Unsupervised Representation Learning via Neural Activation Coding

Yookoon Park[1], Sangho Lee[2], Gunhee Kim[2] and David M. Blei[1]

[1]Computer Science Department, Columbia University, New York, USA

[2]Computer Science Department, Seoul National University, Seoul, South Korea

The Goal of Unsupervised Representation Learning

- Learn an encoder network f_θ on unlabeled data X
  - Which produces representation Z of the data
- Evaluated on its performance on *downstream tasks* e.g. classification
  - Downstream models take Z as input
- Commonly simple linear models are used in downstream
- E.g. pretrain a CNN encoder on unlabeled natural images
  - Attach a linear classifier to the encoder to solve an image classification task

Unsupervised Representation Learning So Far

- Self-supervised learning: formulate *pretext* tasks
  - Generate artificial pseudo-labels to train the encoder
    - Predict spatial context (Doersch et al., 2015)
    - Solve jigsaw puzzle (Noroozi and Favaro, 2016)
    - Predict image rotations (Gidaris et al., 2018)
- Recently, contrastive representation learning
  - Maximize the mutual information between the data and representation $I(X, Z)$
    - Instance discrimination (Wu et al. 2018)
    - Constrative predictive coding (Oord et al. 2018)
    - Momentum contrast (He et al. 2020)
    - SimCLR (Chen et al. 2020)

Our Approach: Neural Activation Coding (NAC)

- **Novelty**: maximize the *nonlinear expressivity* of the encoder
  - A fundamentally new perspective for unsupervised representation learning
- To this end, we formulate a communication problem over a noisy channel
  - Leads to maximum nonlinear expressivity for ReLU encoders
- NAC learns *both* continuous and discrete representations of data
  - Evaluated on 1. linear classification and 2. nearest neighbor search
- Unsupervised encoder pretraining for deep generative models

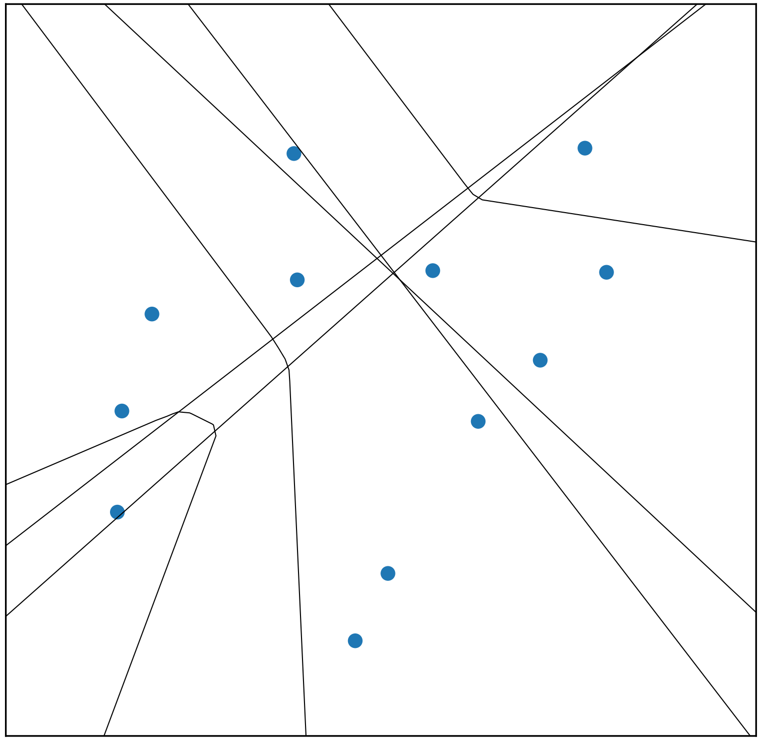
Nonlinear Expressivity of Neural Networks

- ReLU activation networks are piece-wise linear functions
- They divide the input space into a set of locally linear regions
- Nonlinear expressivity $\approx$ # of distinct linear regions (Pascanu et al., 2013)

Why Nonlinear Expressivity?

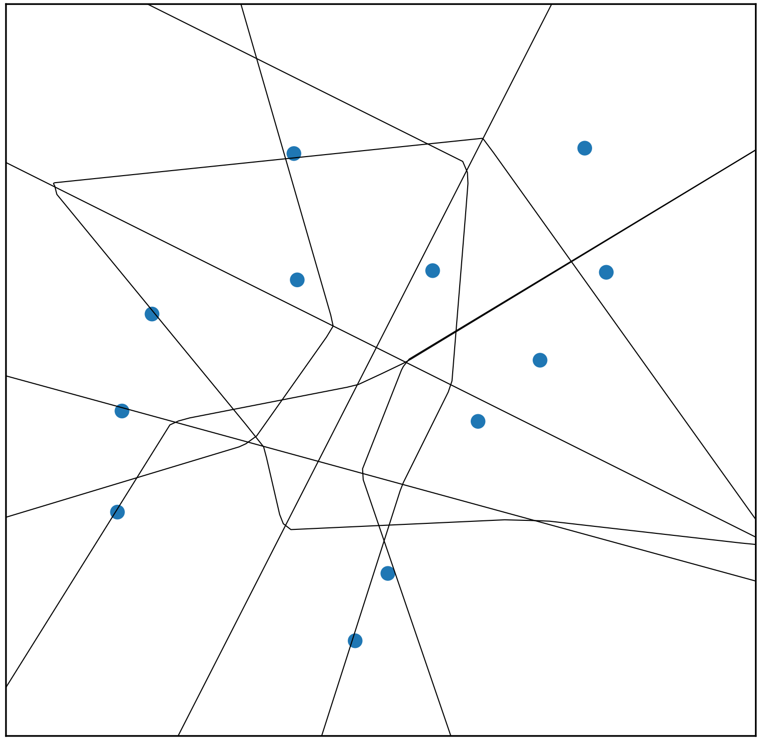
- Visualize linear regions of a ReLU encoder

Low nonlinear expressivity



At initialization

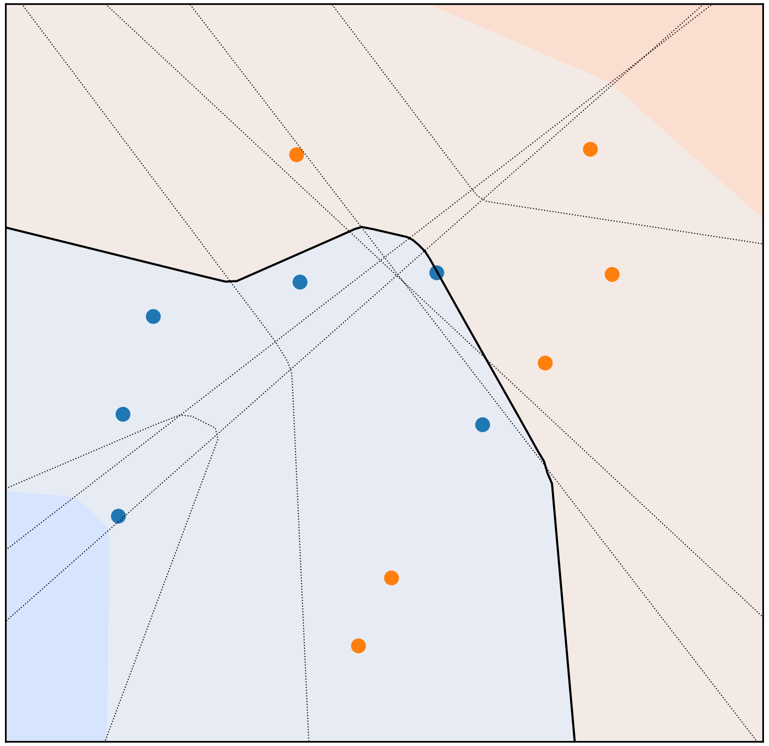
High nonlinear expressivity



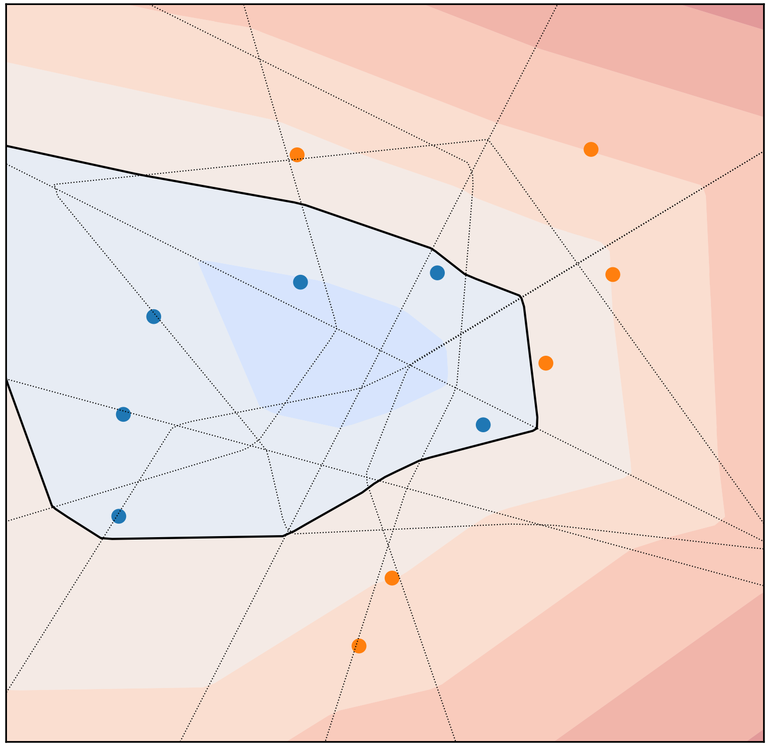
After NAC training

Why Nonlinear Expressivity?

- Solving downstream linear classification



High training error



Zero training error

Activation Code and Nonlinear Expressivity

- A ReLU activation encoder

$$\mathbf{a}^{(l)} = \mathbf{W}^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)},$$
$$\mathbf{h}^{(l)} = \mathrm{ReLU}(\mathbf{a}^{(l)}), \quad l = 1, 2, \ldots, L$$

- We define the ***activation code*** as: $\mathbf{c}^L = \mathrm{sgn}(\mathbf{a}^L) \in \{-1, 1\}^D$

- Each activation codeword is associated with a linear region of the encoder

Activation Code and Nonlinear Expressivity

- The encoder maps the training examples to activation codewords

$$\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n \xrightarrow{\text{Encoder } f_\theta} \mathbf{c}_1, \mathbf{c}_2, ..., \mathbf{c}_n$$

- The Hamming distance between two codewords $d_H(\mathbf{c}_i, \mathbf{c}_j) = (D - \langle \mathbf{c}_i, \mathbf{c}_j \rangle)/2$
  $\approx$ the number of linear regions between $\mathbf{x}_i, \mathbf{x}_j$

- **High distance between codewords → high number of linear regions**
  → high nonlinear expressivity

Neural Activation Coding (NAC)

- Communication problem over a noisy channel $\mathbf{X} \rightarrow \widetilde{\mathbf{X}} \rightarrow \mathbf{C} \rightarrow \widetilde{\mathbf{C}}$

Sender $\mathbf{x}_i \xrightarrow{\text{Augmentations}} \widetilde{\mathbf{x}}_i \xrightarrow{\text{Encoder } f_\theta} \mathbf{c}_i \xrightarrow{\text{Noisy channel}} \widetilde{\mathbf{c}}_i$ **Receiver**

Training data Perturbed data Activation code Noisy code

- Maximize the mutual information $I(\mathbf{X}, \widetilde{\mathbf{C}}) = \mathbb{E}_{P_\theta(\mathbf{x},\tilde{\mathbf{c}})}\left[\log \frac{P_\theta(\tilde{\mathbf{c}}|\mathbf{x})}{P_\theta(\tilde{\mathbf{c}})}\right]$

- **Learning for noise-robust activation codewords**
  → maximum distance codewords → maximum nonlinear expressivity

Mutual Information Lower-bound

- Amortized variational inference: introduce an inference network $Q_\phi(\tilde{\mathbf{c}}|\mathbf{x})$

$$\mathbb{E}_{P_\theta(\mathbf{x},\tilde{\mathbf{c}})}[\log P_\theta(\tilde{\mathbf{c}}|\mathbf{x})] \geq \mathbb{E}_{P_\theta(\mathbf{x},\tilde{\mathbf{c}})}[\log Q_\phi(\tilde{\mathbf{c}}|\mathbf{x})]$$

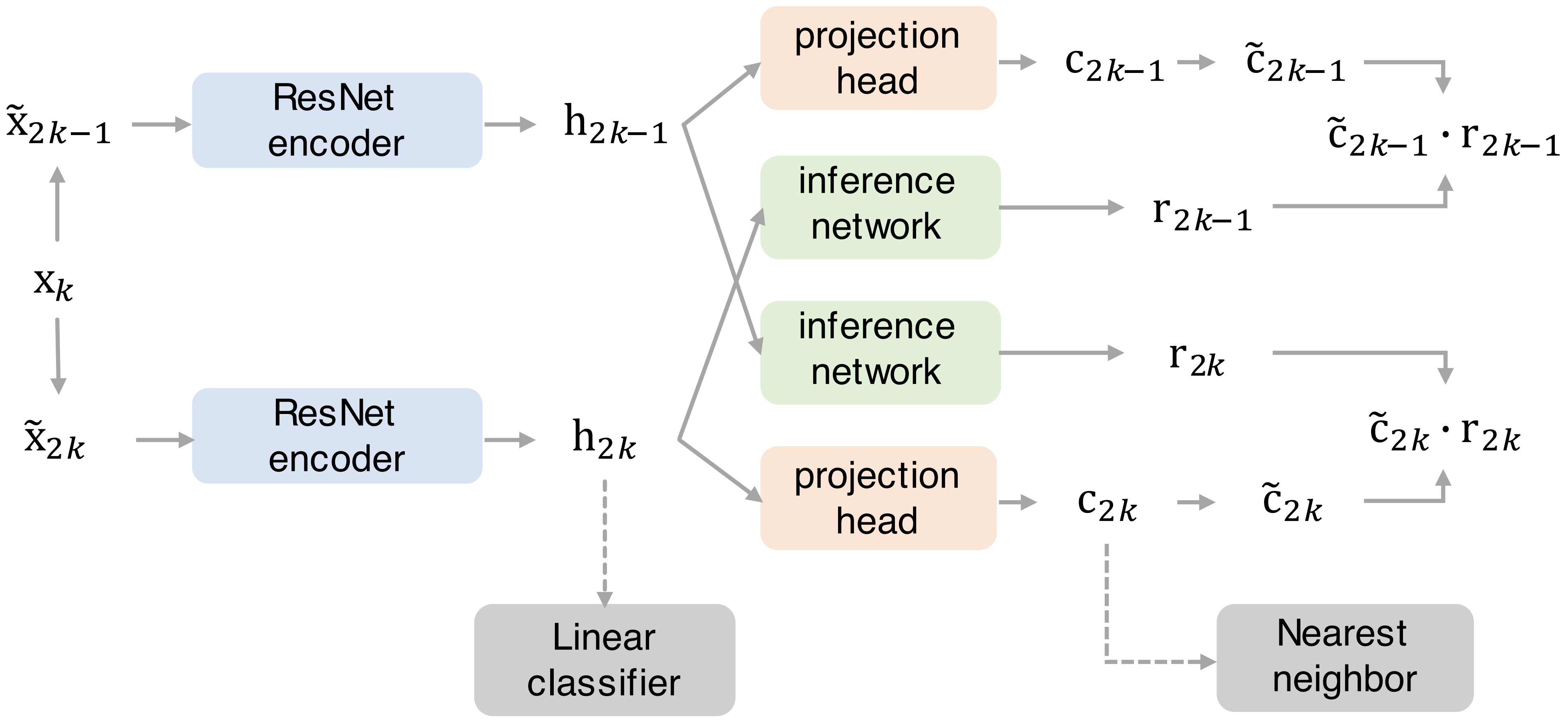
- Subsampling (Poole et al., 2019)

$$\mathbb{E}_{\tilde{\mathbf{c}}}\left[\log \frac{1}{P_\theta(\tilde{\mathbf{c}})}\right] \geq \mathbb{E}_{\tilde{\mathbf{c}},\mathbf{c}_1,\ldots\mathbf{c}_{2K}}\left[\log \frac{1}{\frac{1}{2K}\sum_{k=1}^{2K} P(\tilde{\mathbf{c}}|\mathbf{c}_k)}\right]$$

- Optimization using continuous relaxation to the activation code

$$\mathbf{c} = \mathrm{sgn}(\mathbf{a}) \leftarrow \mathbf{z} = \tanh(\mathbf{a})$$

Model Architecture



Experiments

- NAC learns both *continuous* and *discrete* representations of data

- We evaluate them respectively on
  1. Linear classification on CIFAR-10 / ImageNet-1K
  2. Nearest neighbor search on CIFAR-10 / FLICKR-25K

- Can enhanced encoder expressivity improve the training of VAEs?

Linear Image Classification

- ResNet-50 encoder + linear classifier

Linear classification accuracy (%)

Model	CIFAR-10	ImageNet-1K
InsDis (Wu et al., 2018)	80.8	54.0
SimCLR (Chen et al., 2020a)	92.8*	66.6
MoCo-v2 (Chen et al., 2020b)	91.6*	**67.5**
NAC	**93.9**	65.0

* Re-implemented for multi GPU training

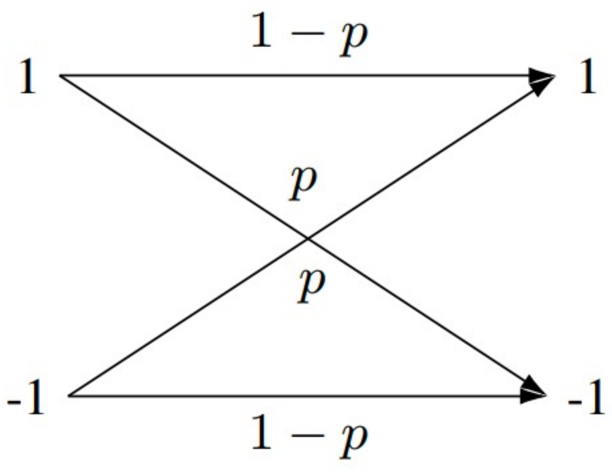
Nearest Neighbor Search using Deep Hash Codes

Mean average precision (%) on nearest neighbor retrieval

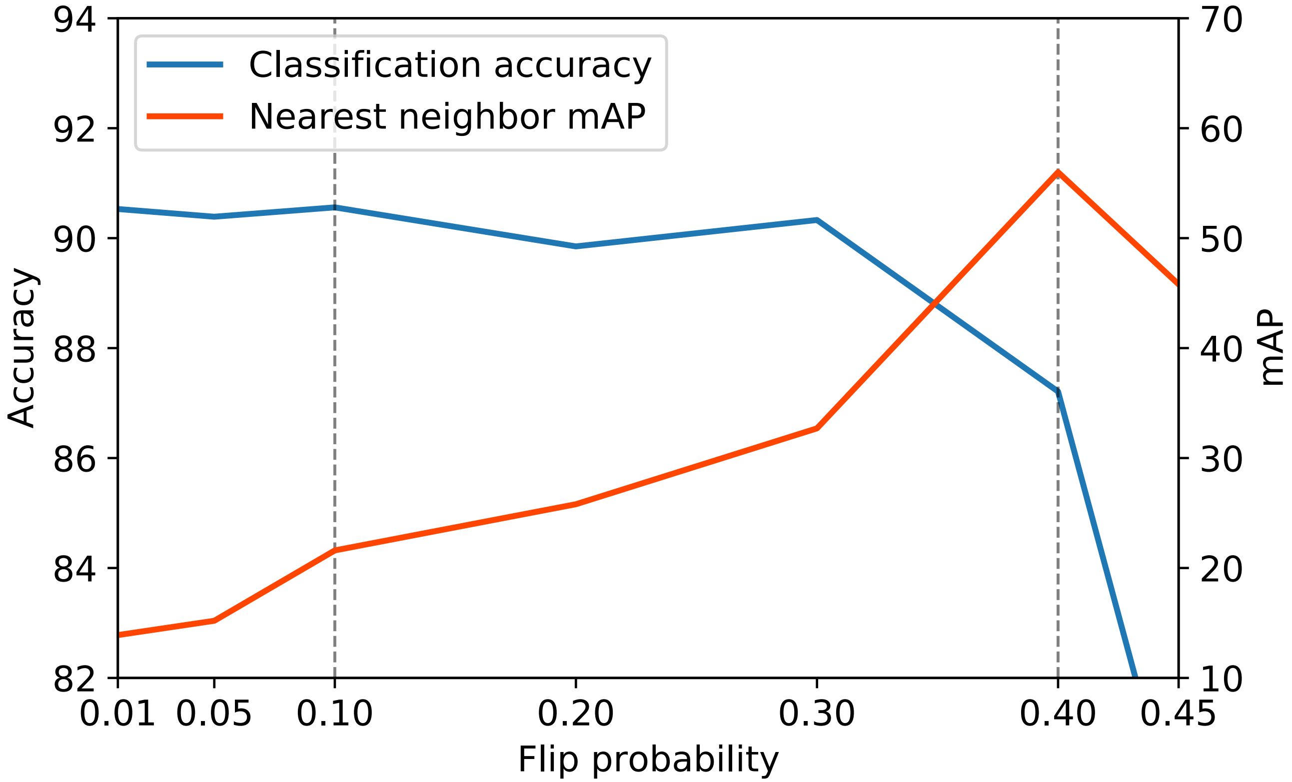
Model	CIFAR-10	FLICKR-25K
Deep hashing methods		
DeepBit (Lin et al., 2016)	25.3	59.3
SSDH (Yang et al., 2018)	26.0	66.2
DistillHash (Yang et al., 2019)	29.0	70.0
Contrastive learning methods		
MoCo-v2 (Chen et al., 2020b)	32.3	65.0
SimCLR (Chen et al., 2020a)	34.2	65.4
NAC	**40.5**	**70.8**

Effect of Symmetric Noise Channel on CIFAR-10

- Low noise level (≈ 0.1) is favorable for classification
- High noise level (≈ 0.4) benefits nearest neighbor search performance



Symmetric noise channel



Encoder Pretraining for Variational Autoencoders (VAEs)

- VAEs suffer from *encoder suboptimality* (Cremer et al., 2018)
  1. Random initialization → *cold start* problem
  2. The encoder is updated only once each iteration

- NAC pretraining improves the training of VAEs
  - High encoder expressivity at initalization → faster convergence, better inference

Encoder init.	Loglikelihood	KL divergence
Random	-3202	33.0
SimCLR	-3174	38.9
MoCo-v2	-3103	32.2
NAC	**-2865**	**71.8**

Thank you

Unsupervised Representation Learning via Neural Activation Coding

Yookoon Park, Sangho Lee, Gunhee Kim and David M. Blei

Code available at https://github.com/yookoon/nac

References

Pascanu, R., Montufar, G., and Bengio, Y. On the number of response regions of deep feed forward networks with piece-wise linear activations. In ICLR, 2013.

Doersch, C., Gupta, A., and Efros, A. A. Unsupervised visual representation learning by context prediction. In *ICCV*, 2015.

Lin, K., Lu, J., Chen, C.-S., and Zhou, J. Learning compact binary descriptors with unsupervised deep neural networks. In CVPR, 2016.

Yang, E., Deng, C., Liu, T., Liu, W., and Tao, D. Semantic structure-based unsupervised deep hashing. In IJCAI, 2018.

Cremer, C., Li, X., and Duvenaud, D. Inference suboptimality in variational autoencoders. In *ICML*, 2018.

Oord, A. v. d., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Yang, E., Liu, T., Deng, C., Liu, W., and Tao, D. Distillhash: Unsupervised deep hashing by distilling data pairs. In CVPR, 2019.

Poole, B., Ozair, S., Oord, A. v. d., Alemi, A. A., and Tucker, G. On variational bounds of mutual information. In ICML, 2019.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In ICML, 2020a.

Chen, X., Fan, H., Girshick, R., and He, K. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020b.