On Compact Codes for Spatially Pooled Features

Yangqing Jia Oriol Vinyals Trevor Darrell
UC Berkeley EECS, {jiayq,vinyals,trevor}@berkeley.edu

1. Summary

- The learning community has been in favor of feature extraction pipelines that use feedforward and over-complete representations.
- We link such pipeline with the Nyström sampling view to analyze the effect of the dictionary size (over-completeness) on the final classification performance.
- We derived a bound that predicts the performance of large codebook sizes with smaller experiments.
- Such a view leads to novel algorithms with complex feature extraction pipelines with efficient, scalable clustering algorithms.

2. Background

Vision:

- Simple clustering methods are effective in dictionary learning in single-layer networks [Coates et al. ICML11].
- Deeper models are built on layers of feedforward encoding methods.

Speech:

- Acoustic modeling was one of the first adopters to feedforward networks, but over-complete representations were not explored until recently.
- Adding more layers of coding is also helpful to achieve better modeling [Vinyals et al, IS13].

3. The Nyström Method

Let $\mathbf{C}$ be an $n \times n$ PSD matrix. The Nyström method defines:

$$\mathbf{C}' = \mathbf{E}\mathbf{W}^{+}\mathbf{E}^{\top},$$

where $\mathbf{E}$ is a $n \times k$ matrix with columns randomly sampled from $\mathbf{C}$:

$$\mathbf{E} = \begin{pmatrix} | & | & & | \\ \mathbf{c}_{\pi(1)} & \mathbf{c}_{\pi(2)} & \cdots & \mathbf{c}_{\pi(k)} \\ | & | & & | \end{pmatrix},$$

and $\mathbf{W}$ is the square $k \times k$ matrix by picking the same k columns and k rows from $\mathbf{C}$.

The matrix $\mathbf{C}'$ is a good approximation to $\mathbf{C}$ and the error is bounded by:

$$\|\mathbf{C} - \mathbf{C}'\|_F \leq \|\mathbf{C} - \mathbf{C}_r\|_F + \epsilon \max(n\mathbf{C}_{ii}),$$

valid if $k \geq 64r/\epsilon^4$. Thus,

$$\|\mathbf{C} - \mathbf{C}'\|_F \leq O + M\left(\frac{1}{k}\right)^{\frac{1}{4}},$$

4. Feature Encoding

- Consider common pipelines in feature coding, e.g. rectified linear units (ReLU) to encode feature $\mathbf{x}$ with dictionary $\mathbf{D}$:

$$\mathbf{c}(\mathbf{x}) = \max(0, \mathbf{x}^{\top}\mathbf{D})$$

- Suppose we take $\mathbf{D} = \mathbf{X}$ (all possible features) to have the best local coding so

$$\mathbf{C} = \max(0, \mathbf{X}^{\top}\mathbf{X})$$

which defines a linear kernel

$$\mathbf{K} = \mathbf{C}\mathbf{C}^{\top}$$

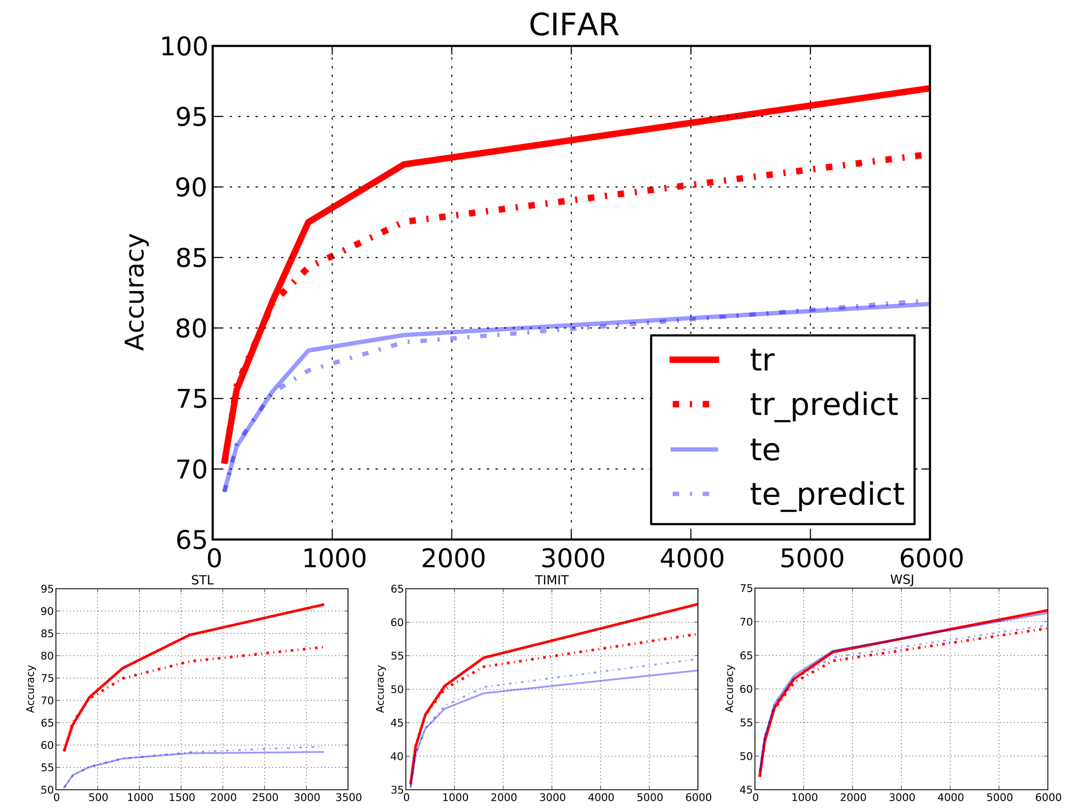
But we need a compact codebook!

- Applying Nyström method to $\mathbf{C}$ (instead of $\mathbf{K}$) we obtain

$$\begin{aligned} \mathbf{C}' &\approx \mathbf{C} = \mathbf{E}\mathbf{W}^{+}\mathbf{E}^{\top}, \text{ and} \\ \mathbf{K}' &\approx \mathbf{K} = \mathbf{C}'\mathbf{C}'^{\top} = \mathbf{E}\mathbf{W}^{+}\mathbf{E}^{\top}\mathbf{E}\mathbf{W}^{+}\mathbf{E}^{\top} \\ &= \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{\top}. \end{aligned}$$

5. Size Matters

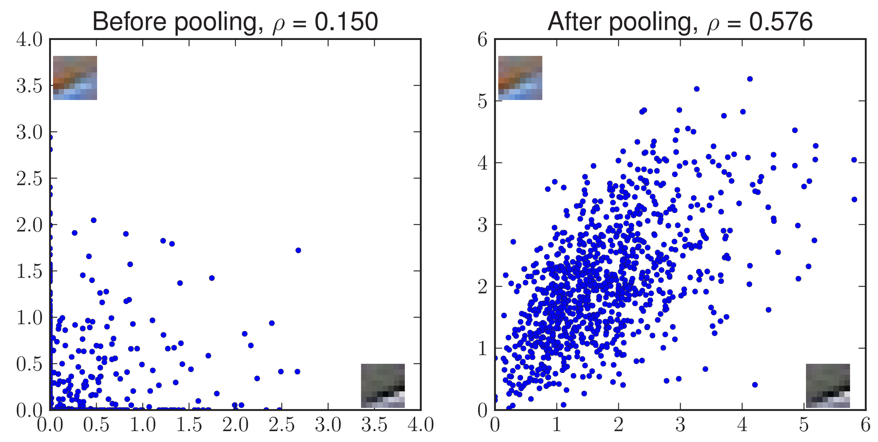
- We explain the good performance of codebooks learned by K-means [Coates et al. ICML11] or even randomly selected patches.
  - Using $\mathbf{E}$ is equivalent to using $\mathbf{D} = RP$ assuming $\mathbf{\Lambda} = \mathbf{I}$.
  - Whitening makes $\mathbf{\Lambda}$ more diagonal.
- We can bound the error in accuracy as a function of dictionary size.
  - The bound on $\mathbf{K}'$ is in the same form as that on $\mathbf{C}'$.
  - Overall classification accuracy is (approximately) proportional to $\|\mathbf{K} - \mathbf{K}'\|_F$.
- On various datasets, larger codebook sizes exhibit diminishing returns, with our method giving a good estimation of the accuracy.



STL TIMIT WSJ

6. Pooling-invariant Dict Le

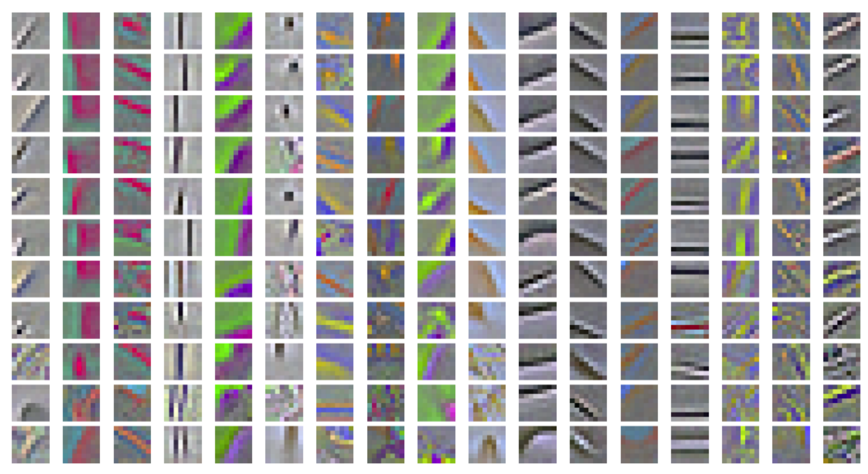
- Image Feature extraction almost always involve more than encoding.
- Conventional unsupervised methods focus on patch-based dictionary learning [Coates et al. ICML11], but pooling adds complications to the statistics of obtained features:



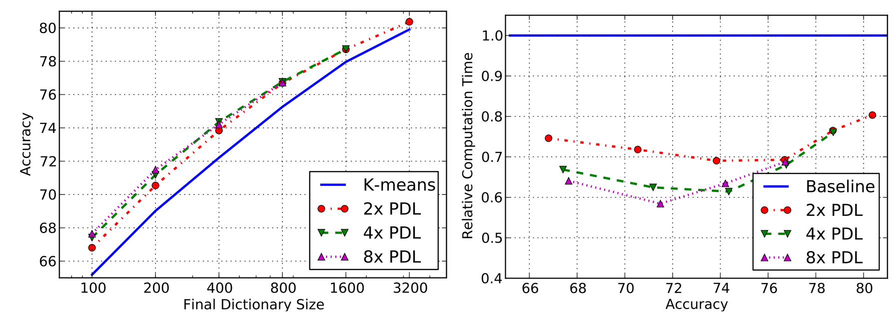
- The Nyström sampling view suggests efficient ways to learn pooling-invariant dictionaries.
- We used a two-stage clustering algorithm to learn such a dictionary:
  1. an overshooting dictionary with patch-based K-means;
  2. reducing the dictionary with affinity propagation (using covariance of pooled outputs as similarities).

6. Results

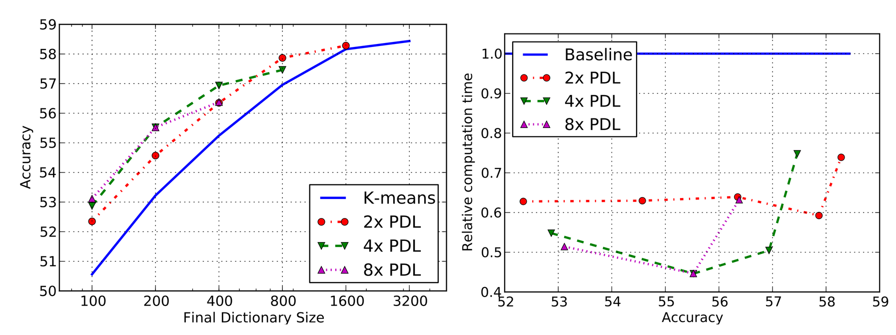
- Learned codes (first row) and pruned codes (codes below):



- Accuracy gain under fixed codebook sizes (left) and speedup under fixed accuracies (right):



CIFAR



STL

- (Note that the method is purely unsupervised.)

7. References

- O. Vinyals, Y. Jia, T. Darrell. *Why Size Matters: Feature Coding as Nystrom Sampling*. ICLR 2013.
- A. Coates, A. Ng. *The Importance of Encoding versus Training with Sparse Coding and Vector Quantization*. ICML 2011.
- O. Vinyals, L. Deng. *Are Sparse Representations Rich Enough for Acoustic Modeling?*. Interspeech 2012