Interpretable Bag-of-Visual Words Networks for Adversarial Example Detection

Krishna Kanth Nakka and Mathieu Salzmann
CVLab, EPFL

Presented by Krzysztof Lis, CVLab, EPFL

Research Goal

Understanding how deep neural network makes predictions, as well as when and why they make errors

Detection of malicious samples in the case of adversarial attacks

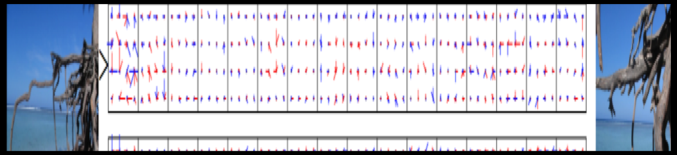
Structured Representations

A Text Retrieval Approach to Object Matching in Video

[Sivic et al., 2003]

Aggregating local descriptors into compact codes

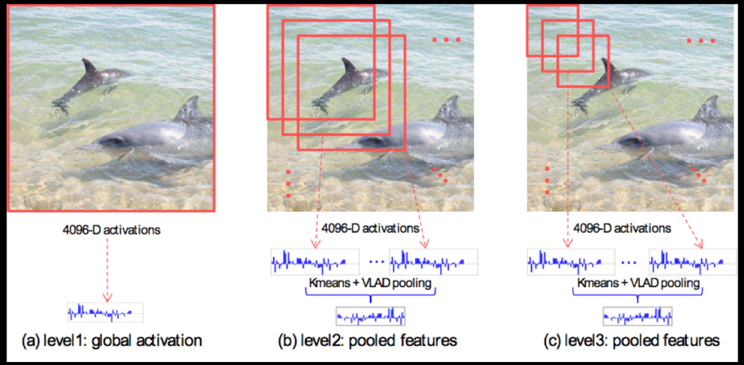
[Herve et al., 2010]



Deep Structured Networks

Multiscale Orderless Pooling

[Gong et al., 2014]



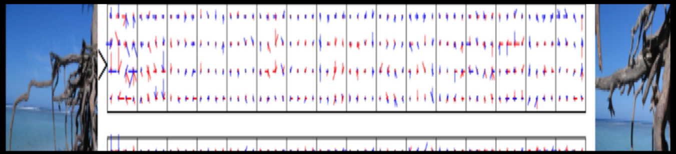
Structured Representations

A Text Retrieval Approach to Object Matching in Video

[Sivic et al., 2003]

Aggregating local descriptors into compact codes

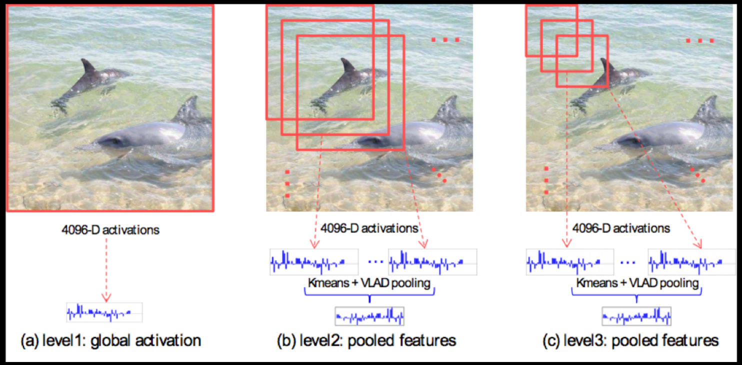
[Herve et al., 2010]



Deep Structured Networks

Multiscale Orderless Pooling

[Gong et al., 2014]



Structured Representations

Structured Representations

Stage 1

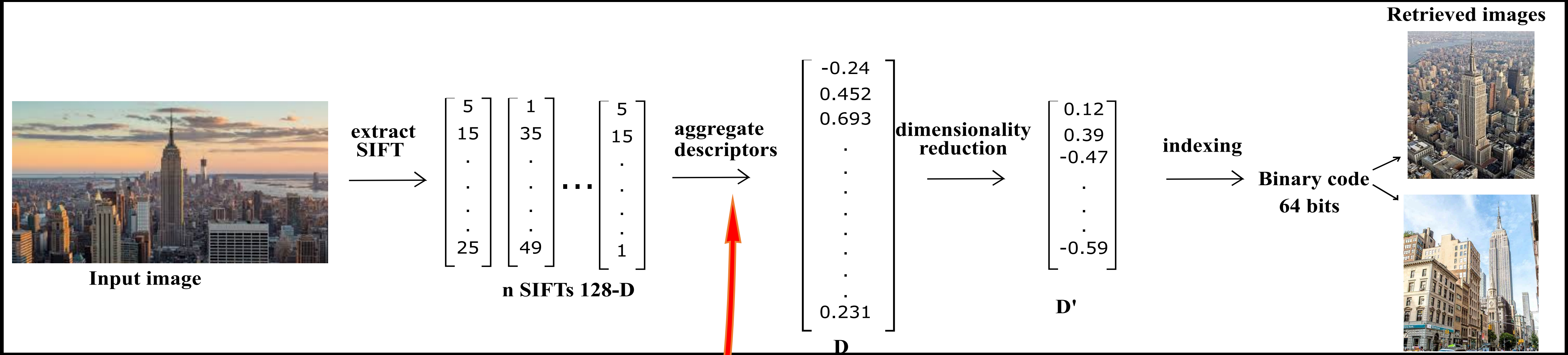
Local features

Stage 2

Aggregation

Stage 3

Retrieval



Large scale image retrieval framework

Different types of Aggregation: BoW, VLAD, Fisher

Herve et al., Aggregating local descriptors into compact codes, 2010

Classical Bag of Words Model

- Consider image $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}, \ldots\ldots, \mathbf{x}_n], \mathbf{x}_i \in \mathbf{R}^D$
- Learn codebook: $\mathbf{B}$, k-means clustering of descriptors

$$\mathbf{B} = [\mathbf{b}_1, \mathbf{b}_2, \ldots\ldots, \mathbf{b}_K], \mathbf{b}_i \in \mathbf{R}^D$$

- Histogram vector: $g(\mathbf{x}_i)$

$$\boxed{g(\mathbf{x}) = [\delta(\mathbf{b}_1 = NN(\mathbf{x})), \ldots\ldots, \delta(\mathbf{b}_N = NN(\mathbf{x}))]}$$

$$\delta(true) = 1 \text{ and } \delta(false) = 0.$$

- Feature aggregation: $G(\mathbf{X}) \in \mathbf{R}^D$

$$G(\mathbf{X}) = \sum_{n=1}^{N} g(\mathbf{x}_n).$$

Net Bag of Words layer in CNNs

Soft assignment policy of features to the codewords

$$\mathbf{h}(\mathbf{x}) = [a_0(\mathbf{x}), a_1(\mathbf{x}), \cdots, a_K(\mathbf{x})]^T$$

$$a_k(\mathbf{x}) = \frac{e^{-\alpha\|\mathbf{x}-\mathbf{b}_k\|^2}}{\sum_{k'} e^{-\alpha\|\mathbf{x}-\mathbf{b}_{k'}\|^2}}.$$

x is a D-dimensional feature vector extracted at final convolutional layer
B is learnable **Codebook** with K codewords each of D-dimension
h(x) – BoW representation

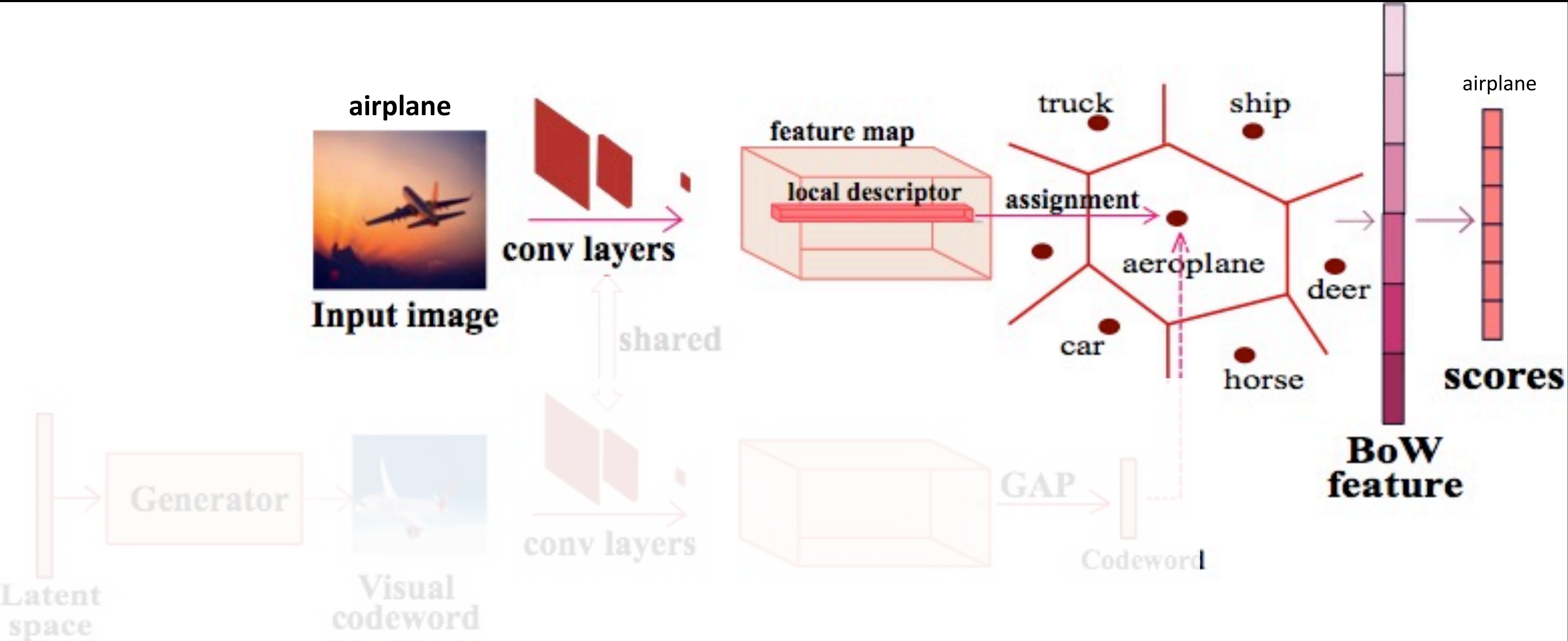
Interpretable Codebook

Interpret the decisions of a CNN by assigning a visual representation to the codewords

We propose to learn semantic codewords in an end-to end manner using pre-trained generative adversarial networks

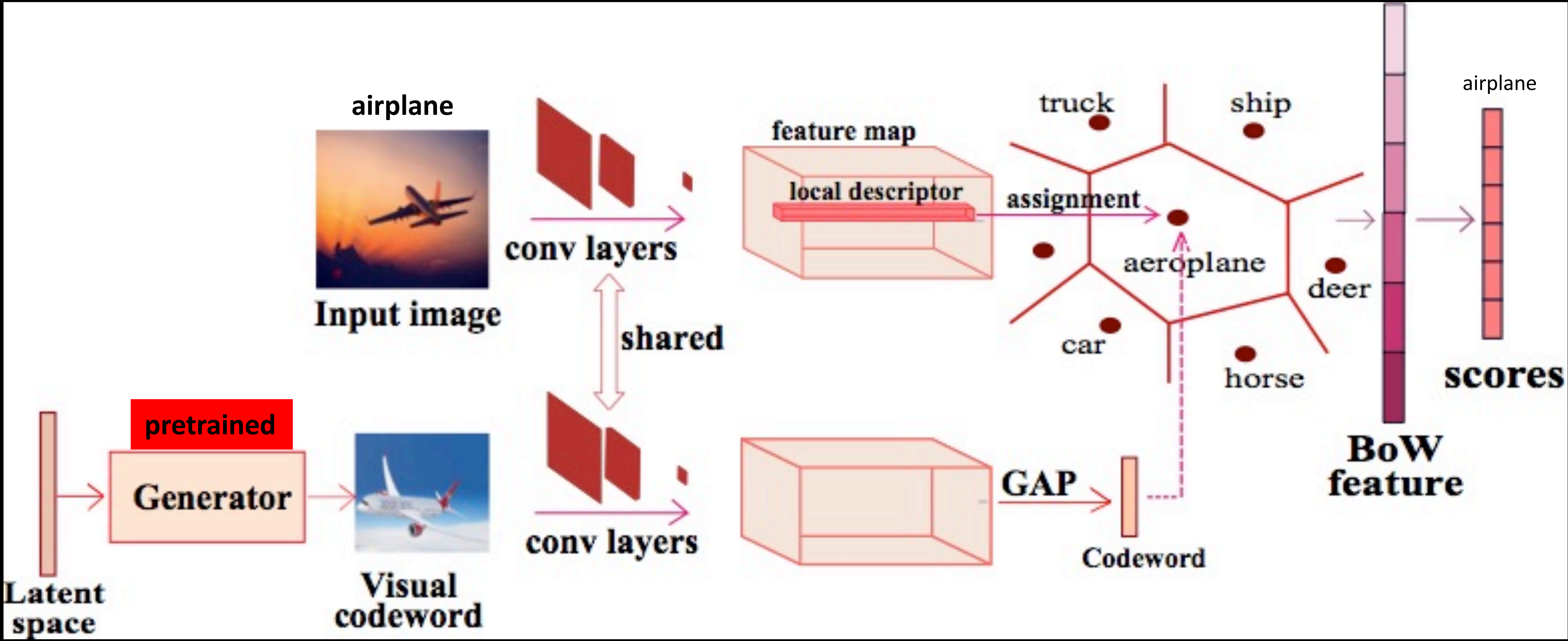
Interpretable Codebook

NetBoW model with BoW layer on the top of the network

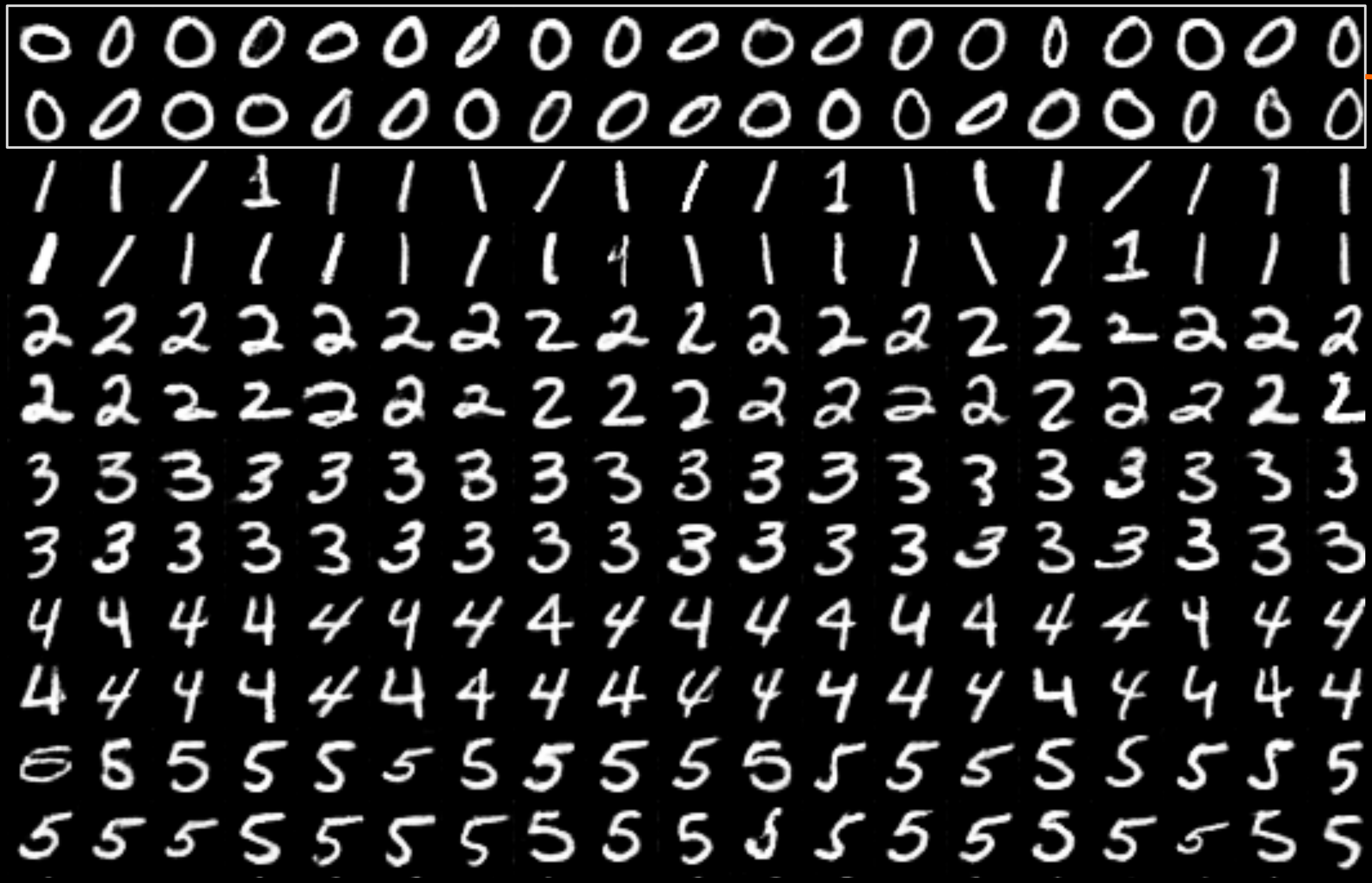


Interpretable Codebook

Jointly optimize the latent space of GAN and BoW network



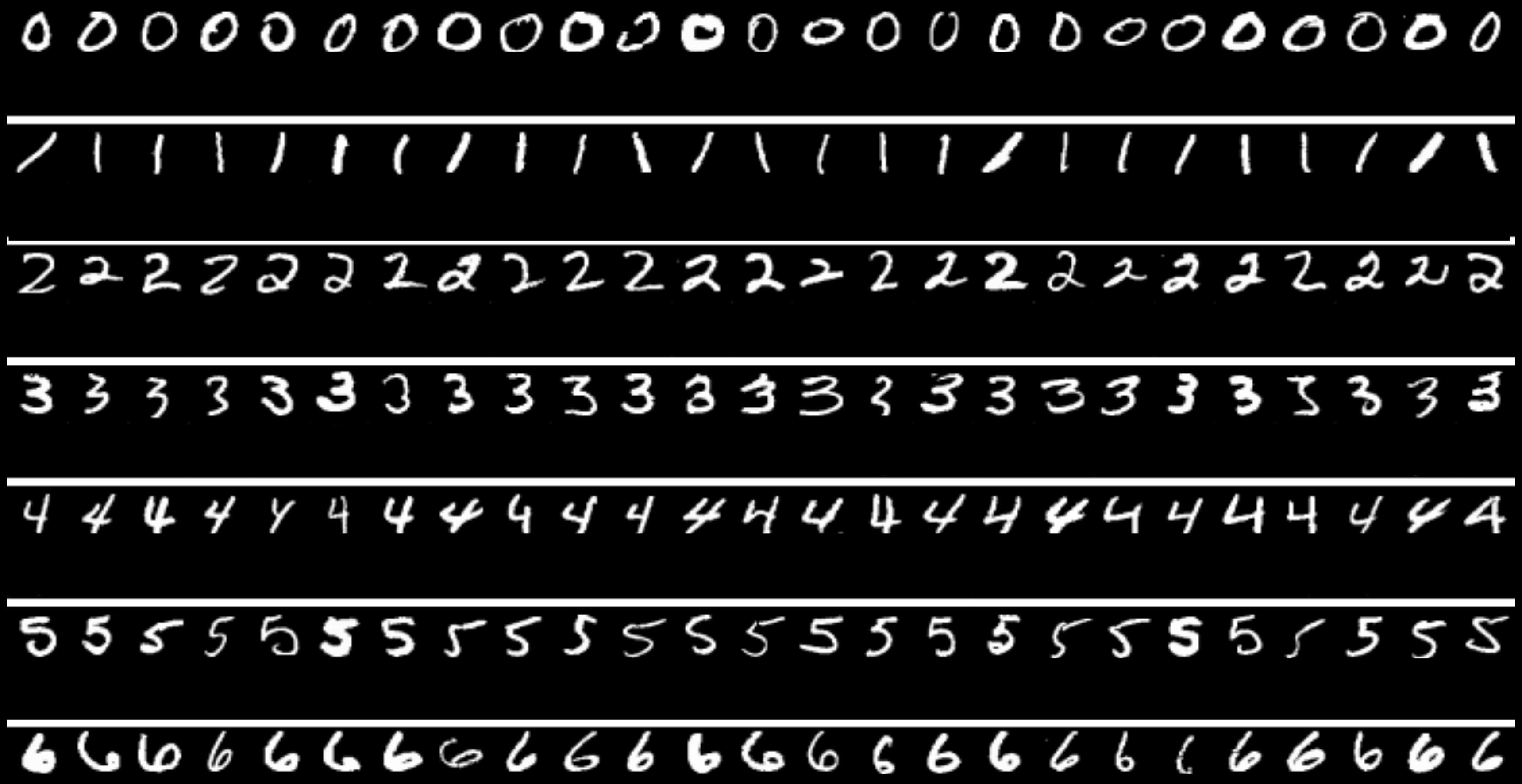
Visualization of Semantic Dictionaries



Learned codewords per class on MNIST dataset

Codewords captures different shapes and orientation in the dataset

Interpreting CNN decisions on Clean Samples



Every block of two rows

First row:
Input images from MNIST dataset

Interpreting CNN decisions on Clean Samples

Every block

First row:
Input images from MNIST dataset

Second row: **activated codewords** on MNIST dataset

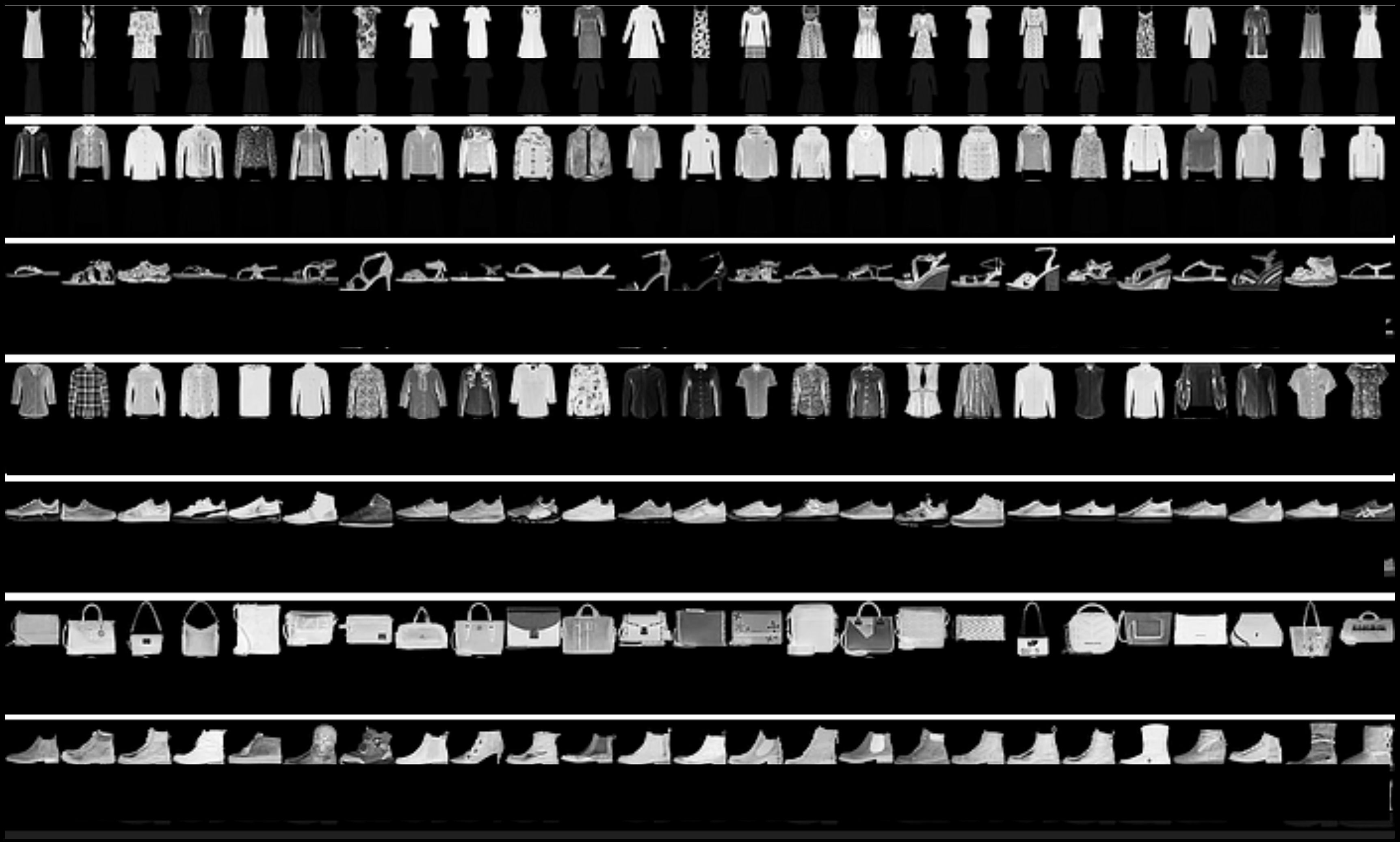
Activated codewords are similar in shape and orientation to input samples

Visualization of Semantic Dictionaries

Learned codewords per class on FMNIST dataset

Codewords captures different shapes and orientation in the dataset

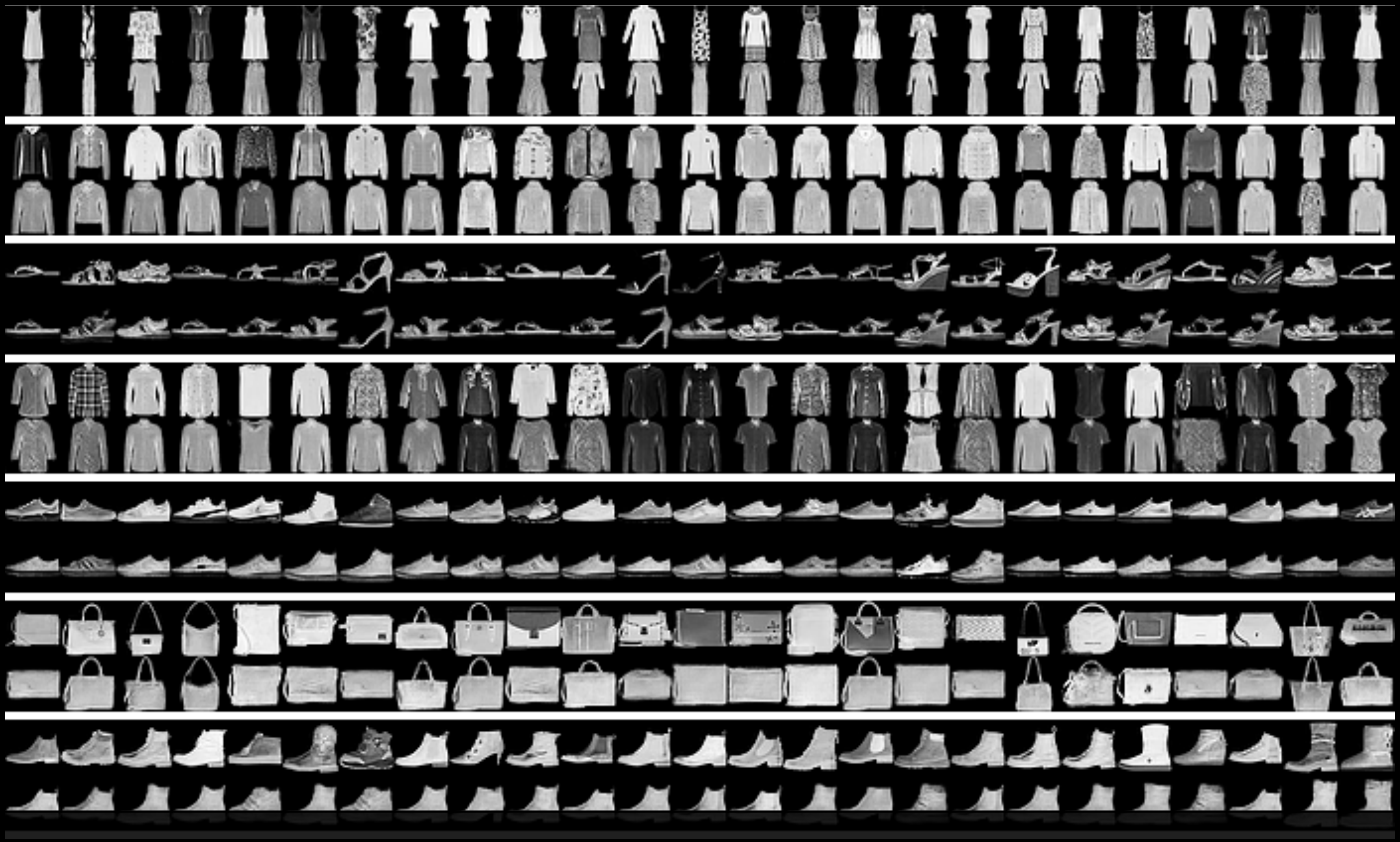
Interpreting CNN decisions on Clean Samples



Every block of two rows

First row:
Input images from
MNIST dataset

Interpreting CNN decisions on Clean Samples



Every block

First row:
Input images from MNIST dataset

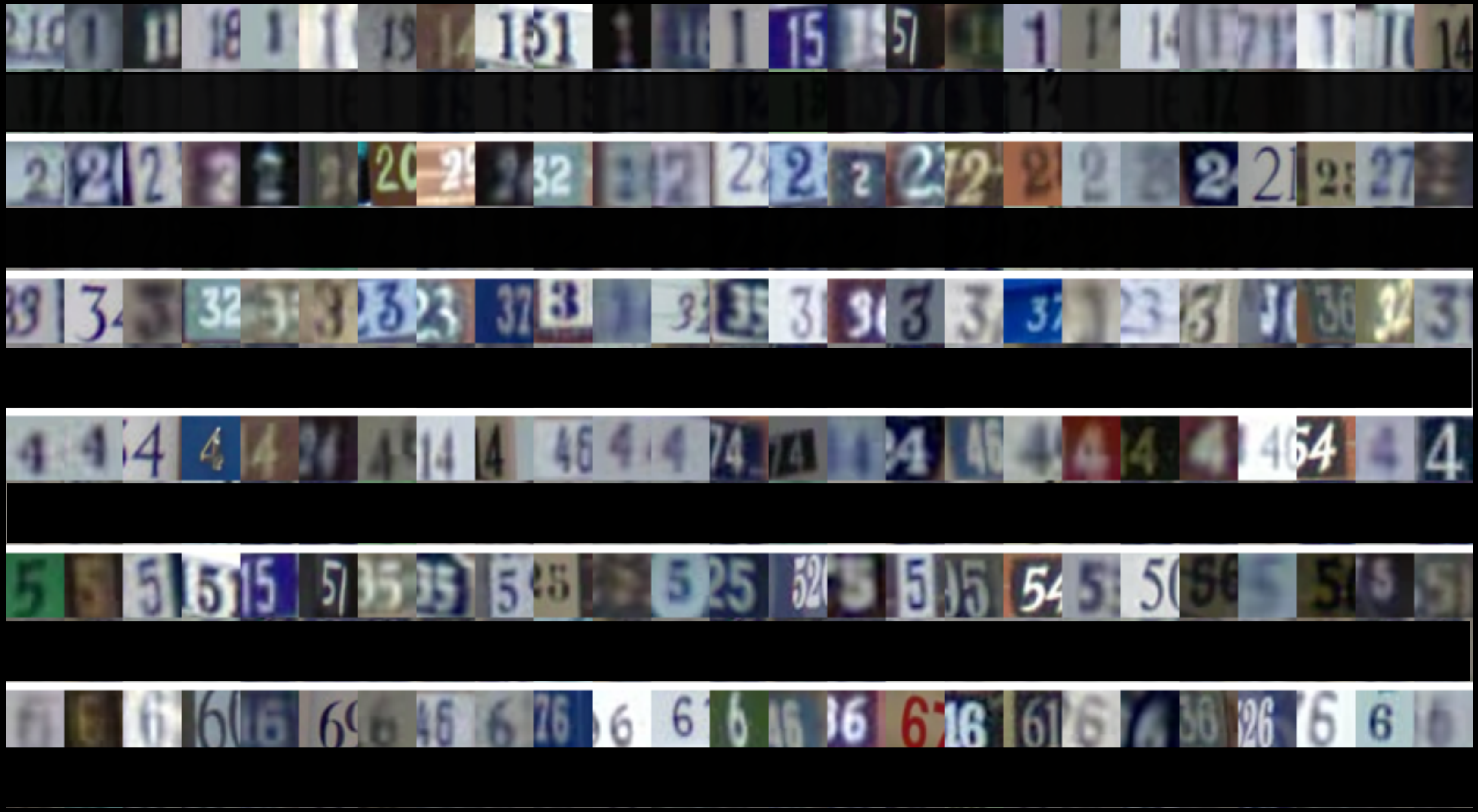
Second row: **activated codewords** on MNIST dataset

Visualization of Semantic Dictionaries

Learned codewords per class on SVHN dataset

Codewords captures variance present in the dataset

Interpreting CNN decisions on Clean Samples



Every block of two rows

First row:
Input images from
SVHN dataset

Interpreting CNN decisions on Clean Samples

Every block

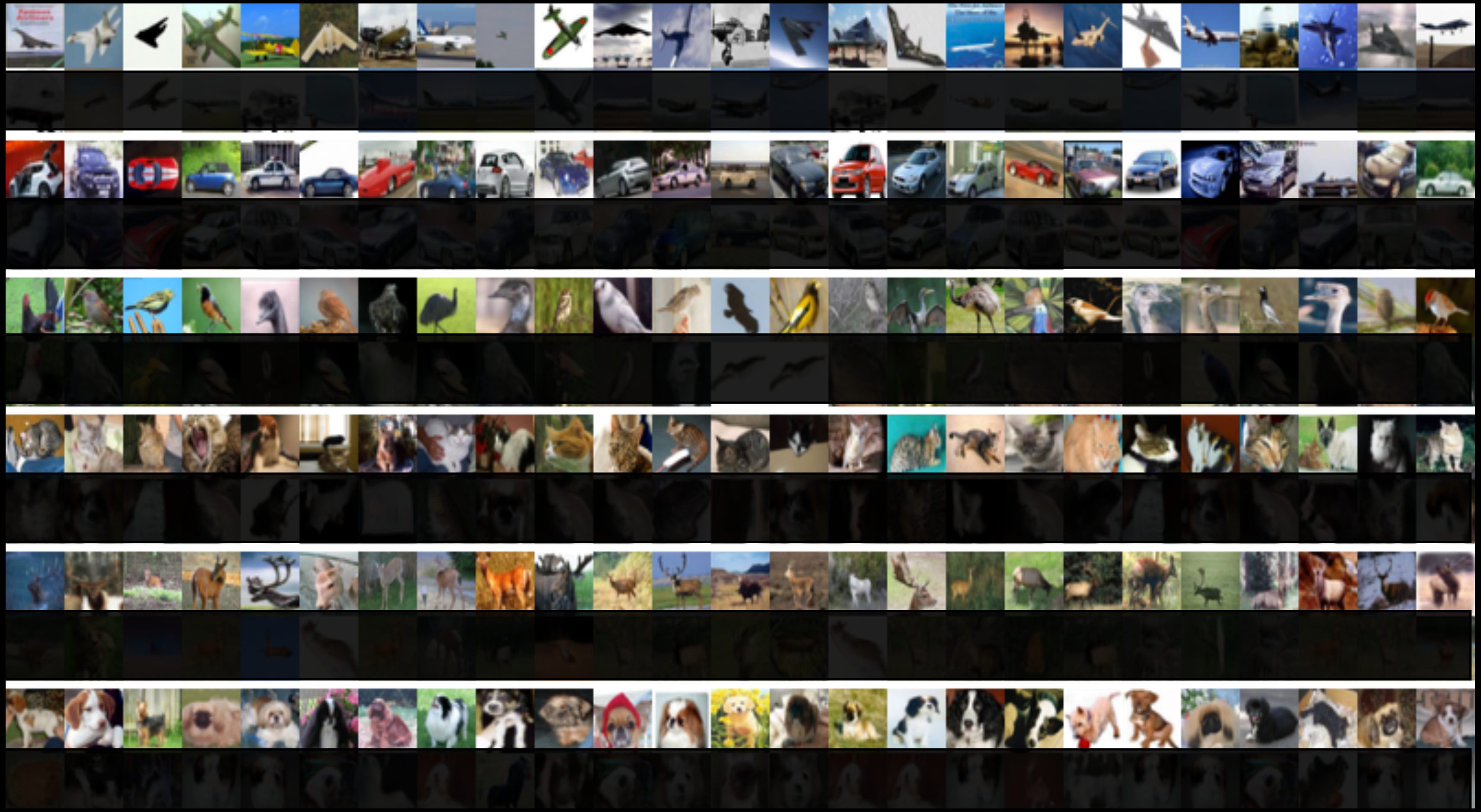
First row:
Input images from SVHN dataset

Second row: **activated codewords** on SVHN dataset

Visualization of Semantic Dictionaries

Learned codewords on CIFAR-10 dataset, ordered by class labels

Interpreting CNN decisions on Clean Samples

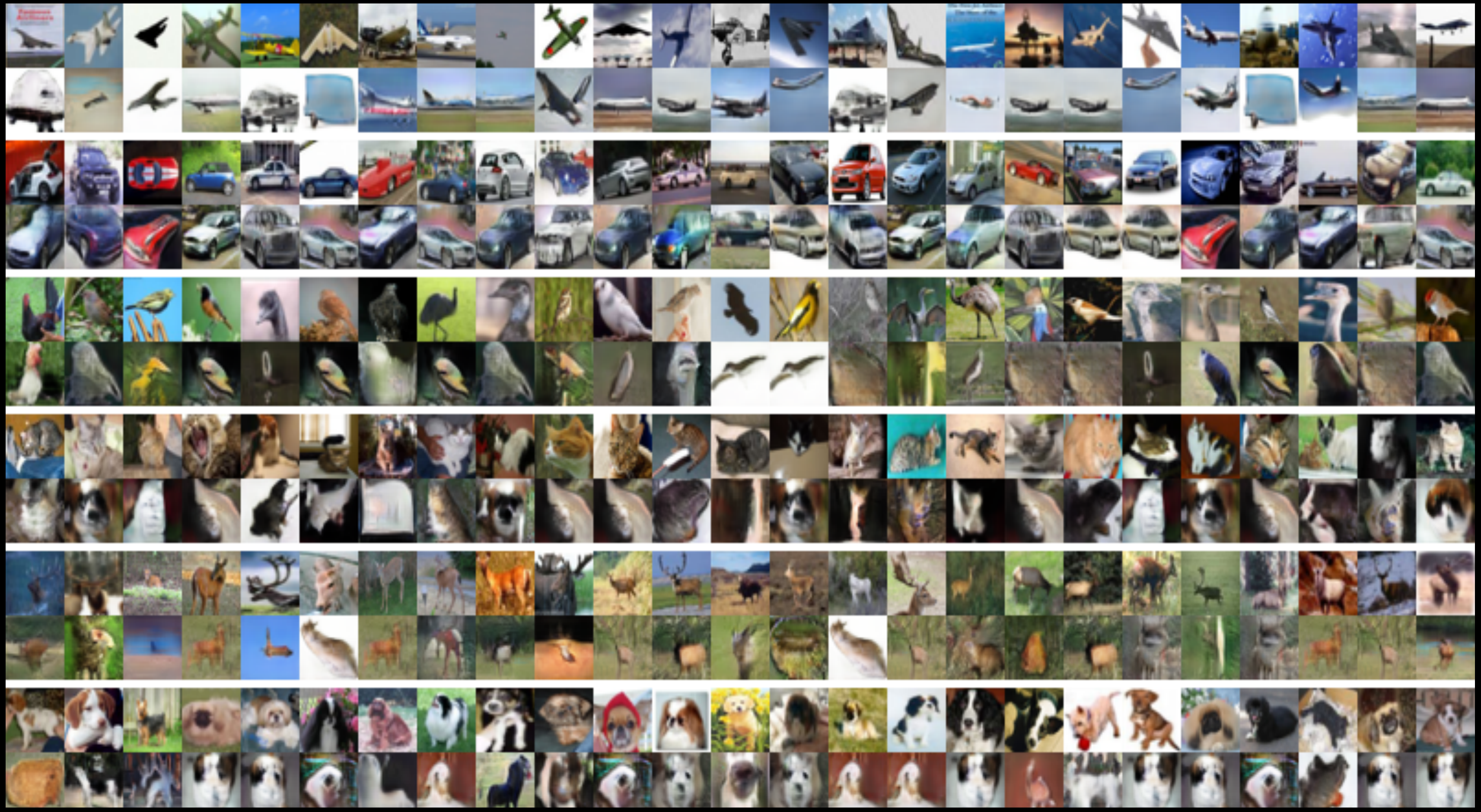


Every block of two rows

First row:
Input images from CIFAR-10 dataset

Activated codewords are similar to class label of the input image

Interpreting CNN decisions on Clean Samples



Every block

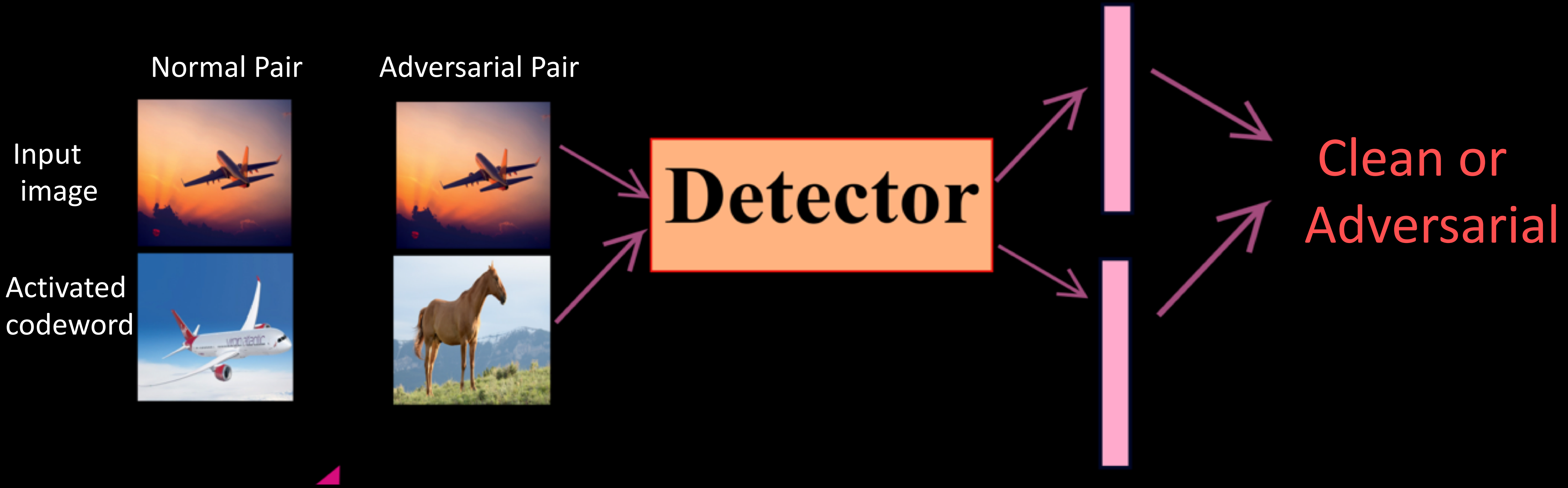
First row:
Input images from CIFAR-10 dataset

Second row: **activated codewords** on CIFAR-10 dataset

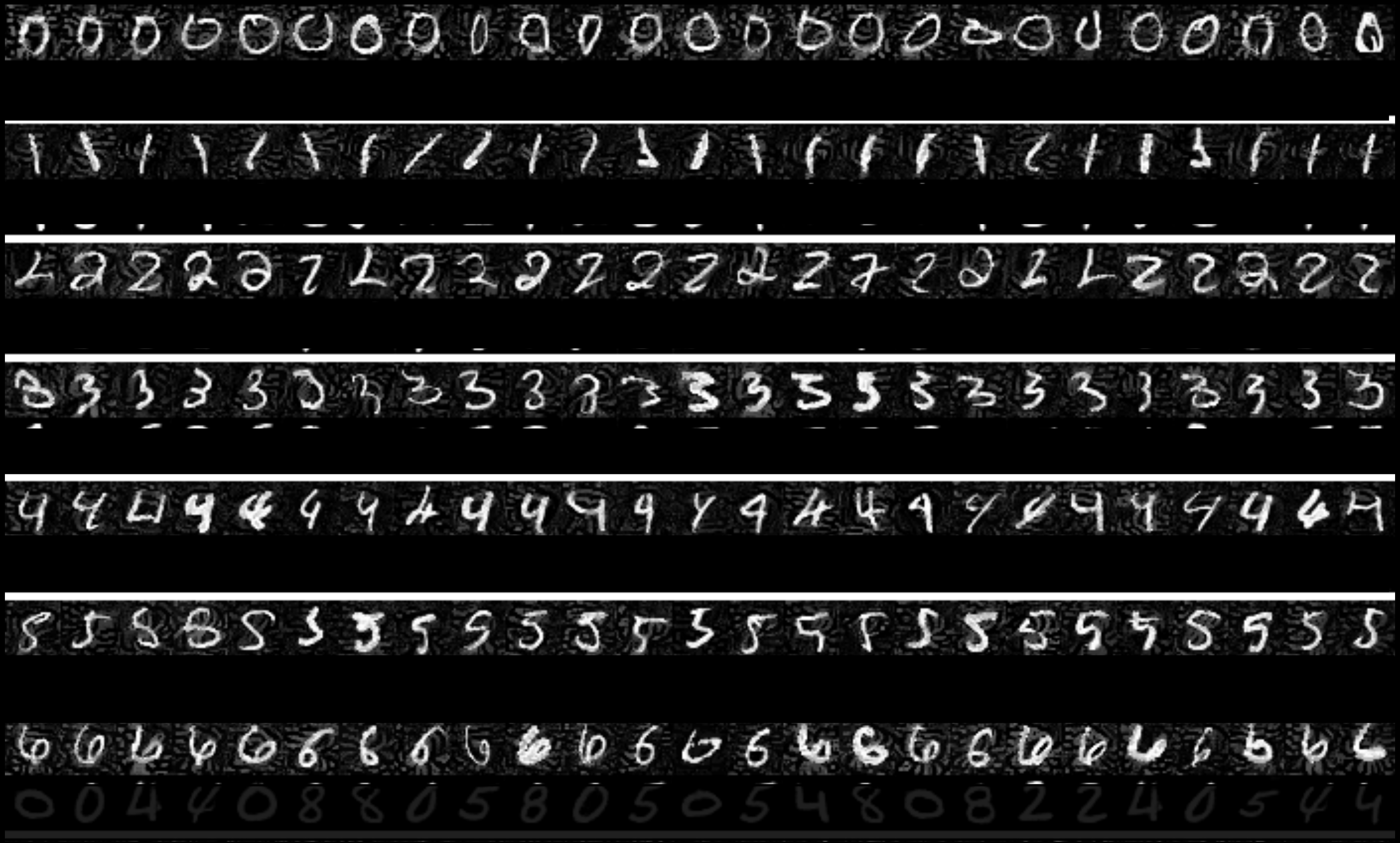
Adversarial Attack Detection

Adversarial Attack Detector

Computes similarity between activated visual codeword and input image using a Siamese network



Interpreting CNN decisions on Adversarial Samples (IFGSM



Every block of two rows

First row:
Adversarial images from MNIST dataset

Interpreting CNN decisions on Adversarial Samples (IFGSM

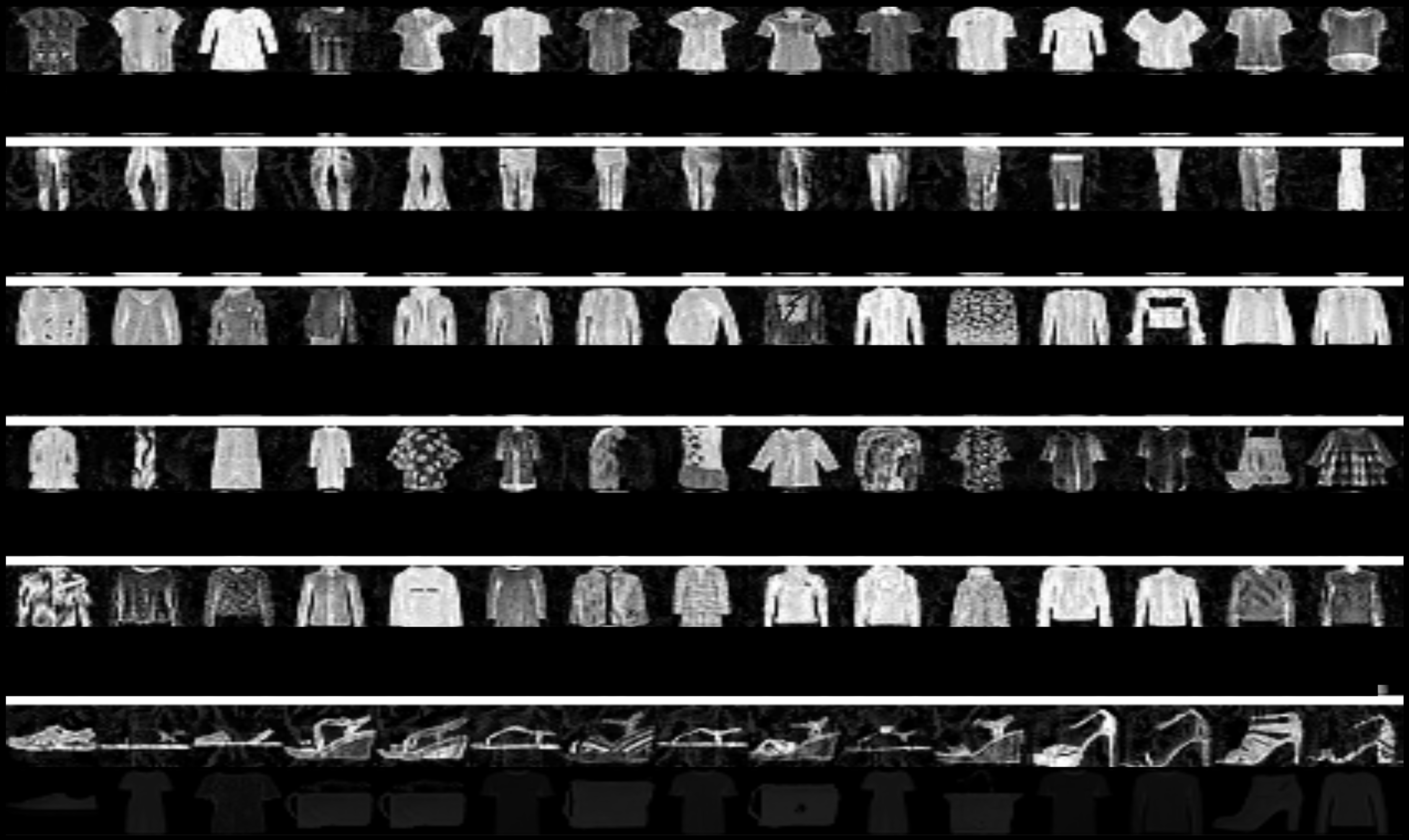
Every block

First row:
Input images from MNIST dataset

Second row: **activated codewords** on MNIST dataset

Activated codewords are dissimilar to input image in class labels

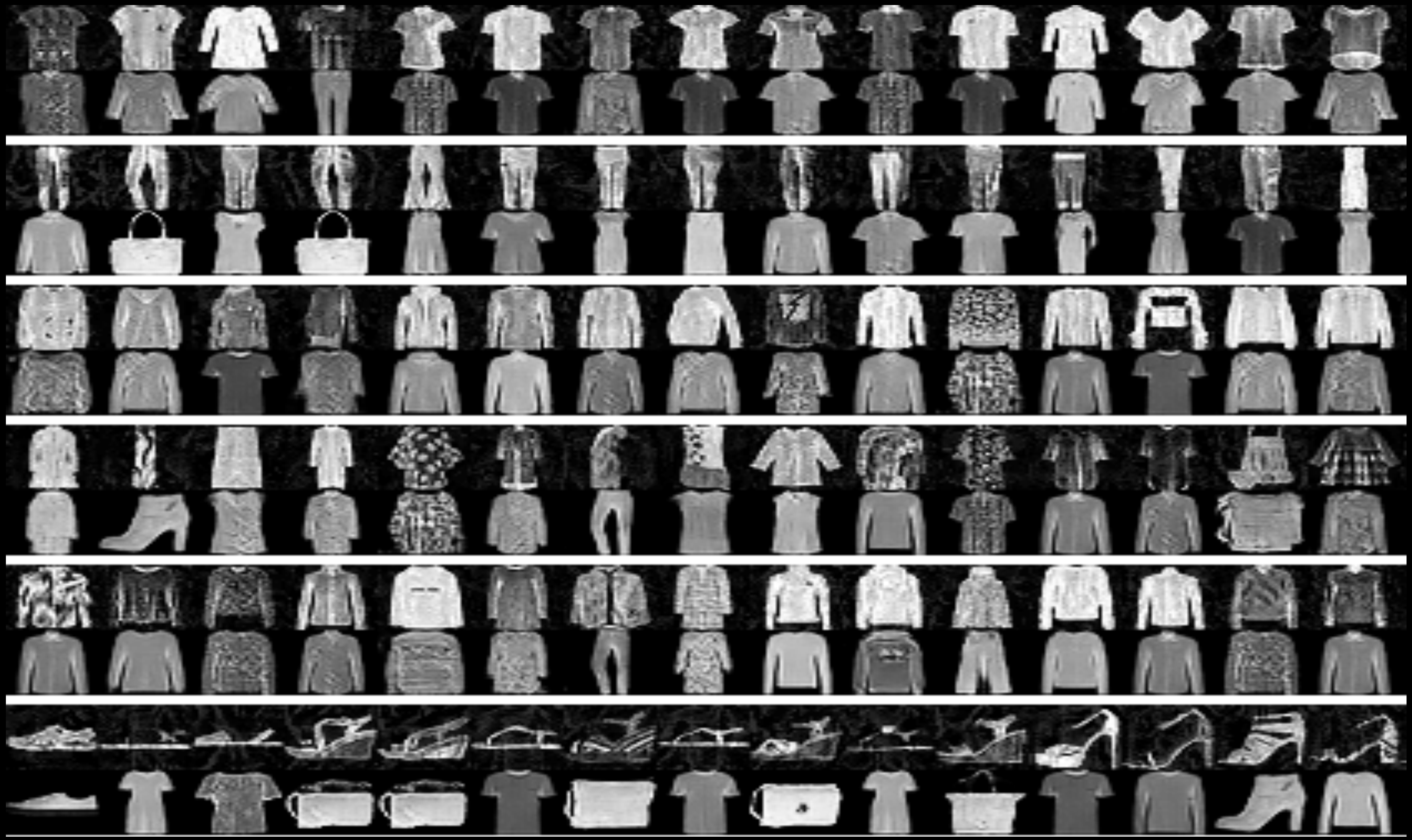
Interpreting CNN decisions on Adversarial Samples (IFGSM



Every block of two rows

First row:
Adversarial images
from FMNIST dataset

Interpreting CNN decisions on Adversarial Samples (IFGSM



Every block

First row:
Input images from FMNIST dataset

Second row: **activated codewords** on FMNIST dataset

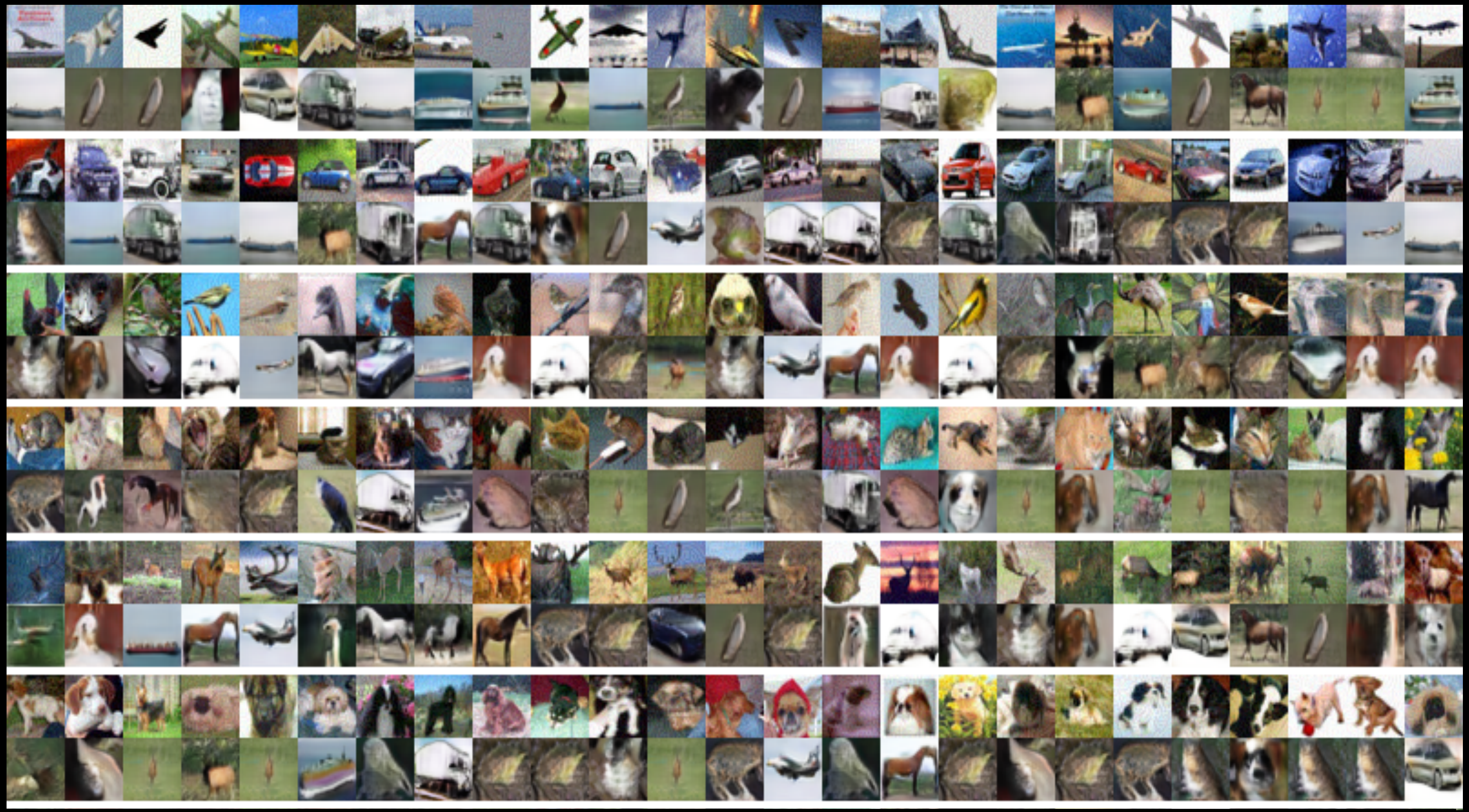
Activated codewords are dissimilar to input image in class labels

Interpreting CNN decisions on Adversarial Samples (IFGSM

Every block of two rows

First row:
Adversarial images
from CIFAR dataset

Interpreting CNN decisions on Adversarial Samples (IFGSM



Every block

First row:
Input images from
CIFAR dataset

Second row: **activated codewords** on CIFAR dataset

Activated codewords are dissimilar to input image in class labels

RESULTS

Adversarial Attack Detection on BoW Networks

AUROC

Dataset	Feature	FGSM	BIM-a	BIM-b	CW
MNIST	MD[1]	100.0	100.0	100.0	67.18
	Ours-GAN				
F-MNIST	MD[1]	98.5	85.49	89.4	72.7
	Ours-GAN				
SVHN	MD[1]	99.7	83.0	93.4	89.6
	Ours-GAN				
CIFAR-10	MD[1]	97.9	84.7	84.5	91.6
	Ours-GAN				

1. K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial Attacks, NIPS 2018

RESULTS

Adversarial Attack Detection on BoW Networks

AUROC

Dataset	Feature	FGSM	BIM-a	BIM-b	CW
MNIST	MD[1]	100.0	100.0	100.0	67.18
	Ours-GAN	100.0	100.0	100.0	100.0
F-MNIST	MD[1]	98.5	85.49	89.4	72.7
	Ours-GAN	100.0	100.0	100.0	97.9
SVHN	MD[1]	99.7	83.0	93.4	89.6
	Ours-GAN	100.0	96.2	100.0	96.5
CIFAR-10	MD[1]	97.9	84.7	84.5	91.6
	Ours-GAN	99.9	97.5	99.9	96.3

1. K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial Attacks, NIPS 2018

RESULTS

Adversarial Attack Detection on Base Networks

AUROC

Dataset	Feature	FGSM	BIM-a	BIM-b	CW
MNIST	MD[1]	100.0	100.0	100.0	99.96
	Ours-GAN				
F-MNIST	MD[1]	98.2	94.0	99.1	95.8
	Ours-GAN				
SVHN	MD[1]	99.8	81.3	99.4	92.6
	Ours-GAN				
CIFAR-10	MD[1]	97.5	76.7	99.9	96.4
	Ours-GAN				

1. K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial Attacks, NIPS 2018

RESULTS

Adversarial Attack Detection on Base Networks

AUROC

Dataset	Feature	FGSM	BIM-a	BIM-b	CW
MNIST	MD[1]	100.0	100.0	100.0	99.96
	Ours-GAN	100.0	100.0	100.0	100.0
F-MNIST	MD[1]	98.2	94.0	99.1	95.8
	Ours-GAN	100.0	100.0	100.0	97.9
SVHN	MD[1]	99.8	81.3	99.4	92.6
	Ours-GAN	100.0	96.3	100.0	96.9
CIFAR-10	MD[1]	97.5	76.7	99.9	96.4
	Ours-GAN	99.7	96.9	99.8	97.2

1. K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial Attacks, NIPS 2018

RESULTS

Generalization to Different Attacks

AUROC

Dataset	Feature	Train	FGSM	BIM-a	BIM-b	CW
MNIST	LID	FGSM	91.2	65.5	64.3	29.0
	MD	FGSM	100.0	99.9	99.9	34.7
	Ours-GAN	BIM-a	100.0	100.0	100.0	97.5
F-MNIST	LID	FGSM	93.9	82.2	82.5	65.0
	MD[1]	FGSM	98.5	87.8	90.5	64.0
	Ours-GAN	CW	97.3	91.1	95.8	97.9
SVHN	LID	FGSM	99.2	77.5	79.8	75.1
	MD[1]	FGSM	99.7	69.5	78.7	79.1
	Ours-GAN	BIM-a	91.4	96.2	91.3	94.7
CIFAR-10	LID	FGSM	89.2	66.5	68.3	66.0
	MD[1]	FGSM	97.9	65.3	80.3	60.0
	Ours-GAN	BIM-a	86.9	97.5	95.0	95.4

1. K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial Attacks, NIPS 2018

Results

White Box Detector Attacks

Expose the detector network to attacker accuracy

Dataset	Attack Detector	Success Rate without AT	Success Rate with AT
MNIST	FGSM CW	0.0 100.0	0.0 8.0
F-MNIST	FGSM CW	2.8 100.0	0.0 27.2

Adversarial training (AT) of the detector makes robust to white box detector attackls

RESULTS

Detecting Out-of-distribution Samples

In-Dataset	Out-Dataset	Baseline	MD[1]	Ours-GAN
SVHN	CIFAR-10	87.4	95.3	
	LSUN	89.1	99.3	
	Tiny ImageNet	90.0	98.8	
MNIST	Not-MNIST	77.1	85.8	
	OMNIGLOT	82.1	99.3	
	CIFAR	79.8	99.7	

AUROC

1. K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial Attacks, NIPS 2018

RESULTS

Detecting Out-of-distribution Samples

In-Dataset	Out-Dataset	Baseline	MD[1]	Ours-GAN
SVHN	CIFAR-10	87.4	95.3	97.3
	LSUN	89.1	99.3	99.9
	Tiny ImageNet	90.0	98.8	99.9
MNIST	Not-MNIST	77.1	85.8	99.9
	OMNIGLOT	82.1	99.3	100.0
	CIFAR	79.8	99.7	100.0

AUROC

1. K. Lee, K. Lee, H. Lee, and J. Shin. A simple unified framework for detecting out-of-distribution samples and adversarial Attacks, NIPS 2018

Conclusion

By providing visual representation to BoW codeword filters in deep CNNs,

- We interpret the decisions of a CNN
- Leverage the activated codewords to detect adversarial and Out of distribution examples

Thank You

Multi-stage Training → End-to-end Training

ConvNet Features + SVMs

MultiScale Orderless Pooling [Gong et al., 2014]

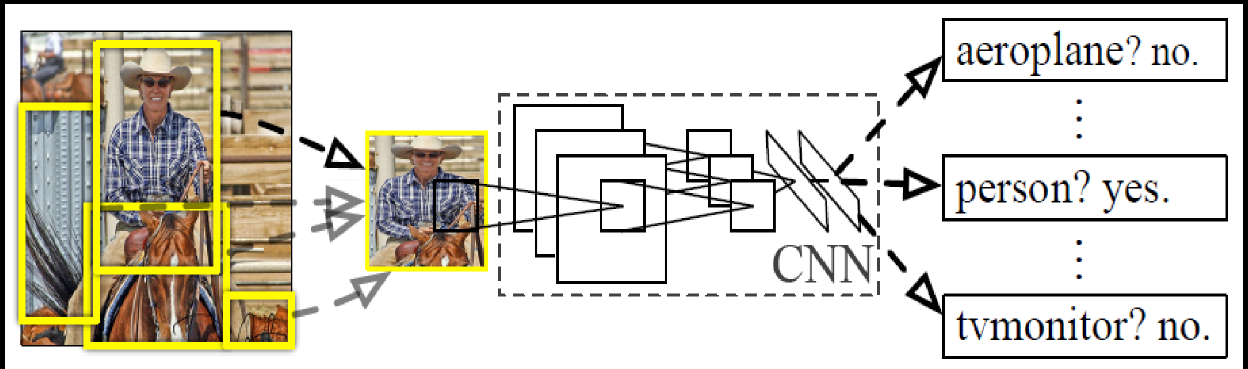
Deep Structured Representation

NetVLAD [Arandjelović et al., 2016]

FisherNet [Peng et al., 2017]

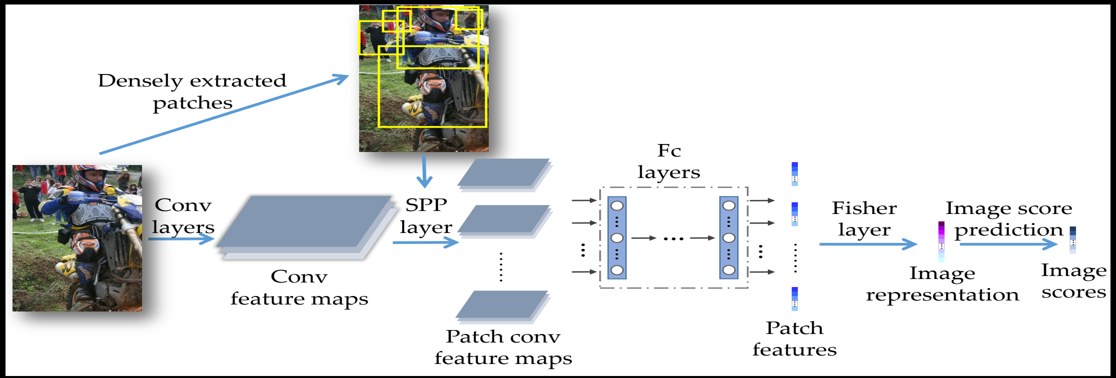
Stage 1 — Features

Stage 2 — VLAD+SVMs



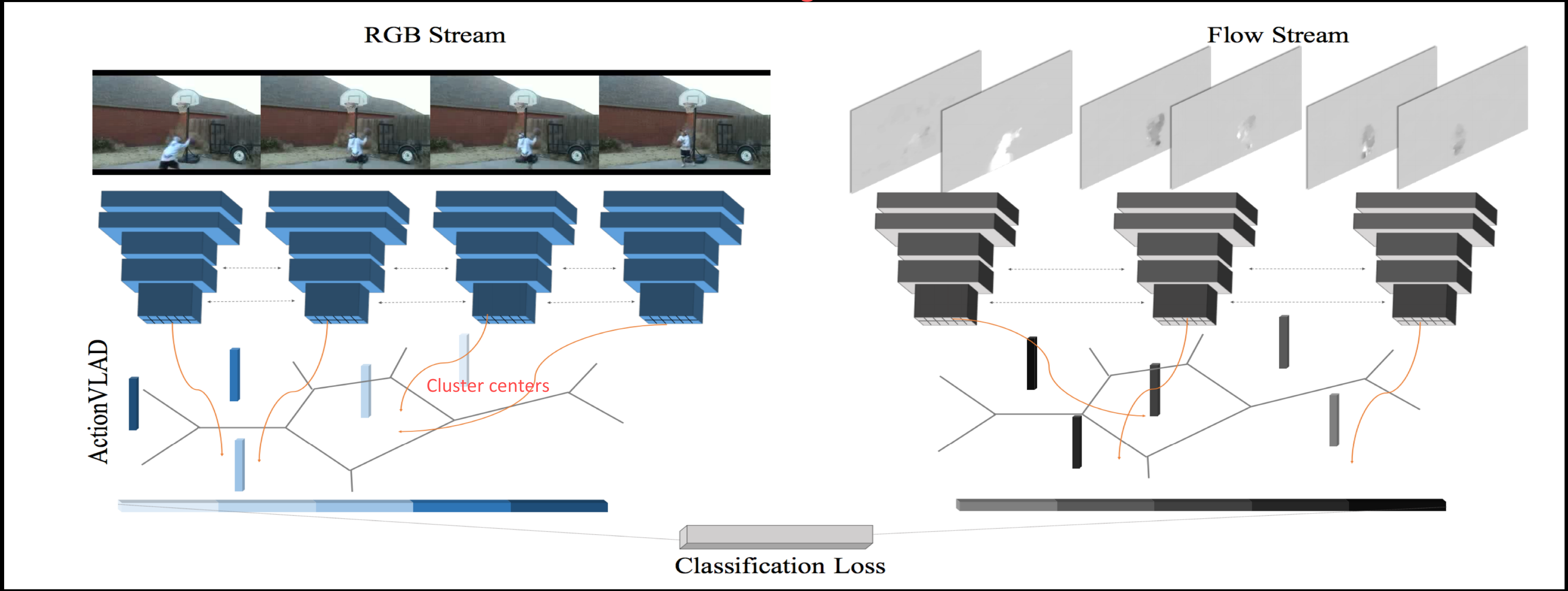
End-to-end

Joint Feature + Classification



Action VLAD: Spatio-temporal aggregation

Action recognition



Features are pooled across space and time using the ActionVLAD pooling layer

Rohit et al., ActionVLAD: Learning spatio-temporal aggregation for action classification, 2017