Probabilistic Label Relation Graphs with Ising Models

Nan Ding
Google Inc.
dingnan@google.com

Jia Deng
University of Michigan
jiadeng@umich.edu

Kevin P. Murphy
Google Inc.
kpmurphy@google.com

Hartmut Neven
Google Inc.
neven@google.com

Abstract

We consider classification problems in which the label space has structure. A common example is hierarchical label spaces, corresponding to the case where one label subsumes another (e.g., animal subsumes dog). But labels can also be mutually exclusive (e.g., dog vs cat) or unrelated (e.g., furry, carnivore). To jointly model hierarchy and exclusion relations, the notion of a HEX (hierarchy and exclusion) graph was introduced in [8]. This combined a conditional random field (CRF) with a deep neural network (DNN), resulting in state of the art results when applied to visual object classification problems where the training labels were drawn from different levels of the ImageNet hierarchy (e.g., an image might be labeled with the basic level category "dog", rather than the more specific label "husky"). In this paper, we extend the HEX model to allow for soft or probabilistic relations between labels, which is useful when there is uncertainty about the relationship between two labels (e.g., an antelope is "sort of" furry, but not to the same degree as a grizzly bear). We call our new model pHEX, for probabilistic HEX. We show that the pHEX graph can be converted to an Ising model, which allows us to use existing off-the-shelf inference methods (in contrast to the HEX method, which needed specialized inference algorithms). Experimental results show significant improvements in a number of large-scale visual object classification tasks, outperforming the previous HEX model.

1. Introduction

Classification is a fundamental problem in machine learning and computer vision. In this paper, we consider how to extend the standard approach to exploit structure in the label space. For example, consider the problem of classifying images of animals. The labels may be names of animal types (e.g., dog, puppy, cat), or attribute labels (e.g., yellow, furry, has-stripes). Many of these labels are not semantically independent of each other. For example, a puppy is also a dog, which is a hierarchical or subsumption relation; an animal cannot be both a dog and a cat, an exclusive relation; but an animal can be yellow and furry, which is a non-relation.

In [8], an approach called Hierarchy and EXclusion (HEX) graphs was proposed for compactly representing such constraints between the labels. In particular, a probabilistic graphical model with deterministic or hard constraints between the binary label nodes was proposed. These hard constraints cut down the feasible set of labels from 2^n (where n is the number of labels) to something much smaller, allowing for efficient exact inference. For example, if all labels are mutually exclusive, the HEX graph is a clique, and there are only $n + 1$ valid label configurations. This graphical model can be combined with any standard discriminative classifier (such as deep neural networks), resulting in a conditional random field (CRF) model with label constraints.

In this paper, we extend the HEX model by allowing for "soft" relationships between the labels. We call this the pHEX model. The pHEX model has five main advantages compared to the HEX model. First, it is a more realistic model, since the relationship between most labels is "soft". For example, a lion may be mostly yellow, but it could also be another color. Second, the pHEX model is easier to train, since the likelihood function is smoother. Third, we show how to perform inference in the pHEX model by converting it to an Ising model, and then using standard off-the-shelf tools such as belief propagation, or the emerging quantum optimization technology [11]. This is in contrast to the HEX case, which needed a specialized (and rather complex) algorithm to perform inference. Fourth, we show how to combine binary labels with k-ary labels, something that wasn't possible with the original HEX model. Finally, we show that the pHEX model outperforms the HEX model on a variety of visual object classification tasks.

2. Related work

There has been a lot of prior work on exploiting structure in the label space; we only have space to mention a few key papers here. Conditional random fields [12, 24] and structural SVMs [23] are often used in structured prediction problems. In addition, in transfer learning [20, 19], zero-shot learning [13, 17], and attribute-based recognition [1, 26, 21], consistency between visual predictions and semantic relations are often enforced.

More closely related to this paper is work that exploits hierarchical structure (e.g., [27, 14, 25, 16]), exclusive relations [5], or both of them [6, 15]. Recently [8] proposed the HEX graph approach, which subsumes a lot of prior work by modeling hierarhical and exclusive relations using graphical models. We discuss this in more detail in Section 3, since it forms the foundation for the current paper.

3. The HEX model

In a nutshell, HEX graphs are probabilistic graphical models with directed and undirected edges over a number of binary variables. Each binary variable represents a label and takes value from $\{-1, 1\}$. Each edge or no-edge between any two labels represents one of three label relations: exclusion, hierarchy and non-relation. The combination of all pairwise label relations allows the HEX graph to characterize the legal and illegal state space of labels, as we explain below.

3.1. HEX relations

The three types of label relations in the HEX graph are defined as follows:

Exclusion When two nodes are connected by an *undirected edge*, this is called an exclusive relation. It means that the two labels cannot be both equal to 1. For example, an animal cannot be both a *cat* and a *dog*. So *cat* and *dog* are mutually exclusive. The legal state space for exclusion is:

$$S^e \triangleq \{(-1,-1),(-1,1),(1,-1)\}. \tag{1}$$

Hierarchy When two nodes are connected by a *directed edge* from y_1 to y_2, this is called a subsumption (hierarchical) relation. It means that if y_2 is 1 then y_1 must be 1 as well. For example, a *puppy* is always a *dog*. So *dog* subsumes *puppy*. The corresponding legal state space for subsumption is:

$$S^h \triangleq \{(-1,-1),(1,-1),(1,1)\}. \tag{2}$$

No relation When two nodes are *not connected* by any edge, we say there is no relation between them. This means that the two labels are independent of each other. For example, *carnivore* and *yellow* are independent properties of animals. In this case, the legal state space for the two variables contains all 4 possible configurations:

$$S^o \triangleq \{(-1,-1),(-1,1),(1,-1),(1,1)\}. \tag{3}$$

3.2. HEX graph as a graphical model

To mathematically formulate the HEX model, assume we have a set of n possible labels, represented as the bit vector $\mathbf{y} = \{y_1, \ldots, y_n\}$, where $y_i \in \{-1, +1\}$. Also, assume we have an input feature vector $\mathbf{x} = \{x_1, \ldots, x_d\}$, and some discriminative model which maps this to the score vector $\mathbf{z} = \{z_1, \ldots, z_n\}$, where z_i is the "local evidence" for label y_i. (The mapping from $\mathbf{x}$ to $\mathbf{z}$ is arbitrary; in this paper, we assume it is represented by a deep neural network parameterized by $\mathbf{w}$, which we will denote by $\mathbf{z} = DNN(\mathbf{x}; \mathbf{w})$.) Given this, we can define the model as follows:

$$p(\mathbf{y} \mid \mathbf{x}) = \frac{1}{Z(\mathbf{x})} \prod_{i=1}^{n} \psi(y_i, z_i) \prod_{(i,j)\in G} \phi_a(y_i, y_j), \tag{4}$$

where $\psi(y_i, z_i) = 1/(1 + \exp(-2y_i z_i))$ is the logistic function, and $\phi(y_i, y_j)$ is the (edge-specific) potential function, defined below: (We use the notation ϕ_a to represent an "absolute" or deterministic potential, to distinguish it from the soft or probabilistic potentials we use later, denoted by ϕ_p.)

- Exclusion

$$\phi_a^e(y_1, y_2) = \begin{cases} 1 & (y_1, y_2) \in S^e \\ 0 & (y_1, y_2) = (1, 1); \end{cases} \tag{5}$$

- Hierarchy

$$\phi_a^h(y_1, y_2) = \begin{cases} 1 & (y_1, y_2) \in S^h \\ 0 & (y_1, y_2) = (-1, 1); \end{cases} \tag{6}$$

- No relation

$$\phi_a^o(y_1, y_2) = 1 \ \ \forall (y_1, y_2). \tag{7}$$

4. Probabilistic HEX models

In this section, we introduce an extension of the HEX model to allow for soft or probabilistic relationships between labels. The basic idea is to relax the hard constraints, by replacing the value 0 (corresponding to illegal combinations) in the definitions of the

potential functions with a value $0 \leq q \leq 1$, representing how strongly we wish to enforce the constraints. (This is somewhat analogous to the approach used in Markov logic networks [18], which relax the hard constraints used in first order logic.) In principle, q can be estimated from data along with the parameters of the unary potentials (discussed in Section 6), but in this paper, we either tie the q's across all edges, or set them based on prior knowledge of the strength of the relations.

4.1. Probabilistic HEX relations

For clarity, we now explicitly specify the form of the two new factors we introduce. We use the generic parameter q to represent the strength of this relation, although this could easily be made edge/ label dependent.

Probabilistic exclusion The potential function of the two variables y_1, y_2 under probabilistic exclusion is defined as:

$$\phi_p^e(y_1, y_2; q) = \begin{cases} 1 & (y_1, y_2) \in S^e \\ q & (y_1, y_2) = (1, 1), \end{cases} \tag{8}$$

where $0 \leq q \leq 1$. When $q = 1$, Equation (8) reduces to the non-relation in Equation (7), where y_i and y_j are independent. When $q = 0$, Equation (8) reduces to the hard exclusion relation Equation (5), where $(y_1, y_2) = (1, 1)$ is strictly prohibited.

Probabilistic hierarchy For hierarchy (subsumption), we define

$$\phi_p^h(y_1, y_2; q) = \begin{cases} 1 & (y_1, y_2) \in S^h \\ q & (y_1, y_2) = (-1, 1), \end{cases} \tag{9}$$

where $0 \leq q \leq 1$. This reduces to the unconstrained relation when $q = 1$; and reduces to the hard subsumption relation when $q = 0$.

Probabilistic exclusions and subsumptions can be seen as a probabilistic mixture of absolute exclusions, subsumptions, and non-relations, where

$$\phi_p^e(y_1, y_2; q) = q\phi_a^o(y_1, y_2) + (1 - q)\phi_a^e(y_1, y_2),$$
$$\phi_p^h(y_1, y_2; q) = q\phi_a^o(y_1, y_2) + (1 - q)\phi_a^h(y_1, y_2).$$

Therefore, the combination of probabilistic label relations generalizes the absolute label relations in the HEX graph.

4.2. Converting pHEX models to Ising models

The main disadvantage of this relaxation is that we lose the ability to perform tractable exact inference. However, we now show that we can formulate pHEX models as Ising models, which opens up the door to using standard tractable approximate inference methods.

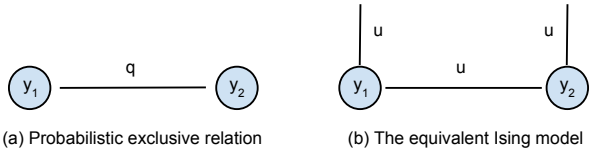

Figure 1. (a) Probabilistic exclusive relations in a pHEX graph with $\phi(1, 1) = q$; (b) the coefficients on the nodes and the edge of the equivalent Ising model, where $q = \exp(-4u)$.

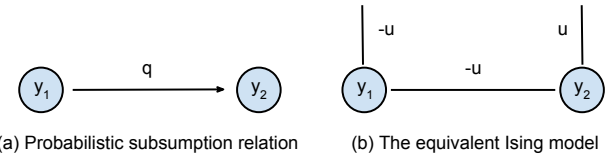

Figure 2. (a) Probabilistic subsumption relations in a pHEX graph with $\phi(-1, 1) = q$; (b) the coefficients of the equivalent Ising model, where $q = \exp(-4u)$.

The Ising model was first proposed in statistical mechanics to study ferromagnetism [3]. Mathematically, it is essentially an undirected graphical model which defines the joint distribution of configurations of n binary random variables $\mathbf{y}$ in graph G by a Boltzmann distribution,

$$p_\beta(\mathbf{y}) = \frac{1}{Z_\beta} \exp(-\beta E(\mathbf{y})), \tag{10}$$

where Z_β is the normalization constant, and β is a temperature variable that will be omitted later by fixing it to 1. $E(\mathbf{y})$ is the energy function of the configuration $\mathbf{y}$, which takes into account local energy potentials $h_i y_i$ as well as pairwise energy potential $J_{ij} y_i y_j$,

$$E(\mathbf{y}) = \sum_{(i,j) \in G} J_{ij} y_i y_j + \sum_{i=1}^{n} h_i y_i. \tag{11}$$

To convert a pHEX graph to an Ising model, we first show how to convert the factor functions $\phi_p(y_1, y_2)$ for the pairwise probabilistic relations to the equivalent pairwise energy functions $E(y_1, y_2)$ of an Ising model.

Consider an Ising model of two variables in Figure 1(b), where $u \geq 0$ are the weights on the local potentials and the pairwise potential. The resulting pairwise

energy function of this Ising model is,

$$E_p^e(y_1, y_2; u) = uy_1y_2 + uy_1 + uy_2 = \begin{cases} -u & (y_1, y_2) \in S^e \\ 3u & (y_1, y_2) = (1, 1). \end{cases} \quad (12)$$

Clearly, Equation (12) looks very similar to Equation (8). In fact, by letting $q = \exp(-4u)$ and $\phi_p^e(y_1, y_2; q) \propto \exp(-E_p^e(y_1, y_2; u))$, we can show they are equivalent up to a constant factor. To see this, let (y_1, y_2) be a legal label pair, and (y_1', y_2') be an illegal pair. We have

$$\phi_p(y_1, y_2)/\phi_p(y_1', y_2') = 1/q = e^{4u},$$
$$\exp(-E(y_1, y_2) + E(y_1', y_2')) = e^{u+3u} = e^{4u}.$$

A larger u means a stronger exclusion between the two labels. When $u \to +\infty$, Equation (12) reduces to the hard exclusive relation; conversely when $u = 0$, Equation (12) reduces to the non-relation.

Similarly, the equivalent Ising model of the probabilistic subsumption is shown in Figure 2(b), where,

$$E_p^h(y_1, y_2; u) = -uy_1y_2 - uy_1 + uy_2 = \begin{cases} -u & (y_1, y_2) \in S^h \\ 3u & (y_1, y_2) = (-1, 1). \end{cases} \quad (13)$$

We set $\phi_p^h(y_1, y_2; q) \propto \exp(-E_p^h(y_1, y_2; u))$ and $q = \exp(-4u)$.

The product of the pairwise factor functions $\phi_p(y_i, y_j; q_{ij})$ can now be written in terms of the sum of pairwise energy functions $E(y_i, y_j; u_{ij})$:

$$\prod_{(i,j)\in G} \phi_p(y_i, y_j, q_{ij}) \propto \exp\left(-\sum_{(i,j)\in G} J_{ij}y_iy_j - \sum_{i=1}^{n} h_iy_i\right),$$

where

$$J_{ij} = \begin{cases} u_{ij}, & (i,j) \in ex. \\ -u_{ij}, & (i,j) \vee (j,i) \in sub. \end{cases} \quad (14)$$

$$h_i = \sum_{\{j|(i,j)\in ex.\}} u_{ij} - \sum_{\{k|(k,i)\in sub.\}} u_{ki} + \sum_{\{l|(i,l)\in sub.\}} u_{il}. \quad (15)$$

Here $ex.$ denotes the set containing all exclusive relations and $sub.$ the set containing all subsumption relations. Note that all the pairs $(i,j) \in ex.$ satisfy $i < j$, and pairs $(i,j) \in sub.$ means i subsumes j.

To incorporate local evidence into the model, we can rewrite Equation (4) as follows:

$$p(\mathbf{y} \mid \mathbf{z}) \propto \exp\left(\sum_{i=1}^{n} \log \psi(y_i, z_i) - \sum_{(i,j)\in G} E(y_i, y_j; u_{ij})\right)$$
$$= \exp(-\sum_{(i,j)\in G} J_{ij}y_iy_j - \sum_{i=1}^{n}(h_i - z_i)y_i),$$

where J_{ij} and h_i are from Equation (14) and Equation (15). Note that we omitted a constant from the $\log \psi$ term, because it will be canceled out by the normalization constant Z. By defining $h_i' = h_i - z_i$, we can "absorb" the local evidence into the Ising model, and use standard inference methods.

4.3. Inference in pHEX models

At test time, we need to compute the marginal distribution per label, $p(y_i \mid \mathbf{z})$. In multi-label classification problems, a label y_i is predicted to be true if $p(y_i \mid \mathbf{z}) \geq 0.5$. In multi-class classification problems, the label

$$y^* = \underset{i=1}{\overset{n}{\operatorname{argmax}}}\, p(y_i \mid \mathbf{z})$$

is predicted to be the true label. At training time, we need $p(y_i \mid \mathbf{z})$ as well as the term $p(y_i | y_j = 1, \mathbf{z})$, where some of the true observed labels (*e.g.* for node j) are set to their desired target states.

Exact inference in pHEX models is usually intractable, when the graphs are loopy, and the legal states are not sparse. Since $p(\mathbf{y} \mid \mathbf{z})$ is an Ising model, we can apply any off-the-shelf inference method, including mean-field inference (MF), loopy belief propagation (LBP), and Markov Chain Monte Carlo (MCMC) methods [4]. In practice, we find that the standard LBP algorithm works consistently well, so we use it as our main inference algorithm in our experiments. We give the details below.

We define the belief on each label y_i to be $b_i(-1)$ and $b_i(1)$, and the message from y_i to its neighbour y_j to be $m_{i\to j}(-1)$ and $m_{i\to j}(1)$. Then the algorithm iterates through all beliefs and messages with updates,

$$b_i(1) \propto \exp(-h_i') \prod_{j\in N(i)} m_{j\to i}(1),$$
$$b_i(-1) \propto \exp(h_i') \prod_{j\in N(i)} m_{j\to i}(-1),$$

where $N(i)$ denotes the neighbours of i, and

$$m_{j\to i}(1) \propto \exp(-J_{ij})\frac{b_i(1)}{m_{i\to j}(1)} + \exp(J_{ij})\frac{b_i(-1)}{m_{i\to j}(-1)},$$
$$m_{j\to i}(-1) \propto \exp(-J_{ij})\frac{b_i(-1)}{m_{i\to j}(-1)} + \exp(J_{ij})\frac{b_i(1)}{m_{i\to j}(1)}.$$

To maintain numerical stability, we normalize b_i and $m_{j\to i}$ throughout inference, and we perform updates in the log domain. After all beliefs have converged or a maximum number of iterations has been reached, we estimate the marginal probabilities by $p(y_i = 1|\,\mathbf{z}) = b_i(1)$.

The inference of $p(y_i|y_j = 1, \mathbf{z})$ is almost the same as above except we set $b_j(1) = 1$ and $b_j(-1) = 0$ to represent the fact that node j is clamped to state 1. (We can easily extend this procedure if we have multiple clamped nodes.)

5. Mutually exclusive and collectively exhaustive relations

In addition to allowing soft relations, our pHEX framework offers another advantage over HEX graphs: it is easy to enforce a new type of constraint, namely Mutually Exclusive and Collectively Exhaustive (MECE) relations, used in the multi-class softmax model. In HEX graphs, there is no way to express the notion of "collectively exhaustive", *i.e.*, one of the mutually exclusive classes must be true. HEX graph thus has to maintain an additional "none of the above" state.

In the pHEX graph, we handle the MECE relation of k nodes using a single multinomial variable with k possible states. Although an undirected graphical model with multinomial nodes is strictly speaking not an Ising model, a slight variant on the standard LBP algorithm can still be applied for efficient approximate inference.

For simplicity, we only illustrate the inference algorithm for pHEX graphs with one multinomial label node, since this will be used in later experiments. Further generalization to pHEX graphs with multiple multinomial nodes is straightforward and follows similar procedures.

Let us denote the multinomial node by $c = \{c_1, \ldots, c_k\}$. The node and message updates for the standard binary nodes are the same as before. The belief of the multinomial node c is updated as,

$$b_c(i) \propto \exp(-h'_i) \prod_{j \in N(c)} m_{j\to c}(i)$$

for state $i \in \{1, \ldots, k\}$ in which $y_{c_i} = 1$. Here $N(c) = \cup_{i=1}^{k} N(c_i)$ is the neighbour set of the multinomial node. The message from a standard node j to the multinomial node c is,

$$m_{j\to c}(i) \propto \exp(\sum_{s=1}^{k} J_{jc_s} - 2J_{jc_i}) \frac{b_j(1)}{m_{c\to j}(1)} + \exp(-\sum_{s=1}^{k} J_{jc_s} + 2J_{jc_i}) \frac{b_j(-1)}{m_{c\to j}(-1)}$$

for state i. The message from the multinomial node to a standard node j is,

$$m_{c\to j}(1) \propto \sum_{i=1}^{k} \exp(\sum_{s=1}^{k} J_{jc_s} - 2J_{jc_i}) \frac{b_c(i)}{m_{j\to c}(i)},$$

$$m_{c\to j}(-1) \propto \sum_{i=1}^{k} \exp(-\sum_{s=1}^{k} J_{jc_s} + 2J_{jc_i}) \frac{b_c(i)}{m_{j\to c}(i)}.$$

As in the standard LBP algorithm, we normalize b_c, $m_{j\to c}$ and $m_{c\to j}$ and update them in the log domain. After the algorithm converges, the marginal probability of a node c_k in clique c is $p(y_{c_k} = 1|\,\mathbf{z}) = b_c(k)$.

6. Learning

An important property of the (p)HEX model is that not all the target labels need to be specified during training. For example, consider a data set of images. It is more common for a user to use basic level category names, such as "dog", than very specific names such as "husky" or "beagle". Furthermore, a user may not label everything in an image. So the absence of a label is not evidence of its absence.

To model this, we allow some of the labels to be unobserved or hidden during training. For example, if we clamp the "husky" label to true, and leave all other label nodes unclamped, the hard constraints will force the "dog" label to turn on, indicating that this instance is an example of both the husky class and the dog class. However, if we clamp the "dog" label to true, we will not turn on "husky" or "beagle", since the relation is asymmetric. We can also clamp labels to the off state, if we know that the corresponding class is definitely absent. For example, turning on "dog" will turn off "cat" if they are mutually exclusive. (In the pHEX case, the "illegal" states are down weighted, rather than given zero probability.)

Let the input scores for the b'th training instance be $\mathbf{z}^b$, and let the subset of target labels be $\mathbf{t}^b = (t_1^b, \ldots, t_m^b)$, where we have assumed that m labels are observed in every instance for notational simplicity. A natural loss function is the negative log likelihood of the observed labels given the inputs:

$$L = -\sum_{b=1}^{N} \sum_{j=1}^{m} \log p(y_{t_j}^b = 1|\,\mathbf{z}^b).$$

To fit the local classifiers (unary potentials), we first need to derive the gradient of the loss wrt the input scores z_i. The derivative of $\log p(y_{t_j} = 1|\,\mathbf{z})$ over some z_i is,

$$\frac{\partial \log p(y_{t_j} = 1|\,\mathbf{z})}{\partial z_i} = \mathbb{E}_{p(y_i|y_{t_j}=1,\mathbf{z})}[y_i] - \mathbb{E}_{p(y_i|\,\mathbf{z})}[y_i].$$

Therefore, we need to compute the conditional distributions $p(y_i|y_{t_j} = 1, \mathbf{z})$ and marginal distributions $p(y_i|\,\mathbf{z})$ for all i. These correspond to the well-known "clamped" and "unclamped" phases of MRF / CRF learning. We can then backpropagate the gradient into the parameters of the local classifiers themselves.

We can use a similar gradient-based training scheme to estimate the CRF edge parameters. However, in this paper, we simply combine prior edge weights from data with a one-dimensional grid search of rescaling factor.

7. Experiments

In [8], the HEX graphs shows significant improvement over standard softmax and (multi-label) logistic regression models, so in this paper, we will just compare pHEX to HEX. We conduct three experiments.

The first experiment is the standard ImageNet image classification problem [7]. We add hierarchical relations between the labels based on the publicly available WordNet hierarchy. Since WordNet does not have exclusive relations, we assume that any two labels are exclusive if they are not in subsumption relation. Figure 3 Left is an example of the subgraph of "fish". As in [8], we assume that the training labels are drawn from different levels of the hierarchy. In this paper, we show that pHEX with (constant) soft relations improves on HEX, especially when leaf labels are rarely present in the training set.

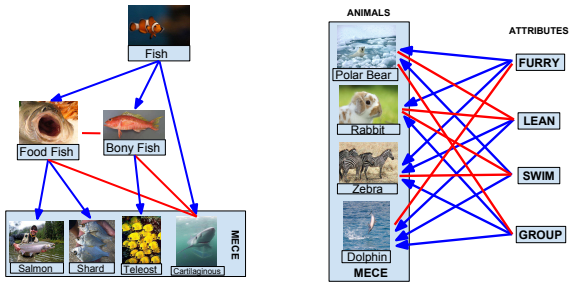

Figure 3. Left: An illustration of the (p)HEX graph based on the WordNet hierarchy in the ImageNet experiments. Right: An illustration of the (p)HEX graph in the Animal with Attributes experiments. The blue directed edges denote the subsumption relations; and the red undirected edges denote the exclusive relations. An MECE relation (multinomial node) is placed in the final pHEX graph.

The second experiment is a zero-shot learning task, in which we must predict unseen classes at test time, leveraging known relations between the class labels and attributes of the class. We use the Animals with Attributes dataset [13]. Following [8], we first assume that all object classes are mutually exclusive. We then add subsumption relations from a predicate (or attribute) to an object if the binary predicate of the object is 1, and add exclusive relations between predicate and objects if the binary predicate of the object is 0. See the illustration in Figure 3 Right. In this paper, we relax the hard constraints and show that pHEX can work significantly better than HEX. Finally, the third experiment is another zero-shot learning task, this time on the PASCAL VOC/ Yahoo images with attributes dataset [10]. Again, we show that pHEX can significantly outperform HEX.

7.1. Experimental setup

Table 1. The Ising coefficients u as well as the corresponding strengths of the label relations q used in pHEX graphs in the experiments, where $q = \exp(-4u)$.

u	0	0.1	0.3	0.5	0.7	1.0	1.5
q	1	0.67	0.30	0.14	0.06	0.02	0.002

In our experiments, we used two types of pHEX graphs. For the ImageNet experiments, we use the same constant edge strength for all edges; we vary this edge paramter u across the ranges shown in Table 1, and plot results for each value. For the zero-shot experiments, we consider constant edge weights, but we also consider variable edge weights, which we derive by scaling the prior edge weight (derived from the data) by a global scale factor u, which we again vary across a range.

Note that, since all three tasks are evaluated on test labels in a multi-class setting, we add a MECE relation into the pHEX graphs In particular, for the ImageNet dataset, we add a multinomial node on the 1000 leaf labels; in the Animal with Attributes dataset, we add a multinomial node on the 50 animal classes; and in the VOC/Yahoo dataset, we add a multinomial node on the 32 object classes. After adding MECE relations, we remove the replicated soft exclusive relations from the pHEX graph.

7.2. ImageNet classification experiments

In this section, we use the ILSVRC2012 dataset [7], which consists of 1.2M training images from 1000 object classes. These 1000 classes are mutually exclusive leaf nodes of a semantic hierarchy based on WordNet

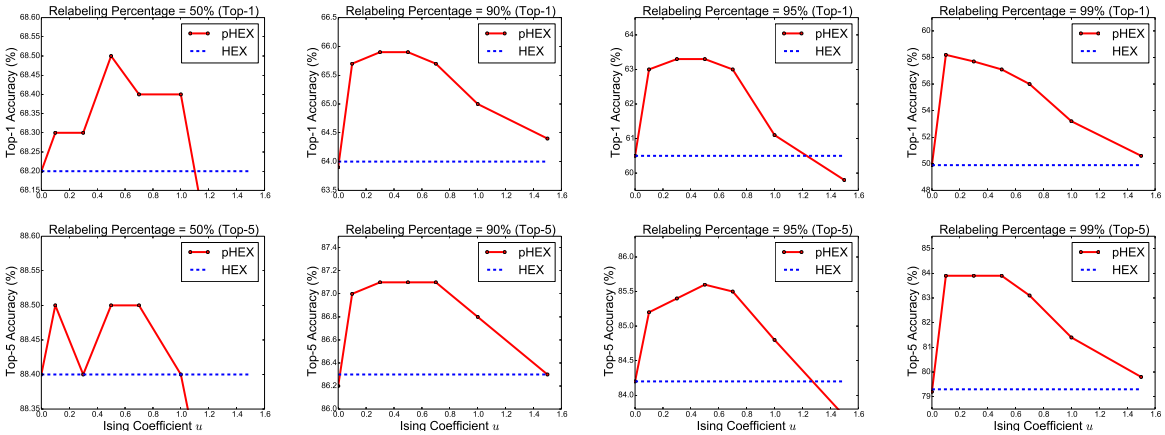

Figure 4. Top-1 (top) and Top-5 accuracies (bottom) vs relation strength u for the ImageNet classification experiment. The results of the pHEX graphs are in the red solid curves, and the results of the HEX graphs are in the blue dashed horizontal lines. From left to right: relabeling 50%, 90%, 95%, 99%.

that has 860 internal nodes. As in [8], we evaluate the recognition performance in the multiclass classification at the leaf level, but allow the training examples to be labeled at different semantic levels. Since ILSVRC2012 has no training examples at internal nodes, we create training examples for internal nodes by relabelling $\{50\%, 90\%, 95\%, 99\%\}$ of the leaf examples to their immediate parents based on the WordNet Hierarchy. Since the ground truth for test set is not released for ILSVRC2012, we use 10% of the released validation set as our validation set and the other 90% as our test set.

The underlying feed-forward network that we use is based on a deep convolutional neural network GoogLeNet [22]. Since GoogLeNet is such a large model, we adopt the following staged training procedure. First we pre-train a CNN with a HEX graph as the top layer until convergence. Then we fine tune the entire model with pHEX graph layers of different coefficients u on top. This can be thought of as a form of curriculum learning [2] by training with a simpler model (HEX graph) with exact inference first.

Figure 4 shows the Top-1 (top row) and Top-5 (bottom row) accuracies across classes as a function of u, for the relabeling experiments. For comparison, the Top-1 (top row) and Top-5 (bottom row) accuracies without relabeling (i.e., the standard ImageNet setup) is 70.1% and 90.0% respectively. Not surprisingly, relabeling (i.e., only providing some labels at the leaves, and using coarser grained categories for the rest) hurts performance (as estimated by leaf-level accuracy). However, in this regime (which occurs commonly in practice), pHEX generally outperforms HEX, especially for 90%, 95% and 99% relabeling, where the accuracies improve by 2%, 3% and 8% respectively. (Note that a 1% difference in performance is considered statistically significant on this problem due to the large size of this dataset.)

At first, it might seem odd that relaxing the hard constraints imposed by the hierarchy can help, since the hierarchy provided by WordNet is supposed to be correct. However, [8] observed that too few training examples labeled at leaf nodes (especially at 99% relabeling) may confuse the leaf models, especially at the beginning of the training. As the algorithm runs longer, it becomes harder to recover from a bad local minimum because the constraints in the HEX graph are hard constraints. By contrast, in the pHEX graph, the weaker relations between internal nodes and leaf nodes make the resulting posterior distribution smoother, so it is easier to overcome bad local minima for the pHEX graph in later iterations.

It is also interesting to see that the optimal value of u appears to depend on the relabeling percentage. When a larger portion of training examples are relabeled, e.g. 99% relabeling, the optimal relation coefficient becomes weaker ($u = 0.1$). This indicates that weaker label relations are preferred when there is more uncertainty in the leaf labels.

On the other hand, when u is large, the label relations become quite certain and the the pHEX graph becomes closer to the HEX graph. In the case of $u = 1.5$ ($q \simeq 0.002$), the performance of pHEX graph can become worse than HEX graph, probably due to the inability to perform exact inference in the pHEX graph.

7.3. Zero shot learning experiments

We use two datasets to illustrate zero shot learning. The first is the Animals with Attributes dataset [13], which includes images from 50 animal classes. For each animal class, it provides both binary and continu-

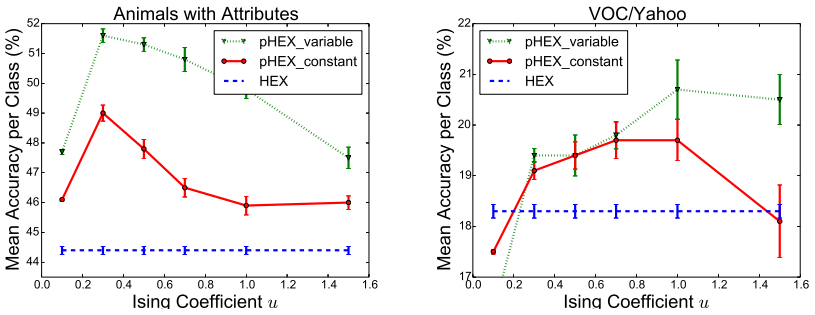

Figure 5. Mean accuracy per class vs relation strength u for the Zero-shot Learning Experiments. Left: animals with attributes. The results of the pHEX with variable edge weights are in the dotted green, and the ones with constant edge weights are in solid red. The results of the HEX graphs are in the blue dashed horizontal lines. Right: VOC/Yahoo images with attributes.

ous predicates for 85 attributes. We convert the binary predicates to constant (soft) relations, and the continuous predicates to variable soft relations by a monotonic mapping function. The details of the mapping are provided in the supplementary material. We evaluate the zero-shot setting where training is performed using only examples from 40 animal classes (with 24295 images) and testing is on classifying the 10 unseen classes (with 6180 images). Our experimental results are based on 5-fold cross validation. The underlying network is a single-layer network whose inputs come from the recently released DECAF features [9].

The second dataset is the aPascal-aYahoo dataset [10], which consists of a 12695 image subset of the PASCAL VOC 2008 dataset and 2644 images that were collected using the Yahoo image search engine. The PASCAL part serves as training data and has 20 object classes. The Yahoo part serves as test data and contains 12 different object classes. Each image has been annotated with 64 binary attributes that characterize shape, material and the presence of important parts of the visible object. We convert them to binary and continous predicates for attributes per object by averaging the image annotations for every object (details in supplementary material). The underlying network is again a single-layer network whose inputs come from the features that the authors of [10] extracted from the objects bounding boxes (as provided by the PASCAL VOC annotation) and released as part of the dataset. Once again we use 5-fold cross validation and compares constant soft relations and variable soft relations with hard relations.

Figure 5 shows the mean accuracy per class (along with standard errors) vs u. We see that pHEX is generally significantly outperforming HEX. In particular, when $u \in [0.1, 1.5]$ for Animals with Attributes and $u \in [0.3, 1.0]$ for VOC/Yahoo, the difference is statistically significant at the 5% level according to a paired t-test. The accuracies of the pHEX graph get closer to the ones of the HEX graph as u becomes larger and the pHEX graph approaches to the HEX graph. Moreover, the pHEX models with variable soft relations improves over the ones with constant soft relations by 2% for Animals with Attributes and 1% for VOC/Yahoo. This demonstrates the value of adding additional information in the variable probabilistic label relations in transfer learning.

7.4. Speed comparison of HEX vs pHEX

In the ImageNet experiements, the cost of HEX and pHEX is similar, since most of the time is spent evaluating the underlying deep CNN. In the two zero-shot learning experiments, the inference time of the pHEX graph is about the same as the one of the HEX graph. Furthermore, many other algorithms such as quantum annealing [11] (which are faster and/or more accurate than loopy belief propagation) have been devised for Ising models which we could try in the future.

8. Conclusions

In this paper, we studied object classification with probabilistic label relations. In particular, we proposed the pHEX graph, which naturally generalizes the HEX graph. The pHEX graph is equivalent to an undirected Ising model, which allows for efficient approximate inference methods. We embed the pHEX graph on top of a deep neural network, and show that it outperforms the HEX graph on a number of classification tasks which require exploiting label relations.

There are several possible future directions of this work. One idea is to learn the Ising coefficients of the pHEX graph together with the underlying neural network parameters. Another is to combine the pHEX graph into a larger framework which exploits spatial relations between objects.

References

[1] Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid. Label-embedding for attribute-based classification. In *CVPR*, pages 819–826. IEEE, 2013. 2

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009. 7

[3] K. Binder. Ising model, 2001. Encyclopedia of Mathematics, Springer, ISBN 978-1-55608-010-4. 3

[4] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 4

[5] X. Chen, X.-T. Yuan, Q. Chen, S. Yan, and T.-S. Chua. Multi-label visual classification with label exclusive context. In *ICCV*, pages 834–841. IEEE, 2011. 2

[6] B. Dalvi, E. Minkov, P. P. Talukdar, and W. W. Cohen. Automatic gloss finding for a knowledge base using ontological constraints. In *WSDM*, WSDM '15, pages 369–378. ACM, 2015. 2

[7] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei. Imagenet large scale visual recognition challenge 2012, 2012. www.image-net.org/challenges/LSVRC/2012. 6

[8] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classfication using label relation graphs. In *ECCV*, pages 48–64, Zurich, Switzerland, September 2014. Springer. 1, 2, 6, 7

[9] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. Decaf: A deep convolutional activation feature for generic visual recognition. In *ICML*, 2014. 8

[10] A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth. Describing objects by their attributes. In *CVPR*, 2009. 6, 8

[11] T. Kadowaki and H. Nishimori. Quantum annealing in the transverse ising model. *Physics Review E*, 58, 1998. 1, 8

[12] J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, ICML '01, pages 282–289, San Francisco, CA, USA, 2001. 2

[13] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958. IEEE, 2009. 2, 6, 7

[14] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *CVPR*, pages 1–7. IEEE, 2007. 2

[15] F. Mirzazadeh, S. Ravanbakhsh, B. Xu, N. Ding, and D. Schuurmans. Embedding inference for structured multilabel prediction. In *NIPS*, 2015. 2

[16] V. Ordonez, J. Deng, Y. Choi, A. C. Berg, and T. L. Berg. From large scale image categorization to entry-level categories. In *ICCV*, 2013. 2

[17] M. Palatucci, G. Hinton, D. Pomerleau, and T. M. Mitchell. Zero-shot learning with semantic output codes. In *NIPS*, 2009. 2

[18] M. Richardson and P. Domingos. Markov logic networks. *Machine Learning*, 62:107–136, 2006. 3

[19] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011. 2

[20] M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele. What helps where - and why? semantic relatedness for knowledge transfer. In *CVPR*, pages 910–917. IEEE, 2010. 2

[21] V. Sharmanska, N. Quadrianto, and C. H. Lampert. Augmented attribute representations. In *ECCV*, pages 242–255, 2012. 2

[22] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions, 2014. arXiv:1409.4842. 7

[23] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, Dec. 2005. 2

[24] S. V. N. Vishwanathan, N. N. Schraudolph, M. W. Schmidt, and K. P. Murphy. Accelerated training of conditional random fields with stochastic gradient methods. In *ICML*, ICML '06, pages 969–976, New York, NY, USA, 2006. ACM. 2

[25] C. Wu, I. Lenz, and A. Saxena. Hierarchical semantic labeling for task-relevant rgb-d perception. In *Proceedings of Robotics: Science and Systems*, Berkeley, USA, July 2014. 2

[26] F. Yu, L. Cao, R. Feris, J. Smith, and S.-F. Chang. Designing category-level attributes for discriminative visual recognition. In *CVPR*, Portland, OR, June 2013. 2

[27] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *ICCV*, pages 1–8. IEEE, 2007. 2