing, a second region responds to tool-related stimuli, and a third region handles scene information. These distributed cortical regions are activated exclusively by features belonging to their corresponding target objects. Inspired by this biological observation, we propose DisCNNs, a family of headless CNNs tailored for single-category feature learning.

Training such networks in a distributed manner to extract visual features solely for a target category carries substantial research significance. In this work, we first analyze the geometric property of fea-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ture embedding immediately preceding the classification head in conventional CNNs. By analogy, we derive an equivalent loss for cross-entropy loss on headless networks without the Softmax layer. Building upon this observation, we further propose a novel loss function, termed N2O loss, which constrains visual-feature embeddings of the target positive class within a compact ball, and simultaneously projects embeddings of dissimilar negative samples toward the origin. For categories with highly distinct characteristics, such as tools, animals, human faces, and scenes, multiple such headless networks can be trained independently. We term these distributed networks DisCNNs. Through two complementary experiments, we validate that the convolutional layers of DisCNN optimized with N2O loss selectively extract features exclusively from the target positive class and fully ignore dissimilar features not only from negative references but also from unseen categories during training. Enabled by this distributed architecture inspired by the ventral visual pathway, each DisCNN isolates features of its target positive class and achieves effective feature disentanglement from features of dissimilar categories. Since each DisCNN focuses exclusively on positive-feature learning, the number of feature maps produced by the convolutional layers can be drastically reduced, even down to a single channel, yielding an extremely lightweight network architecture. Furthermore, this distributed and lightweight architecture enables all category-specific DisCNN instances to conduct feed-forward inference in parallel for classification tasks. Theoretically, this yields an N-fold speedup relative to conventional CNNs, where N denotes the number of target categories. We verify that the ℓ_2-norms of positive-class feature embeddings can serve as a metric indicating the presence of positive-class features. The proposed model achieves strong generalization on the corresponding test dataset. When evaluated on unseen categories, samples with features dissimilar to the positive class are mapped toward zero embeddings, while samples sharing consistent positive-class features are compactly clustered within the same embedding region as positive samples. Finally, we conduct a preliminary experiment to verify that addressing object-detection tasks based on DisCNNs is straightforward, as the problem essentially simplifies to object localization. Collectively, our work demonstrates that the biologically inspired distributed paradigm provides a promising alternative for single-category feature learning, offering lightweight inference, parallel computation, and strong generalization toward both seen and unseen categories.

2 Model, dataset, and loss functions

2.1 Model architecture

DisCNN comprises four convolutional layers with 3×3 kernels, following the VGG architecture proposed in Simonyan & Zisserman (2014). Unlike conventional multi-class CNNs that learn universal feature representations for all categories, our model exclusively learns discriminative abstract features for a single positive target class, thus requiring significantly fewer feature channels. For instance, core visual attributes of typical categories such as car bodies, wheels, windows, license plates, and vehicle lights can be effectively encapsulated within only a handful of feature channels. Accordingly, we configure the convolutional layers of DisCNN to output merely eight feature maps, constructing a lightweight variant termed DisCNN-8. Furthermore, the number of output channels can be further reduced to one, where the entire feature information of the target category is compressed into a single high-level abstract representation; we denote this ultra-compact variant as DisCNN-1. Except for the removed softmax layer, the fully connected (FC) layers of DisCNN adopt the standard design of conventional CNNs. Overall, the network architecture contains seven hidden layers, including four convolutional layers and three FC layers. The detailed layer configurations are presented in Table 1. For intuitive comparison, the last column of the table reports the standard VGG model trained under a 10-way classification setting, which yields 128 convolutional feature maps, equivalent to approximately 12 channels per category on average. As expected, DisCNNs contain substantially fewer parameters than the VGG model trained for the 10-way classification task. As listed in Table 2, DisCNN-8 has merely 0.149 M parameters, achieving a drastic parameter reduction compared with the VGG baseline, which strongly demonstrates the superior lightweight property of our proposed DisCNN architecture.

2.2 Dataset

We train our model on the STL-10 dataset, introduced in Coates et al. (2011). STL-10 consists of ten object categories: airplane, bird, car, cat, deer, dog, horse, monkey, ship, and truck. Each

Table 1: Model Architectures. Conv layers are formatted as "conv⟨kernel size⟩-⟨channels⟩".

DisCNN-8	DisCNN-1	VGG
input(96×96 RGB)	input(96×96 RGB)	input(96×96 RGB)
Conv3-64, BN, ReLU, MaxPool	Conv3-64, BN, ReLU, MaxPool	Conv3-64, BN, ReLU, MaxPool
Conv3-32, BN, ReLU, MaxPool	Conv3-32, BN, ReLU, MaxPool	Conv3-128, BN, ReLU, MaxPool
Conv3-16, BN, ReLU, MaxPool	Conv3-16, BN, ReLU, MaxPool	Conv3-128, BN, ReLU, MaxPool
Conv3-8, BN, ReLU, MaxPool	Conv3-1, BN, ReLU, MaxPool	Conv3-128, BN, ReLU, MaxPool
FC-288, ReLU	FC-36, ReLU	FC-4096, ReLU
FC-128, ReLU	FC-32, ReLU	FC-2048, ReLU
FC-16, ReLU	FC-16, ReLU	FC-1000, Softmax

Table 2: Number of parameters (in millions).

Network	DisCNN-8	DisCNN-1	VGG
Number of parameters	0.149	0.028	3.096

category has 500 training images and 800 test images. All images, cropped from ImageNet Deng et al. (2009), have a fixed resolution of 96×96 pixels. Obviously, the ten categories can be grouped into two superclasses: vehicles and mammals. Samples from the same superclass exhibit high feature similarity, whereas objects across different superclasses share few overlapping visual features. Leveraging this property, we select car (a vehicle category) as the target positive class and two mammal classes (bird and cat) as negative references. We then train DisCNN on this customized subset to learn feature embeddings exclusively for cars.

Remark 1. *The training dataset for DisCNN must satisfy two conditions: positive samples possess shared intrinsic features, and samples from negative classes contain no overlapping visual features with the target positive class.*

2.3 Loss function

Lemma 1. *The zero vector is orthogonal to every vector $\mathbf{v} \in \mathbb{R}^d$, since the inner product satisfies $\langle \mathbf{0}, \mathbf{v} \rangle = 0$ for any $\mathbf{v}$ Strang (2016).*

This lemma is a well-known result, and we omit its proof.

Lemma 2. *A binary softmax layer is appended to the head of DisCNN-8 in Table 1. Training is carried out on a binary-class dataset with cross-entropy loss, after which the softmax layer is discarded. Let $\boldsymbol{f}_{c_1}$ and $\boldsymbol{f}_{c_2}$ denote the model embeddings corresponding to the two classes. It follows that*

(a) $\boldsymbol{f}_{c_1}$ and $\boldsymbol{f}_{c_2}$ have similar average ℓ_2-norms.
(b) $\boldsymbol{f}_{c_1}$ and $\boldsymbol{f}_{c_2}$ are aligned among themselves, respectively.
(c) $\boldsymbol{f}_{c_1}$ and $\boldsymbol{f}_{c_2}$ are orthogonal.

Proofs are deferred to the supplementary material.

Definition 1. *Let $\boldsymbol{f}_c \in \mathbb{R}^{N_c \times d}$ be the matrix of batch-wise feature embeddings for class c, where each row corresponds to a single embedding vector. The mean and standard deviation of the ℓ_2-norms of embeddings for class c are defined as*

$$\mu_c = \frac{1}{N_c} \sum_{i=1}^{N_c} \|\boldsymbol{f}_{c,i}\|_2, \quad \sigma_c = \sqrt{\frac{1}{N_c} \sum_{i=1}^{N_c} \left(\|\boldsymbol{f}_{c,i}\|_2 - \mu_c\right)^2}. \tag{1}$$

Given two embedding matrices $\boldsymbol{f}_p \in \mathbb{R}^{N_p \times d}$ and $\boldsymbol{f}_n \in \mathbb{R}^{N_n \times d}$, the average cosine similarity over index-aligned pairs $(\boldsymbol{f}_{p,i}, \boldsymbol{f}_{n,i})$ is

$$\mathrm{Cos}(\boldsymbol{f}_p, \boldsymbol{f}_n) = \frac{1}{N} \sum_{i=1}^{N} \frac{\boldsymbol{f}_{p,i} \boldsymbol{f}_{n,i}^\top}{\|\boldsymbol{f}_{p,i}\|_2 \|\boldsymbol{f}_{n,i}\|_2}, \tag{2}$$

where $N = \min(N_p, N_n)$.

Lemma 3. *Based on Lemma 2, we define the following loss function for DisCNN-8 on the binary classification task, which yields optimization dynamics equivalent to cross-entropy training. The loss reads*

$$\mathcal{L} = \frac{1}{\mu_{c_1}} + \sigma_{c_1} + \frac{1}{\mu_{c_2}} + \sigma_{c_2} + \lambda \mathrm{Cos}(\boldsymbol{f}_{c_1}, \boldsymbol{f}_{c_2}), \tag{3}$$

with empirically chosen hyperparameter $\lambda = 2$.

Proofs are deferred to the supplementary material.

Definition 2. *Our objective is to enforce the model to activate exclusively for positive-class features and drive contrasting negative embeddings toward the origin. The proposed N2O loss is derived by slightly modifying equation 3 as*

$$\mathcal{L}_{\mathrm{N2O}} = \frac{1}{\mu_p} + \sigma_p + \mu_n + \sigma_n + \mathrm{Cos}(\boldsymbol{f}_p, \boldsymbol{f}_n) \tag{4}$$

where subscripts p and n denote the positive and negative classes, respectively.

Remark 2. *The loss function in equation 4 maximizes the ℓ_2-norms of positive-class feature embeddings under a low-standard-deviation constraint, while driving negative-class embeddings toward the origin. Orthogonality between positive and negative embeddings is maintained throughout optimization. By Lemma 1, this orthogonality property still holds even as negative-class embeddings converge to zero vectors.*

Furthermore, we conduct ablation experiments on a variant of $\mathcal{L}_{\mathrm{N2O}}$ loss that retains only the first three terms in Equation (4); see the supplementary material.

3 Main results

3.1 DisCNN extracts exclusively positive-class features

Theorem 1. *Let $\{\boldsymbol{x}_i\}_{i=1}^N$ denote a set of input images sampled from the data distribution $\mathcal{D}$. For positive-class samples, DisCNN yields feature embeddings $\boldsymbol{f}_p \in \mathbb{R}^d$ that lie within a compact ball of radius r. By contrast, the feature embeddings $\boldsymbol{f}_n \in \mathbb{R}^d$ of negative-class samples shrink toward the origin.*

Proof: We train DisCNN-8 with back-propagation using our proposed N2O loss, where $\{\text{car}\}$ serves as the positive class and $\{\text{bird}, \text{cat}\}$ constitute the negative classes. We feed the training and test datasets into DisCNN-8 respectively and collect the statistics of embedding ℓ_2-norms, summarized in Table 3.

Table 3: Statistics of ℓ_2-norms. $\boldsymbol{c}_p$ denotes the centroid of positive embeddings.

Dataset	Metric	Count	Mean	Std	Max	Min	Median
Training	$\|\boldsymbol{f}_p\|_2$	500	12.4380	0.1147	12.7368	12.0582	12.4272
	$\|\boldsymbol{c}_p\|_2$	—	12.4327	—	—	—	—
	$\|\boldsymbol{f}_n\|_2$	1000	0.0000	0.0000	0.0000	0.0000	0.0000
Test	$\|\boldsymbol{f}_p\|_2$	800	11.2299	2.9100	14.3649	0.0000	12.2005
	$\|\boldsymbol{c}_p\|_2$	—	11.2237	—	—	—	
	$\|\boldsymbol{f}_n\|_2$	1600	0.2340	1.1393	11.8479	0.0000	0.0000

As shown in Table 3, on the training set, positive-class feature embeddings are mapped far from the origin , while negative embeddings collapse to the origin. This behavior retains good generalization on the test set.

In addition, we evaluate the cosine similarity and ℓ_2-distance between each positive-class embedding and its centroid, and the corresponding statistics are summarized in Table 4.

As can be observed from Table 4, on both the training and test sets, positive-class embeddings concentrate within a compact ball of small radius r centered at their centroid, with near-aligned directions emanating from the origin, though the radius is larger on the test set.

Table 4: Statistics of cosine similarities and ℓ_2-distances with the centroid.

Dataset	Metric	Count	Mean	Std	Median	Max / Min
Training	Cosine	500	0.9996	0.0006	0.9997	1.0000 / 0.9877
	ℓ_2-distance	500	0.3539	0.1336	0.3383	1.9521 / 0.1102
Test	Cosine	800	0.9774	0.1419	0.9997	1.0000 / 0.0000
	ℓ_2-distance	800	1.6829	2.6898	0.6414	12.4327 / 0.1215

Combining the above two lines of evidence, we conclude that the theorem holds.

□

Although Theorem 1 guarantees that the feature embeddings of negative-class samples contract toward the origin, it remains unclear whether this contraction is achieved by completely ignoring negative-sample features or by extracting negative-class features and then mapping them to the origin. Therefore we propose Theorem 2 to prove that the former is indeed the case.

Theorem 2. *Feature maps $\{\boldsymbol{F}_i\}_{i=1}^{M} \in \mathbb{R}^{m\times m}$ generated by the convolutional layers of DisCNN encode features for the positive class and negative classes.*

Proof: First, we construct a subset of the Tiny-ImageNet-200 dataset containing 180 categories by removing classes visually similar to cars, such as wagons, bullet trains, police vans, and tractors. From this filtered subset, we randomly sample six images per class and combine them with car samples from the STL-10 dataset to assemble our training set. The test set is built following the identical protocol. We then conduct two experiments on this newly curated dataset to validate our theorem.

Experiment A. We train DisCNN using the N2O loss, where the car category is defined as the positive class and all remaining samples are treated as negatives. Then we evaluate the model on the test set and statistically compare the ℓ_2-norms and cosine similarities of embeddings from positive and negative samples, as summarized in Table 5.

Table 5: Statistics of ℓ_2-norms and cosine similarities (experiment A).

Metric	Count	Mean	Std	Median	Max / Min
$\|\boldsymbol{f}_p\|_2$	800	8.5817	1.7171	9.0455	10.4946 / 0.0000
$\|\boldsymbol{f}_n\|_2$	1080	0.0004	0.0146	0.0000	0.4812 / 0.0000
Cosine	800	0.9825	0.1229	0.9997	1.0000 / 0.0000

Evidently, the negative samples in the training set are heterogeneous and share no common features. Thus, it is impossible for the model to learn representative negative-class features. Nevertheless, the proposed N2O loss remains effective under this setting, as shown in Table 5. This validates our theorem: the convolutional layers only extract positive-class features.

Experiment B. In this experiment, we test DisCNN-8, which is trained with car as the positive class and bird andcat as negative classes, on the same test set used in Experiment A. The test set of Experiment A contains numerous samples that share no feature similarities with birds or cats, including mushrooms, umbrellas, and electric fans. If the model only learns negative representations from bird and cat features and maps these learned features toward the origin, such training-unseen negative samples should fail to collapse correspondingly. As reported in Table 6, these unseen negative samples still contract effectively toward the origin, which further validates our theorem.

□

Remark 3. *Based on Theorem 2, DisCNN feature embeddings achieve explicit feature disentanglement compared with embeddings obtained via cross-entropy. Specifically, positive-class features are fully disentangled from all dissimilar features, and these dissimilar features yield zero model activation.*

Table 6: Statistics of ℓ_2-norms and cosine similarities (Experiment B).

Metric	Count	Mean	Std	Median	Max / Min
$\|\boldsymbol{f}_p\|_2$	800	10.4777	3.0763	11.5801	13.8118 / 0.0000
$\|\boldsymbol{f}_n\|_2$	1080	0.2359	1.1591	0.0000	11.7885 / 0.0000
Cosine	800	0.9625	0.1854	0.9996	1.0000 / 0.0000

3.2 THE FEATURE EMBEDDINGS ℓ_2-NORMS SERVE AS A METRIC INDICATING THE PRESENCE OF POSITIVE-CLASS FEATURES

Theorem 3. *Under Theorem 2, the ℓ_2-norms of DisCNN feature embeddings can serve as a metric indicating the presence probability of positive features, which increases monotonically with the activation strength $\|\boldsymbol{f}_p\|_2$.*

Proof: Firstly, we evaluate the trained DisCNN-8 on a subset of the STL-10 test set consisting of {car, bird, cat}. We bin the ℓ_2-norms of embeddings f_p for all test samples as illustrated in Figure 1. For each bin, we visualize car samples sorted in ascending order of their ℓ_2-norms in Figure 2. As observed from Figure 2, samples within the bin of maximum ℓ_2-norm mostly possess complete car-relevant features. In contrast, samples with smaller ℓ_2-norms either contain only partial car-like features or depict cars with atypical appearances.

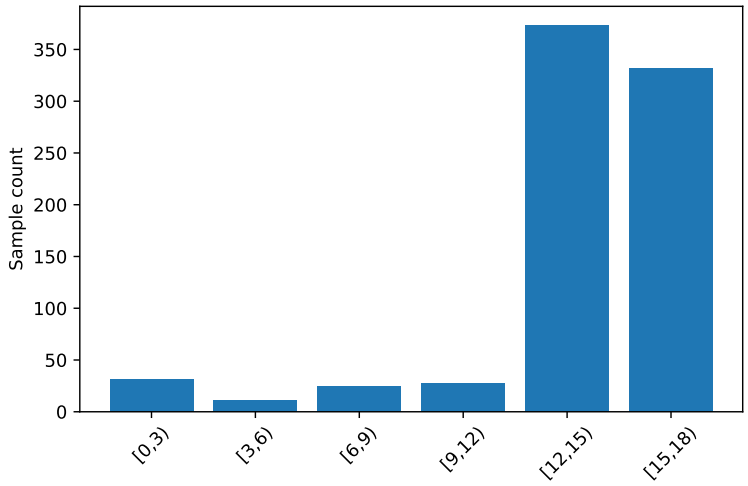

Figure 1: Histogram of car sample counts in each embedding ℓ_2-norm bin.

Secondly, we evaluate the trained DisCNN-8 on a subset of the STL-10 test set consisting of {car, truck, ship, bird, cat, dog, horse}. Specifically, truck and ship serves as training-unseen quasi-positive classes visually similar to car, while dog and horse are adopted as negative classes to balance the dataset distribution. We compute statistical metrics for the ℓ_2-norms of embeddings of car, truck, and ship, as summarized in Table 7.

Table 7: Statistics of ℓ_2-norms for Car, Truck and Ship.

Class	Count	Mean	Std	Max	Min	Median
Car	800	13.3985	3.2835	17.4446	0.0000	14.4058
Truck	800	11.6809	4.4195	16.6284	0.0000	13.7389
Ship	800	8.5037	5.5875	16.0014	0.0000	10.5730

Accordding to Table 7, we first observe that the training-unseen quasi-positive classes truck and ship activate the model at a considerable rate. This arises because truck and ship share many similar visual features with car, such as windows, wheels, lights, license plates. Further inspection reveals

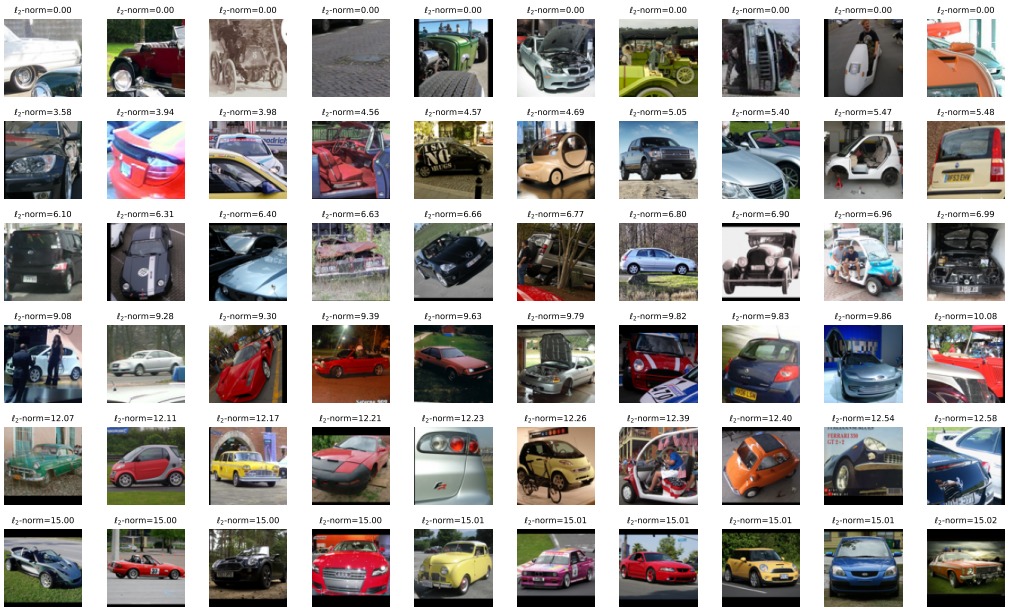

Figure 2: Visualization of car samples grouped by embedding ℓ_2-norm bins. Each row shows 10 samples from the same bin.

that the ℓ_2-norms of car embeddings are on average larger than those of truck embeddings. This discrepancy becomes more pronounced with ship. This also indicates that features of truck and ship samples dissimilar to car-class features are silenced, just like those of other negative samples. Such suppressed features contribute nothing to the ℓ_2-norms of their embeddings.

From two complementary perspectives, these two experiments validate the positive monotonic relationship between the presence probability of positive-class features and the ℓ_2-norm of feature embeddings.

□

3.3 MODEL EVALUATION

The STL-10 test sub-dataset {car, bird, cat} is used to evaluate the DisCNN-8.

Remark 4. *When the test dataset distribution differs from that of the training set, this causes two phenomena: some negative-class embeddings deviate from the origin and move toward positive-class embeddings, while certain positive samples with atypical features may also be mapped near the origin. However, based on Theorem 3, the model retains its effectiveness by comparing the embedding ℓ_2-norm of a sample against an appropriate threshold.*

Thresholds can be chosen according to practical needs. If a lower false positive rate for the positive class is desired, a smaller threshold should be chosen. Conversely, if a lower false-negative rate is required for the negative class, a larger threshold should be selected. We vary the classification threshold and report the corresponding precision, recall and F1 scores in Table 8. The threshold of 5 achieves the best overall performance, yielding the highest F1 for both positive and negative classes.

Moreover, this model demonstrates excellent generalization to training-unseen classes. We test the model on the set {truck, deer, monkey}, treating truck as the quasi-positive class. The threshold of 3 yields reasonably good performance, as shown in Table 9. Unseen samples without any positive class features are mapped to the origin, whereas those with similar features are mapped in the same compact ball as the positive class.

In addition, we perform ablation experiments on an alternative positive class (cat); see the supplementary material.

Table 8: Classification performance under different thresholds.

Threshold	Positive			Negative		
	Precision	Recall	F1	Precision	Recall	F1
0	0.979	0.841	0.905	0.907	0.988	0.946
1	0.974	0.897	0.934	0.944	0.986	0.965
2	0.968	0.925	0.946	0.961	0.983	0.972
3	0.960	0.948	0.954	0.974	0.980	0.977
4	0.950	0.962	0.956	0.981	0.975	0.978
5	0.944	0.970	**0.957**	0.986	0.972	**0.979**
6	0.925	0.976	0.950	0.989	0.963	0.976

Table 9: Classification performance under different thresholds.

Threshold	Positive			Negative		
	Precision	Recall	F1	Precision	Recall	F1
0	0.960	0.840	0.896	0.909	0.978	0.942
1	0.953	0.896	0.924	0.945	0.975	0.960
2	0.944	0.923	0.933	0.961	0.972	0.966
3	0.936	0.942	**0.939**	0.971	0.968	**0.970**
4	0.920	0.961	0.940	0.981	0.961	0.971
5	0.900	0.965	0.931	0.984	0.952	0.967
6	0.879	0.972	0.923	0.988	0.942	0.964

3.4 Exploring the Potential of DisCNN for Object Detection

To detect cars embedded in a complex background, the large image is first evenly partitioned into patches of appropriate size, and then the trained DisCNN-8 is applied to determine whether a patch contains a car according to the embedding ℓ_2-norm of the model. The results are shown in Figure 3.

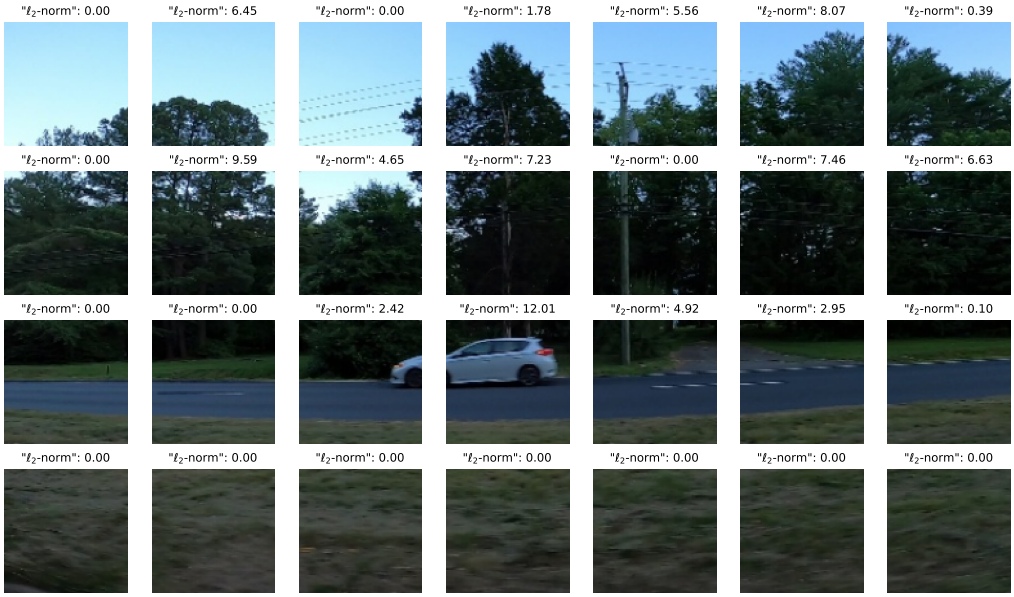

Figure 3: Patches and embedding ℓ_2-norms

Since the ratio of positive to negative samples here is 1:28, quite different from the training data, a considerable portion of the background patches activate the model, as shown in Figure 3. So, an

appropriate threshold shall be chosen according to Remark 4. By sorting the ℓ_2-norms in descending order, as shown in Figure 4, the threshold can be set to 10 to suppress all background patches and locate the car patch, i.e., the patch at the third row and fourth column.



Figure 4: Patches sorted by embedding ℓ_2-norm

Remark 5. *Background patches lacking features similar to those of the positive class (car) fail to activate the model strongly enough, much like negative samples. This property is valuable for object-detection tasks. Note that our experiment only yields coarse-grained localization results. To achieve precise object localization, one may further take the maximum-activation patch obtained in our experiment as a reference to generate candidate bounding boxes with certain offsets in both position and size relative to this reference, and re-evaluate them using the model to localize the patch with maximum activation.*

Remark 6. *DisCNN embeddings constitute pure feature representations for specific categories sharing common characteristics, which are disentangled not only from dissimilar features but also from spatial features such as size and position. Consequently, addressing object-detection tasks based on DisCNNs becomes straightforward, as the problem essentially simplifies to object localization.*

4 Conclusion

This paper presents DisCNN together with its training loss function N2O, establishing a distributed feature-embedding framework for image-classification tasks. The core property of a category-specific trained DisCNN is that its embeddings exclusively encode positive-class features, while features of dissimilar samples are completely suppressed. The proposed architecture aligns closely with the distributed object-recognition mechanism of the human ventral visual pathway. As a fundamental visual-embedding paradigm, DisCNN possesses broad application potential. First, parallel inference based on the distributed architecture brings substantial speed gains for image-classification tasks. Second, it can also enhance classification accuracy. In our ablation experiments, we further train a DisCNN targeting the dog category and achieve comparably high test accuracy. It can therefore be inferred that high overall classification accuracy can be maintained with N DisCNNs, provided that their target features are mutually dissimilar. Third, the feature-disentangling property of DisCNNs may greatly enhance the performance of YOLO-based object detection Redmon et al. (2016), as it simplifies the object-detection problem to object localization. More broadly, for world-model research Assran et al. (2023), since DisCNN embeddings encode only positive-class features, they are disentangled not only from dissimilar features of other categories but also from spatial features such as size and position. Such properties could bring considerable benefits to this line of research.

AI use statement

In this work, generative AI tools were used to assist in manuscript language polishing, academic sentence refinement, code debugging, and scientific figure plotting and beautification. No generative AI tools were used to produce research ideas, technical motivations, methodologies, mathematical derivations, experimental results, or empirical conclusions. All AI-assisted text, code, and visual contents have been fully reviewed, verified, and manually revised by the authors. The authors take full responsibility for the originality, accuracy, and validity of all content in the final manuscript.

Reproducibility statement

All experimental settings, network architectures, and hyperparameters are explicitly presented in the main paper and supplementary materials.

The source code, experimental scripts, and model configurations required for reproducing all results will be publicly released upon paper acceptance.

All datasets used in this work are publicly available standard benchmarks with official public access.

All quantitative results are reported with consistent experimental settings and fixed random seeds to ensure reproducibility. Detailed ablation studies, and the proofs of Lemma 2 and Lemma 3 are provided in the supplementary file.

ACKNOWLEDGMENTS

This section is removed for double-blind review.

REFERENCES

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 15619–15629, 2023.

Adam Coates, Andrew Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 215–223. JMLR Workshop and Conference Proceedings, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia 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. Ieee, 2009.

Michael S Gazzaniga. *The cognitive neurosciences*. MIT press, 2009.

Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. Object recognition with gradient-based learning. In *Shape, contour and grouping in computer vision*, pp. 319–345. Springer, 1999.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 2002.

Yann LeCun, Koray Kavukcuoglu, and Clément Farabet. Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems*, pp. 253–256. IEEE, 2010.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788, 2016.

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

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Gilbert Strang. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, 5th edition, 2016. ISBN 9780980232776.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

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