Sparse Output Coding for Large-Scale Visual Recognition

Bin Zhao Eric P. Xing
School of Computer Science, Carnegie Mellon University
{binzhao, epxing}@cs.cmu.edu

Abstract

Many vision tasks require a multi-class classifier to discriminate multiple categories, on the order of hundreds or thousands. In this paper, we propose sparse output coding, a principled way for large-scale multi-class classification, by turning high-cardinality multi-class categorization into a bit-by-bit decoding problem. Specifically, sparse output coding is composed of two steps: efficient coding matrix learning with scalability to thousands of classes, and probabilistic decoding. Empirical results on object recognition and scene classification demonstrate the effectiveness of our proposed approach.

1. Introduction

A recent trend in visual recognition is the rapid increase of concept space. For example, the SUN [26] database for scene recognition has 899 categories, and ImageNet [10] spans a total number of 21841 image classes. Moreover, the number of categories can even grow further in the future. With such massive number of classes, it is no surprise that classical algorithms such as one-vs-rest, one-vs-one, or kNN, often favored for their simplicity [22, 3], will be brought to their knees not only because of the training time and storage cost they incur [10], but also because of the conceptual awkwardness of such algorithms in massive multi-class paradigms. For example, facing, say, one million classes, should we go ahead and build 1 million classifiers each trained on 1-vs-999999 classes? Just imagine the resultant data imbalance issue at its extreme, let alone the terrible irregularities of the decision boundaries of such classifiers. Clearly, large-scale visual recognition requires new, out-of-box rethinking of classical approaches and more effective yet simple alternatives.

Our goal in this work is to design a multi-class classification method that is both accurate and fast when facing a large number of categories. Specifically, we propose sparse output coding (SpOC), which turns the original large-scale K-class classification into an L-bit code construction problem, where $L = O(\log(K))$ and each bit can be constructed in parallel through a binary off-the-shelf classifier; followed by a decoding scheme to extract the class label.

1.1. Previous Work

The following two lines of research are related to our work.

**Large-scale visual recognition:** Very recently, we have seen successful attempts in large-scale visual recognition [19, 18, 2, 11]. Both [19] and [18] focus on designing high-dimensional feature representation for images, where classifier is trained using conventional one-vs-rest approach. SpOC serves as an important complement to this line of research, in the sense that we could very easily combine our classification method with feature representations learned in [19, 18] to yield even better results. On the other hand, [2, 11] learn tree classifiers, where multiple classifiers are organized in a tree and a test image traverses the tree from root to leaf to obtain its class label. However, tree structured classifiers face the well-known error propagation problem, where errors made close to the root node are propagated through the tree and yield misclassification. On the other hand, SpOC is robust to errors in local classifiers, as a result of the error correcting property of output coding.

**Error Correcting Output Coding:** For a K class problem, error correcting output coding (ECOC) [1] consists of two stages: coding and decoding. An output code B is a matrix of size $K \times L$ over $\{-1,0,+1\}$ where each row of B corresponds to a class $y \in \mathcal{Y} = \{1, \ldots, K\}$. Each column $\beta_i$ of B defines a partition of $\mathcal{Y}$ into three disjoint sets: positive partition (+1 in $\beta_i$), negative partition (-1 in $\beta_i$), and ignored classes (0 in $\beta_i$). Binary learning algorithms are used to construct bit predictor $h_l$ using training data $Z_l = \{(x_1, B_{y_1,l}), \ldots, (x_m, B_{y_m,l})\}$ with $B_{y_l,l} \neq 0$, for $l = 1, \ldots, L$ (throughout the rest of this paper, we use “bit predictor” to denote the binary classifier associated with a column of the coding matrix). Results in [1] suggest that learning a coding matrix in a problem-dependent way is better than using a pre-defined one. However, strong error-correcting ability alone does not guarantee good classification [8], since the performance of output coding is also highly dependent on the accuracy of the individual bit predictors. Consequently, several approaches [23, 8, 13] op-
timizing coding matrix and bit predictors simultaneously have been proposed. However, the coupling of learning coding matrix and bit predictors in a unified optimization framework is both a blessing and a curse. On the one hand, it could directly assess the accuracy of each bit predictor and hence pick the coding matrix that avoids difficult bit prediction problems; on the other hand, simultaneous optimization often results in expensive computation, hindering these approaches from being applied to large-scale multi-class problems. Therefore, SpOC learns coding matrix and bit predictors separately, but still balances error-correcting ability and bit prediction accuracy in learning the coding matrix. Therefore, SpOC is computationally very efficient compared with [23, 8, 13], without sacrificing accuracy.

Given a test instance \( x \), the decoding procedure finds the class \( y \) whose codeword in \( B \) is “closest” to \( h(x) = (h_1(x), \ldots, h_L(x)) \). For binary output coding scenario, where \( B \in \{-1, +1\}^{K \times L} \), either Hamming distance or Euclidean distance could be adopted to measure distance between two codewords. However, in the ternary case, where \( B \in \{-1, 0, +1\}^{K \times L} \), the special 0 symbol indicating ignored classes could raise problems. Specifically, previous attempts in decoding ternary codes [12] either (1) treat “0” bits the same way as non-zero bits, or (2) ignore those “0” bits entirely and only use non-zero bits for decoding. However, neither of the above approaches would prove sufficient. Specifically, treating “0” bits the same way as non-zero ones would introduce bias in decoding, since the distance increases with the number of positions that contain the zero symbol. On the other hand, ignoring “0” bits entirely would discard great amount of information. Probabilistic decoding utilizes zero bits by propagating labels from non-zero bits to zero ones subject to smoothness constraints, and proves effective especially on large scale problems.

1.2. Summary of Contributions

To conclude the introduction, we summarize our main contributions as follows. (1) We propose an approach for large-scale visual recognition, with scalability to hundred or thousand class problems. SpOC is robust to errors in bit predictors, simple to parallel, and its computational time scales logarithmically with the number of classes. (2) We propose a new optimization technique: dual proximal gradient method, for efficiently solving \( l_1 \) regularized optimization with complicated constraints. (3) We propose probabilistic decoding to effectively utilize semantic similarity between visual categories for accurate decoding.

2. Coding

For the sake of scalability to large-scale problems, SpOC decouples the learning processes of coding matrix and bit predictors. However, our proposed approach still balances the error-correcting ability of the coding matrix and classification accuracy for each resulting bit predictor, by utilizing training data and structure information among classes.

2.1. Formulation

As its most attractive advantage, the coding matrix in output coding is usually chosen for strong error-correcting ability. Besides, since output coding is essentially aggregating discriminative information residing in each bit, learning accurate bit predictors is also crucial for its success. Since the high computational cost associated with methods optimizing coding matrix and bit predictors simultaneously [13] renders them unfavorable in large-scale problems, we propose to use training data and class taxonomy, to provide a measure of separability for each binary partition. Specifically, if some classes are closely related but are given different codes in the \( l \)-th bit, the bit predictor \( h_r \) may not be easily learnable. However, a binary partition is more likely to be well solved if the intra-partition similarity is large while the inter-partition similarity is small. Finally, for large-scale multi-class problems, it is crucial to introduce ignored classes, i.e., 0 in the coding matrix. Otherwise, every single class will participate in the training of each bit predictor. As an illustrating example, consider the ImageNet data set with 20\( K \) + classes. With each of the 20\( K \) classes participating in learning a bit predictor, we will likely be facing a binary partition problem where both the positive and negative partitions are populated with data points coming from thousands of different classes. Clearly, learning bit predictor for such binary partition will be extremely difficult, due to the huge intra-partition dissimilarity. Therefore, sparse output coding learns optimal coding matrix as follows,\n
\[
\max_B F_b(B) - \lambda_r F_r(B) - \lambda_2 \sum_{l=1}^{L} \|\beta_l\|_2^2 \tag{1}
\]

\[
s.t. \quad B \in \{-1, 0, +1\}^{K \times L} \tag{2}
\]

\[
\sum_{k=1}^{K} \mathcal{I}(b_{kl} = 1) \geq 1, \quad \sum_{k=1}^{K} \mathcal{I}(b_{kl} = -1) \geq 1, \quad \forall r = 1, \ldots, L \tag{3}
\]

\[
\sum_{l=1}^{L} \mathcal{I}(b_{kl} \neq 0) \geq 1, \quad \forall k = 1, \ldots, K \tag{4}
\]

where \( \mathcal{I} \) is the indicator function. \( F_b(B) \) measures the separability of each binary partition problem associated with columns of \( B \), and reflects the expected accuracy of bit predictors. Moreover, \( F_r(B) \) measures codeword correlation, and minimizing \( F_r(B) \) ensures the strong error-correcting ability of the resulting coding matrix. The \( l_2 \) regularization on each column \( \beta_l \) of \( B \) controls the complexity of each bit prediction problem. \( \lambda_r \) and \( \lambda_2 \) are regularization parameters controlling the relative importance of the 3 competing objectives. Constraints in (2) ensure that each column of the coding matrix defines a binary partition problem, with the freedom of introducing ignored classes. Constraints in (3) ensure that each bit prediction problem has non-empty positive partition and non-empty negative partition. Finally,
constraints in (4) ensure that each class in the original multiclass classification appears in at least one bit prediction problem, so that we could effectively decode this class.

Before presenting details of each part in (1), we would like to make clear that the goal and contribution of this work is effective multi-class categorization with large number of classes, where complex methods would fail, and simplicity prevails. As a result, motivation for design of each piece in (1) is balance between effectiveness and efficiency. Although there might be more sophisticated formulations for pieces in (1), they will very likely increase computational cost, ultimately rendering the method incapable of handling large-scale multi-class problems.

\[ \text{2.1.1 } F_b(B): \text{Separability of Binary Problem} \]

One key issue in designing the coding matrix is to ensure that the resulting bit prediction problems could be effectively solved. The key motivation of our mathematical formulation is to compute the following two measures using semantic relatedness matrix \( S \) (for example, rose is more similar to tulip than truck) for each binary partition problem: intra-partition similarity, and inter-partition similarity.

Specifically, in each binary partition problem, both positive partition and negative partition are composed of data points from multiple classes in the original problem. To encourage better separation, those classes composing the positive partition should be similar to each other. The similar argument goes for those classes composing the negative partition, but they should be different from the former set which composes the positive partition. Specifically, for the \( l \)-th binary partition defined by \( \beta_l \), its separability is measured as

\[
F_b(\beta_l) = \sum_{k=1}^{K} \sum_{k'=1}^{K} \left( I(B_{kl}B_{k'l} > 0)S_{kk'} - I(B_{kl}B_{k'l} < 0)S_{kk'} \right)
\]

(5)

where \( k \in \mathbb{R}^K \) is the all-one vector and \( \circ \) is element-wise product. \( F_b(\beta_l) \) defined above should subtract \( \sum_k S_{kk} \).

However, since this quantity is constant and will not affect optimization of \( B \), we omit this step. Finally,\n
\[
F_b(B) = \sum_{l=1}^{L} \sum_{l=1}^{L} \left( e_K^T \left[ \beta_l \beta_l^T \circ S \right] e_K \right)
\]

(6)

where \( \beta_l \in \mathbb{R}^K \) is the all-one vector and \( \circ \) is element-wise product.

\[ \text{Semantic Relatedness Matrix: } S \] measures similarity between classes using training data and class taxonomy. Let \( \mathcal{X}_i = \{X_1^{(i)}, \ldots, X_{r(i)}^{(i)}\} \) and \( \mathcal{X}_j = \{X_1^{(j)}, \ldots, X_{r(j)}^{(j)}\} \) be two classes from the multi-class categorization problem. Several approaches have been proposed to measure class similarity / distance, such as Hausdorff distance, match kernel \([14, 20]\), divergence between probability distributions \([21]\). Here we adopt the sum match kernel \([14]\), and define data similarity \( S^D \) between classes \( i \) and \( j \) as

\[
S^D_{ij} = \frac{1}{|X_i||X_j|} \sum_{p=1}^{|X_i|} \sum_{q=1}^{|X_j|} K_D(x_p^{(i)}, x_q^{(j)})
\]

where \( K_D \) is a Mercer kernel. Moreover, classes in large-scale multi-class categorization are rarely organized in a flat fashion, but rather with a taxonomical structure \([10, 6]\), such as a tree. Besides, algorithms for learning structure among classes have also been studied \([2]\), although this is beyond the scope of this work. Following \([5]\) we define structural affinity \( A_{ij} \) between class \( i \) and class \( j \) as the number of nodes shared by their two parent branches, divided by the length of the longest of the two branches \( A_{ij} = \text{intersect}(P_i, P_j)/\max(\text{length}(P_i), \text{length}(P_j)) \), where \( P_i \) is the path from root node to node \( i \) and \( P_j \) counts nodes shared by two paths \( P_i \) and \( P_j \). We then construct structural similarity matrix \( S^S = \exp(-\kappa(E - A)) \), where \( \kappa \) is a constant controlling the decay factor, and \( E \in \mathbb{R}^{K \times K} \) is all-one matrix. Finally, semantic relatedness matrix \( S \) is the weighted sum \( S = \alpha S^D + (1 - \alpha)S^S \) with \( \alpha \in [0, 1] \) being the weight. In problems without class taxonomy, we simply set \( \alpha = 1 \).

\[ \text{2.1.2 } F_c(B): \text{Codeword Correlation} \]

Given an example \( x \), the \( L \)-dimensional bit predictor \( h(x) = [h_1(x), \ldots, h_L(x)] \) is computed. We then predict its label \( y \) based on which row in \( B \) is “closest” to \( h(x) \). To increase tolerance of errors occurred in bit predictions, a crucial design objective of the coding matrix is to ensure that the rows in \( B \) are separated as far from each other as possible. Hence, we propose to maximize the distance between rows in \( B \). Equivalently, we could minimize the inner products of the corresponding vectors. Thus, codeword correlation of \( B \) could be computed as following:

\[
F_c(B) = \sum_{k=1}^{K} \sum_{k'=1}^{K} r^T_k r_{k'} = e_K^T (BB^T) e_K
\]

(7)

where \( r_1, \ldots, r_K \) are row vectors of coding matrix \( B \).

\[ \text{2.1.3 Relaxing the Integer Constraints} \]

In problem (1), the integer constraint (2) on \( B \) makes the problem NP-hard to solve. To enable efficient solution, we follow \([8]\) and relax it to \( B \in [-1, +1]^{K \times L} \). To introduce ignored classes in bit predictors, we further regularize \( l_1 \) norm of \( B \). Putting everything together, we get

\[
\min_B -tr(BB^T S) + \lambda_2 e_K^T BB e_K + \lambda_1 \sum_{l=1}^{L} ||B_l||_2^2 + \lambda_1 ||B||_1
\]

s.t. \(-1 \leq B_{kl} \leq 1, \forall k = 1, \ldots, K; \forall l = 1, \ldots, L \) \hspace{1cm} (8)

\[
\frac{K}{k=1} ||B_{kl} + B_{kl}||_2 \geq 2 \sum_{l=1}^{L} ||B_{kl} - B_{kl}||_2, \forall l = 1, \ldots, L \hspace{1cm} (9)
\]

\[
\frac{L}{l=1} ||B_{kl}||_1 \geq 1, \forall k = 1, \ldots, K \hspace{1cm} (10)
\]
where \(\|B\|_1 = \sum_{k,l} |B_{kl}|\), and we equivalently reformulate constraints (3) and (4) to (10) and (11).

2.2. Optimization

The difficulty of solving problem (8) lies in two facts: non-smoothness of \(l_1\) regularization on \(B\), and non-convexity of objective and constraints. However, problem (8) has the special structure that the objective is the difference of two convex functions. Specifically, both \(f(B) = \lambda_r e_k^\top B B^\top e_k + v \lambda_r \sum_{i=1}^L \|\beta_i\|^2 + \lambda_1 \|B\|_1\) and \(g(B) = tr(\hat{B} B^\top S)\) are convex. Constraints (10) and (11) could also be formulated similarly. Therefore, we propose a concave-convex procedure based algorithm, where the non-convexity is handled by constrained concave-convex procedure (CCCP) [24, 7], and the non-smoothness is handled using dual proximal gradient method.

2.2.1 Constrained Concave-Convex Procedure

Given an initial point \(B^0\), the CCCP computes \(B^{t+1}\) from \(B^t\) by replacing \(g(B)\) with its first-order Taylor expansion at \(B^t\), i.e., \(g(B^t) + < \nabla g(B^t), B - B^t >\). For non-smooth functions, it has been shown that the gradient should then be replaced by sub-gradient [7, 28]. Hence, the \(|B_{kl}|\) term appearing in constraints could be replaced with its first-order Taylor expansion at \(B^t\), i.e., \(\text{sign}(B^t_{kl})B_{kl}\). The resulting optimization problem is as follows,

\[
\min_{B} -2\frac{\text{tr}(SB^\top B^\top)}{\gamma} + \lambda_r e_k^\top B B^\top e_k + \lambda_1 \sum_{l=1}^L \|\beta_l\|^2 + \lambda_1 \|B\|_1 \quad (12)
\]

\[
\text{s.t.} \quad -1 \leq B_{kl} \leq 1, \quad \forall k = 1, \ldots, K; \forall l = 1, \ldots, L
\]

\[
2 + \sum_{k=1}^K \sum_{l=1}^L [\text{sign}(B^t_{kl})]B_{kl} \leq 0, \quad \forall l = 1, \ldots, L
\]

\[
2 + \sum_{k=1}^K \sum_{l=1}^L [\text{sign}(B^t_{kl})]B_{kl} \leq 0, \quad \forall l = 1, \ldots, L
\]

\[
1 - \sum_{l=1}^L \text{sign}(B^t_{kl})B_{kl} \leq 0, \quad \forall k = 1, \ldots, K
\]

Algorithm 1 Learning output coding matrix

1. Initialize \(B^0\)
2. repeat
   1. Find \(B^t\) as the solution to problem (12);
   2. Set \(t = t + 1\) and get the new problem (12)
3. until stopping criterion satisfied

2.2.2 Dual Proximal Gradient Method

Denote the objective function in problem (12) as \(F(B)\). Although \(F(B)\) is convex, it is a non-smooth function of \(B\) due to the \(l_1\) regularization imposed on \(B\), making the problem difficult to solve. Traditional algorithms for non-smooth convex optimization include smoothing techniques where the non-smooth term is approximated by a smooth function, and sub-gradient method. However, smoothing technique will lose the sparsity inducing property of the \(l_1\) regularization, and sub-gradient method is known for slow convergence and difficulty in picking step size. On the other hand, proximal gradient method [15] has been the major workhorse for solving un-constrained non-smooth convex optimization problems, due to its fast convergence and low complexity. Unfortunately, the constraints in (12) renders the problem unsuitable for proximal gradient methods, as we cannot easily project the solution to constraints in problem (12). Therefore, to solve problem (12), we first get its dual, then apply proximal gradient method on the obtained dual, as the constraints in the dual problem are much easier for projection. Specifically, define \(\beta = \text{vec}(B) \in \mathbb{R}^{KL}\) as the vector obtained by stacking columns of \(B\). Problem (12) could be equivalently formulated as

\[
\min_{\beta} F_s(\beta) + \lambda_1 \|\beta\|_1, \quad \text{s.t.} \quad A\beta \leq b
\]

with \(F_s(\beta) = -2\frac{\text{tr}(SB^\top B^\top)}{\gamma} + \lambda_r e_k^\top B B^\top e_k + \lambda_1 \sum_{l=1}^L \|\beta_l\|^2\) and \(b \in \mathbb{R}^{2KL + 2L + K}\). Therefore, to solve problem (12), we first get its dual, then apply proximal gradient method on the obtained dual, as the constraints in the dual problem are much easier for projection. Specifically, define \(\beta = \text{vec}(B) \in \mathbb{R}^{KL}\) as the vector obtained by stacking columns of \(B\). Problem (12) could be equivalently formulated as

\[
\min_{\beta} \frac{\partial h(\gamma, \mu)}{\partial \gamma} = -A\nabla F_s^* (-(A^\top \gamma - \mu)) + b
\]

\[
\min_{\beta} \frac{\partial h(\gamma, \mu)}{\partial \mu} = -\nabla F_s^* (-(A^\top \gamma - \mu))
\]

where \(\nabla F_s^* (-(A^\top \gamma - \mu)) = \arg\min_{\beta} [F_s(\beta) + (A^\top \gamma - \mu)^\top \beta]\). Moreover, we could reformulate \(F_s(\beta)\) as follows:

\[
F_s(\beta) = \sum_{l=1}^L \left\{ -2[SB^\top B^\top]_{l}^\top \beta_l + \lambda_1 \beta_l e_k^\top B e_k + \lambda_1 \beta_l \right\}
\]
where $(SB^t)_l$ denotes the $l$-th column of matrix $SB^t$. Therefore, \( \hat{\beta} = \nabla F^{-1} \left( - (A(\gamma - \mu)) \right) \) could be calculated as \( \hat{\beta} = [\hat{\beta}_1 \ldots \hat{\beta}_l]^\top \), with
\[
\hat{\beta}_l = \frac{1}{2}(\lambda_e e_k^\top + \lambda_e I)^{-1}[2(SB^t)_l - (A(\gamma - \mu))]
\] (25)
where $(A(\gamma - \mu))_l$ is the $l$-th column of the matrix formulated by resizing $A(\gamma - \mu)$ into $K \times L$ matrix. Finally, the projected gradient algorithm for solving problem (12) is shown in Algorithm 2, where $P$ represents projection onto the corresponding constraints.

**Algorithm 2** Dual proximal gradient for problem (12)

Choose step size $t > 0$, choose initial $\gamma$ and $\mu$
repeat
Compute $\hat{\beta}$ using Eq. (25)
\[ \gamma = P_{t > 0}(\gamma - t(b - A\hat{\beta})); \mu = P_{\|\mu\| \leq \lambda_e}(\mu - t\hat{\beta}) \]
until convergence

**3. Probabilistic Decoding**

For large-scale multi-class categorization, a sparse output coding matrix is necessary to ensure the learnability of each bit predictor. However, the zero bits in coding matrix also bring difficulty in decoding. For example, consider a 5-class problem in Figure 1. Given a test image from class Husky, if we treat zero bits the same way as non-zero ones, Hamming decoding would prefer Shepherd over Husky. However, Husky is only worse than Shepherd as its codeword has more zero. This effect occurs because the decoding value increases with the number of positions that contain the zero symbol and hence introduces bias. On the other hand, ignoring zero bits entirely would discard precious information for decoding. This is especially important when $K$ is large, where we expect a very sparse coding matrix. For example, in Figure 1, classes Husky and Tiger have only two non-zero bits in their codewords. Since we cannot always have perfect bit predictors, classification errors on bit 1 and bit 4 would severely impair the overall accuracy.

Fortunately, the semantic class similarity $S$ computed using training data and class taxonomy, provides venue for effectively propagating information from non-zero bits to zero ones. For the example in Figure 1, class Husky is more similar to (Shepherd, Wolf) than (Fox, Tiger). The second bit predictor in Figure 1 solves a binary partition of (Shepherd, Wolf) against Fox. Even though class Husky is ignored in training for this bit, the binary partition on images from this class will have a higher probability of being +1, due to the fact that the two positive classes in this binary problem are closely related to class Husky. Therefore, those classes with non-zero bits in the coding matrix, should effectively propagate their label to those initially ignored classes. In this section, we propose probabilistic decoding, to effectively utilize semantic class similarity for better decoding. Specifically, we treat each bit prediction (without loss of generality, say, the $l$-th bit) as a label propagation [29] problem, where the labeled data corresponds to those classes whose codeword’s $l$-th bit is non-zero, and unlabeled data corresponds to those whose $l$-th bit is zero. The goal of label propagation is to define a prior distribution indicating the probability of one class being classified as positive in the $l$-th binary partition. Combining this prior with the available training data, we formulate the decoding problem in ternary ECOC as maximum a posteriori estimation.

**3.1. Formulation**

Given output coding matrix $B \in \{-1, 0, +1\}^{K \times L}$, our decoding method estimates conditional probability of each class $k$ given input $x$ and $L$ bit predictors $\{h_1, \ldots, h_L\}$. Without loss of generality, we assume the bit predictors constructed in the coding stage are linear classifiers, each parameterized by a vector $w$ as $h_l(x) = \text{sign}(w_l^\top x)$. Define $\{c_1, \ldots, c_L\} \in \{-1, +1\}^L$ as a random vector of binary values, representing one possible codeword for instance $x$.

The decoding problem is then to find the class $k$, which maximizes the following conditional probability:
\[
\mathbb{P}(y = k | \{w_l\}, x, \mu) = \sum_{\{c_l\}} \mathbb{P}(y = k | \{c_l\}, x, \mu, \{c_l\}) \cdot \mathbb{P}(\{c_l\} | \{w_l\}, x, \mu)
\]
\[
= \sum_{\{c_l\}} \prod_{l} \mathbb{P}(c_l | k, \mu_k) \prod_l \mathbb{P}(c_l | w_l, x)
\]
\[
= \prod_l \left( \sum_{c_l} \mathbb{P}(c_l = 1 | w_l, x) + (1 - \mu_k) \mathbb{P}(c_l = 0 | w_l, x) \right)
\]
\[
= \prod_l \left( \mu_k \mathbb{P}(c_l = 1 | w_l, x) + (1 - \mu_k) (1 - \mathbb{P}(c_l = 1 | w_l, x)) \right)
\] (26)

where $\{c_l\} = \{c_1, \ldots, c_L\}, \{w_l\} = \{w_1, \ldots, w_L\}$, and $
\mu_k \in [0, 1]$ is the parameter in Bernoulli distribution $\mathbb{P}(c_l = 1 | y = k) = \mu_k$. Moreover, given the learned bit predictors, $\mathbb{P}(c_l = 1 | w_l, x)$ could be computed using a logistic link function as $\mathbb{P}(c_l = 1 | w_l, x) = 1/(1 + \exp(-w_l^\top x))$.

Therefore, we need to learn the Bernoulli parameters $\{\mu_k\}$,
which measures the probability of the $l$-th bit being +1 given the true class as $y = k$. Specifically, for the $l$-th column of the coding matrix, those classes corresponding to $+1$ in the $l$-th bit, i.e., $B_{kl} = 1$, will have $\mu_{kl} = 1$, and similarly those classes corresponding to $-1$, i.e., $B_{kl} = -1$, will have $\mu_{kl} = 0$. However, originally ignored classes (those corresponding to 0 in the coding matrix) will also be likely to have a preference on the value of the $l$-th bit. For the example in Figure 1, the second bit predictor separates (Shepherd, Wolff) from Fox. Clearly, $P(c_l = 1|\text{Shepherd}) = P(c_l = 1|\text{Wolf}) = 0$ and $P(c_l = 1|\text{Fox}) = 1$. Since class Husky is not directly involved in this binary classification problem, a non-informative prior would put $P(c_l = 1|\text{Husky}) = 0.5$. However, if the true class for an instance $x$ is Husky, this bit clearly has a much higher probability of being $-1$ than $+1$, due to the fact that Husky is much closer to Shepherd and Wolf semantically, than Fox. Therefore, we should have $P(c_l = 1|\text{Husky}) < 0.5$, and in such way those classes with non-zero values in the $l$-th bit effectively propagate their labels to those initially ignored classes.

### 3.1.1 Prior Distribution via Label Propagation

Since each column $\beta_i$ in coding matrix has its own labeling (different composition of positive classes, negative classes, and ignored classes), label propagation for each column is independent of others. For $l$-th column $\beta_l$, consider a connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $K$ nodes $\mathcal{V} = \mathcal{L} \cup \mathcal{U}$ corresponding to the $K$ classes, where $\mathcal{L}$ corresponds to the labeled classes, and $\mathcal{U}$ corresponds to ignored classes. Our task is then to assign probabilities of being labeled as positive to nodes in $\mathcal{U}$. Define $\mu_l = (\mu_{l1}, \ldots, \mu_{lK})$ as labels on nodes $\mathcal{V}$, where $\mu_{kl} = 1$ for those classes labeled as $+1$ and $\mu_{kl} = 0$ for those classes labeled as $-1$ in $\beta_l$. Consequently, for any node $k \in \mathcal{V}$, $\mu_{kl} = P(c_l = 1|y = k)$. Intuitively, we want unlabeled nodes that are nearby in the graph to have similar labels, and this motivates the choice of the following quadratic energy function [29]:

$$E(\mu_l) = \frac{1}{2} \sum_{i,j} S_{ij}(\mu_{il} - \mu_{jl})^2,$$

where $S$ is the semantic similarity matrix. To assign probability distribution on $\mu_l$, we form Gaussian field $p_C(\mu_l) = \frac{1}{Z_c} \exp(-CE(\mu_l))$ [29], where $C$ is inverse temperature parameter, and $Z_C = \int_{\mu_l} \exp(-CE(\mu_l))d\mu_l$ is a normalizing constant. Define diagonal degree matrix $D$ as $D_{ll} = \sum_j S_{lj}$ and graph Laplacian $\Delta = D - S$, the Gaussian field defined on $\mu_l$ could be equivalently formulated as $p_C(\mu_l) = \frac{1}{Z_c} \exp(-C\mu_l^\top \Delta \mu_l)$, with $\mu_{kl}$ clamped to 1 on positive classes and 0 on negative classes.

### 3.1.2 Parameter Learning

Given $L$ bit predictors, and training data $Z = (X, Y) = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, we have that

$$\log P(Y|\{w_i\}, X, \mu) = \frac{1}{L} \sum_{i=1}^L \sum_{l=1}^L \log \{\mu_{yl}P_l + (1 - \mu_{yl})(1 - P_l)\}$$

where $P_l = P(c_l = 1|w_i, x_i)$. Combining data likelihood with prior distribution, we get the following optimization problem for learning parameters $\mu$ using MAP estimation

$$\min_{\mu} -\sum_{l=1}^L \sum_{i=1}^m \log \{\mu_{yl}P_l + (1 - \mu_{yl})(1 - P_l)\} + C \sum_{l=1}^L \mu_l^\top \Delta \mu_l$$

s.t. $0 \leq \mu_{kl} \leq 1$, $k = 1, \ldots, K$, $l = 1, \ldots, L$

$$\mu_{kl} = 1$$

$$\mu_{kl} = 0$$

where $\mu = [\mu_1, \ldots, \mu_L]$. Clearly, $\mu_l$ in the above optimization problem is independent of each other, and could therefore be optimized separately. We use projected gradient descent to solve the above optimization problem.

### 3.2 Decoding

Given the learned Bernoulli parameters $\mu$, the inference problem targets to find the label $k^*$ that maximizes the conditional probability: $k^* = \arg \max_k P(y = k|x_1, \ldots, x_N)$, $\mu$. Clearly, given $\mu$, decoding takes linear time with the number of columns in coding matrix, which could be as small as $L = \mathcal{O}(\log K)$. Hence, our proposed probabilistic decoding is very efficient, and promising for large-scale multi-class categorization.

### 4. Experiments

In this section, we test the performance of sparse output coding on two data sets: ImageNet [10] for object recognition, and SUN database [26] for scene recognition.

#### Object recognition on ImageNet

We use two subtrees in ImageNet, with the root node being “flower” and “food”, respectively. The flower image collection contains a total of 0.34 million images covering 462 categories, and the food data set contains a total of 0.93 million images covering 1308 categories. For both data sets, we randomly pick 50% of images from each class as training data, and test on the remaining 50% images.

#### Scene recognition on SUN database

The SUN database is by far the largest scene recognition data set, with 899 scene categories. We use 397 well-sampled categories to run the experiment [26]. For each class, 50 images are used for training and the other 50 for test.

#### Feature representations

For each data set, we start with computing dense SIFT descriptors for each image, and then run $k$-means clustering on a random subset of 1 million SIFT descriptors to form a visual vocabulary of 8192 visual words. Using the learned vocabulary, we employ Locality-constrained Linear Coding (LLC) [25] for feature coding. Finally, a single feature vector is computed for each image using max pooling on a spatial pyramid [17].

#### 4.1. Experiment Design and Evaluation

We compare SpOC against one-vs-rest (OVR), one of the most widely applied frameworks for large-scale visual recognition. Since class taxonomy exists for all data sets,
we also test two hierarchical classifiers. The first hierarchical classifier (HieSVM-1) follows a top-down approach, and trains a multi-class SVM at each node in the class taxonomy [16]. The second one (HieSVM-2) adopts the strategy in [9]. Moreover, we also compare with several output coding based methods. Specifically, we compare with the random dense code output coding (RDOC) proposed in [1], where each element in the coding matrix is chosen at random from \{-1, 1\}, with probability 1/2 for -1 and +1 each. Also, we provide results for the random sparse code output coding (RSOC) in [1], where each element in the coding matrix is chosen at random from \{-1, 0, +1\}, with probability 1/2 for 0, and probability 1/4 for -1 and +1 each. Furthermore, we also report results of the algorithm in [27], which builds dense output codes using spectral decomposition (SpecOC) of the graph Laplacian constructed using the class taxonomy. Finally, to test the impact of probabilistic decoding on SpOC, we report results of SpOC using a simple Hamming distance based decoding strategy, denoted as SpOC-H.

For all algorithms, we train linear SVM using stochastic gradient descent [4]. For output coding based methods, we set code length \(L = 200\) for flower and SUN, and \(L = 300\) for food. For SpOC and SpOC-H, we simply set \(\lambda_c = 1\), \(\lambda_1 = K\) and \(\kappa = 5\) for all data sets. Data similarity matrix \(S^D\) is pre-computed with linear kernel and \(\alpha = 0.5\). For RDOC and RSOC, 1000 random coding matrices are generated for each scenario and the one with the largest minimum pair-wise Hamming distances between all pairs of codewords and does not have any identical columns is chosen. To decode the label for OVR using learned binary classifiers, we pick the class with the largest decision value. For RDOC, RSOC, SpecOC and SpOC-H, we pick the class whose codeword has minimum Hamming distance with the codeword of test data point. Specifically, for decoding in RSOC and SpOC-H, we test both strategies of treating zero bits the same way as non-zero ones and ignoring zero bits entirely, and report the best result of these two methods.

For every data point, each algorithm will produce a list of 10 classes in the descending order of confidence (except HieSVM-1, which only provides the most confident class label), based on which the top-\(n\) accuracy is computed, \(n = 1, 5, 10\) in our case. Specifically, accuracy equals 1 if the true class is within the \(n\) most confident predictions, and 0 otherwise. The overall accuracy for each algorithm is the average over the entire test data set.

### 4.2. Results

Classification results for various algorithms are shown in Table 2, with the following observations: (1) SpOC systematically outperforms OVR. More interestingly, SpOC consists of much less binary classifiers than OVR. This shows that for large-scale visual recognition problems, output coding with a carefully designed coding matrix could outperform OVR, while maintaining cheaper computational cost, due to the error-correcting property introduced in the coding matrix. (2) Both SpOC and OVR beat RDOC and RSOC, revealing the importance of enforcing learnability of each bit predictor, since randomly generated coding matrix could very likely generate difficult binary separation problems. (3) SpOC performs better than SpecOC, which employs a dense coding matrix. The margin between SpOC and SpecOC is even more severe on food, revealing the importance of having ignored classes in each bit predictor. (4) SpOC and OVR both outperform HieSVM-1, where errors made in the higher level of the class hierarchy get propagated into the lower levels, with no mechanism to correct those early errors. On the other hand, the error-correcting property in SpOC introduces robustness to errors made in bit predictors. (5) SpOC-H generates inferior results than SpOC across the board, indicating the necessity of probabilistic decoding. Finally, HieSVM-2 runs into out of memory problems on all three data sets.

**Effect of code length:** To investigate the effect of code length on classification accuracy of SpOC, we test SpOC on the flower data set with various \(L\). According to Table 3, classification accuracy of SpOC improves as the code length increases, as stronger error-correcting ability is accompanied with longer codes. However, the fact that \(L = 200\) performs almost as well as \(L = 400\) demonstrates that SpOC usually requires much less bit predictors compared to the number of classes in the multi-class categorization problem.

**Time complexity:** We compare the time complexity of SpOC with OVR in Table 4. Specifically, computational time for SpOC consists three parts: (1) time for learning output coding matrix, (2) time for training bit predictors, and (3) time for probabilistic decoding. We implement SpOC using MATLAB 7.12 on a 3.40 GHZ Intel i7 PC with 16.0 GB main memory. Bit predictors or binary classifiers are trained in parallel on a cluster composed of 200 nodes. Time for training bit predictors is the summation of time spent on each node. According to Table 4, time for learning coding matrix and probabilistic decoding is almost negligible compared to the time spent on training bit predictors. Moreover, the total CPU time spent for training bit predic-

### Table 1. Data sets details.

<table>
<thead>
<tr>
<th>Data set</th>
<th>#Class</th>
<th>#Train</th>
<th>#Test</th>
<th>#Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flower</td>
<td>462</td>
<td>169691</td>
<td>169691</td>
<td>170006</td>
</tr>
<tr>
<td>Food</td>
<td>1308</td>
<td>467374</td>
<td>467374</td>
<td>170006</td>
</tr>
<tr>
<td>SUN</td>
<td>397</td>
<td>19850</td>
<td>19850</td>
<td>170006</td>
</tr>
</tbody>
</table>

Table 3. Classification accuracy of SpOC vs. code length.
Table 2. Classification accuracy (%) comparison.

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Flower (%)</th>
<th>Food (%)</th>
<th>SUN (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>OVR</td>
<td>27.23</td>
<td>60.05</td>
<td>72.57</td>
</tr>
<tr>
<td>Hier-SVM-1</td>
<td>23.81</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>RDOC</td>
<td>13.09</td>
<td>45.22</td>
<td>59.13</td>
</tr>
<tr>
<td>RSOC</td>
<td>12.88</td>
<td>46.31</td>
<td>60.96</td>
</tr>
<tr>
<td>SpecOC</td>
<td>21.37</td>
<td>52.25</td>
<td>64.09</td>
</tr>
<tr>
<td>SpOC</td>
<td>27.14</td>
<td>61.26</td>
<td>74.33</td>
</tr>
<tr>
<td>OVR-H</td>
<td>30.48</td>
<td>65.72</td>
<td>76.38</td>
</tr>
</tbody>
</table>

Table 4. Time complexity comparison. For SpOC, the time (seconds) before | is for learning coding matrix and decoding, and the time after | is for learning bit predictors. (1E6 = 1 x 10^6)

5. Conclusions

Sparse output coding provides an initial foray into large-scale visual recognition, by turning high-cardinality multi-class classification into a bit-by-bit decoding problem. We also propose probabilistic decoding to decode the optimal class label. Effectiveness of SpOC is demonstrated on object recognition and scene classification, with hundreds or thousands of classes. The fact that SpOC takes less bit predictors than OVR while achieving better accuracy, renders it especially promising for large-scale visual recognition.

Acknowledgements

This research is supported by Google, NSF IIS-0713379, and NSF DBI-0640543, ONR N000140910758 and AFOSR FA9550010247.

References