Learn from Your Neighbor: Learning Multi-modal Mappings from Sparse Annotations

Ashwin Kalyan [1] **Stefan Lee** [1] **Anitha Kannan** [2] **Dhruv Batra** [1 3]

Abstract

Many structured prediction problems (particularly in vision and language domains) are ambiguous, with multiple outputs being 'correct' for an input – *e.g*. there are many ways of describing an image, multiple ways of translating a sentence; however, exhaustively annotating the applicability of all possible outputs is intractable due to exponentially large output spaces (*e.g*. all English sentences). In practice, these problems are cast as multi-class prediction, with the likelihood of only a sparse set of annotations being maximized – unfortunately penalizing for placing beliefs on plausible but unannotated outputs. We make and test the following hypothesis – for a given input, the annotations of its neighbors may serve as an additional supervisory signal. Specifically, we propose an objective that *transfers* supervision from neighboring examples. We first study the properties of our developed method in a controlled toy setup before reporting results on multi-label classification and two image-grounded sequence modeling tasks – captioning and question generation. We evaluate using standard task-specific metrics and measures of output diversity, finding consistent improvements over standard maximum likelihood training and other baselines.

1. Introduction

In many real-world tasks, a single input is associated with multiple correct outputs. For instance, as shown in Fig. 1, multiple captions can accurately describe an image. For tasks with small output spaces, it may be practical to treat the problem under a multi-label formulation – learning to predict the correctness of each possible output based on exhaustive human annotations. However, for structured prediction tasks, the output space is exponentially large (*e.g*. the space of all English sentences) such that collecting exhuastive annotations is intractable even for a single input. Instead, sparse annotations are obtained by collecting human responses – leaving the correctness of a vast majority of possible outputs uncertain.

Figure 1. Many tasks exhibit many-to-many relationships between inputs and outputs. Taking image captioning as an example, a single image can be described with multiple captions (top) and likewise a single caption can accurately describe multiple images (bottom). In this work, we leverage these relationships in the data to learn multi-modal output mappings from sparse annotations.

This problem of *multi-label classification with missing labels* has been addressed in prior work by either imposing structure on the label space such as known label taxonomies (Verma & Jawahar, 2013; Deng et al., 2014), or by imposing constraints on the model parameters (Yu et al., 2014) or the posterior distributions (Lin et al., 2014b) to effectively compress the label space. However, these approaches do not scale to the exponentially large label spaces often seen in structured prediction tasks like sequence-modeling (*e.g*. $|\mathcal{V}|^T$ length–T sentences in captioning where $\mathcal{V}$ is the vocabulary). As a consequence, such tasks are often cast as multi-class problems with parameters learned to maximize the likelihood of a sparse set of human annotations (*i.e*. Maximum Likelihood training) – implicitly enforcing the unreasonable assumption that all outputs that are not annotated must be incorrect.

Much contemporary research has been invested in the more expensive yet viable option of curating massive datasets that

[1]Georgia Tech [2]Curai [3]Facebook AI Research. Correspondence to: Ashwin Kalyan <ashwinkv@gatech.edu>.

Proceedings of the 35^{th} International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018.

leads to better estimation of the true multi-modal mapping with increasing dataset size. For instance, progress in captioning has been largely propelled by the massive COCO (Lin et al., 2014a) dataset containing $\sim 330K$ images with 5 human-provided captions per image. However, as evidenced by the impoverished, generic captions generated by models trained on this dataset (Vijayakumar et al., 2018; Dai et al., 2017), even such large-scale efforts fall short – capturing only a small fraction of all possible outputs.

Overview and Contributions. In this work, we propose a simple approach that enables models to place beliefs on multiple plausible outputs while training only on sparse set of annotations, or in the extreme case only a single annotation per input on tasks where there are many many possible correct outputs. Essentially, our goal is to learn to produce multi-modal outputs from 'uni-modal' annotations. The key inductive bias in our approach is the following – for a given input, the annotations of its neighbors may serve as an additional supervisory signal. Fig. 1 (bottom) demonstrates this intuition for captioning with the caption accurately describing all four depicted scenes. Based on this insight, we propose a novel objective that treats outputs of neighboring inputs to be applicable to the given input to an extent determined by the similarity in the input space. This objective allows us to *transfer* annotations from neighboring examples to provide additional supervision and so contribute towards recovering the underlying multi-modal mapping.

In order to analyze our approach in a tractable domain, we perform a number of multi-label classification experiments with missing labels. First, we evaluate in a toy setting where the data generating distribution is known and find that our method is able to better estimate the true distribution as compared to standard cross-entropy training. We also study multi-label prediction on two real-world datasets – CUB-200 (Wah et al., 2011) and Animals with Attributes (AWA) (Xian et al., 2017) – by sub-sampling attribute annotations. As in the toy setting, we see improvements over baseline methods. Finally, we apply our method to two established image-grounded language generation tasks – image captioning and question generation – which are both sequential prediction tasks with exponentially large label spaces. We evaluate using both standard task-specific metrics for the generated language and criteria that assess the multi-modal nature of the produced outputs. We find consistent improvements over baseline methods on these challenging tasks.

2. Approach

We first establish the notation and succinctly summarize the learning problem before explaining the proposed approach.

Consider a multi-label prediction setting where the goal is to learn a *one-to-many* relationship $f{:}\mathcal{X} \rightarrow 2^{\mathcal{Y}}/\emptyset$ that maps a given input $\mathbf{x}{\in}\mathcal{X}$ to a set of valid outputs, a subset of all possible outputs $\mathcal{Y}$. In our setting, obtaining annotations for each element in $\mathcal{Y}$ is intractable even for a single instance and instead only a sparse set of *positive* annotations are available. Specifically, we assume access only to a dataset of the form $\mathcal{D} = \{(\mathbf{x}_m, \{y_{m,1}, \ldots, y_{m,k}\})\}_{m=1}^{M}$ where $\mathbf{x}_m$ is the input and $\{y_{m,1}, \ldots, y_{m,k}\}$ is a sparse set of labellings with $k << |f(\mathbf{x}_m)|$. In practice, k may vary for each example and often, $k = 1$ *i.e.* only one annotation is available.

The observed dataset $\mathcal{D}$ can be thought of as being produced by a stochastic function g that selects k labels from $f(\mathbf{x})$, the set of all applicable labels for $\mathbf{x}$. In practice, a collection of human annotators often play the role of g, generating a small set of possible outputs for each $\mathbf{x}$ (*e.g.* each providing a single image description in captioning). We therefore summarize the overall learning problem as – *how can we estimate the true multi-modal input-output relationship f while only observing sparse samples from g?*

2.1. Enforcing Adaptive Neighborhood Structures

At a high-level, our approach has two key components – 1) a mechanism that allows us to use outputs of neighboring inputs to provide additional supervision and 2) an appropriate measure of semantic relatedness to define the neighborhood. We now explain both these aspects in detail.

Learning from Neighbors. The predictions $\tilde{y}_m$ outputted by the model for each input $\mathbf{x}_m$ are evaluated using a loss function $\ell : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and the standard objective is to reduce the empirical risk on the training set:

$$\frac{1}{Mk} \sum_{m=1}^{M} \sum_{k} \ell\left(\tilde{y}_m,\ y_{m,k}\right) \tag{1}$$

Due to the presence of multiple annotations, the summations cover both examples (m) and their annotations (k).

Let us begin by assuming access to a function $r : \mathcal{X} \rightarrow \mathbb{R}^d$,

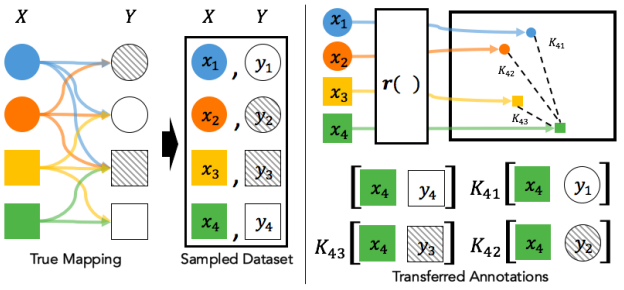


Figure 2. Given a dataset of sparse annotations sampled from a true multi-modal input-output mapping (left), our approach leverages a learned similarity space to perform a soft-transfer of annotations between semantically related inputs (right) – effectively recovering the underlying multi-modal mapping from few samples.

a potentially non-linear transformation that maps inputs to a space where distances correspond to semantic relatedness. Let $\mathcal{K}_{ij}$ be the *similarity* between $\mathbf{x}_i$ and $\mathbf{x}_j$ in this semantic space. Note that all distances and similarities are computed in this semantic space unless otherwise mentioned. Equipped with this semantic space, we now define the neighborhood $\mathcal{N}(\mathbf{x})$ of a data point $\mathbf{x}$ – specifically, let $\mathcal{N}(\mathbf{x})$ to be the set of indices of the N-nearest neighbors of $\mathbf{x}$. Recall that we wish to incorporate the key inductive bias that outputs of semantically similar inputs can be potentially 'correct' for a given input. Thus, we can now write a regularized objective that encourages the model to place beliefs on multiple outputs as:

$$\underbrace{\ell(\tilde{y}_i, y_{i,k})}_{\text{loss w.r.t. own label}} + \underbrace{\frac{\lambda}{|\mathcal{N}(\mathbf{x}_i)|}}_{\text{normalization}} \sum_{j \in \mathcal{N}(x_i)} \overbrace{\mathcal{K}_{ij}}^{\text{similarity to neighbor}} \underbrace{\ell(\tilde{y}_i, y_j)}_{\text{loss w.r.t. neighbor's label}} \tag{2}$$

where λ is a hyper-parameter that controls the importance of additional supervision. The weighting of the additional supervision using the similarity $\mathcal{K}_{ij}$ can be thought of as accounting for the *uncertainty* in the applicability of neighboring output y_j to the input $\mathbf{x}_i$, due to the lack of its annotation. In this work, we set $\mathcal{K}_{ij} = \max\left\{0, \cos\left(r(\mathbf{x}_i), r(\mathbf{x}_j)\right)\right\}$ where cos is the cosine similarity.

Connections to label smoothing. Unlike maximum likelihood training that penalizes unannotated predictions, our objective encourages the model to place beliefs on outputs of neighboring inputs apart from its own annotatation. In a simple $\mathcal{C}$-way classification setting, it is easy to see that this corresponds to label smoothing with class $c \in \mathcal{C}$ assigned a mass proportional to $\sum_{j \in \mathcal{N}(\mathbf{x}), g(\mathbf{x}_j)=c} \mathcal{K}_{ij}$. Thus, our loss redistributes mass in a systematic, input-aware fashion unlike Szegedy et al. (2015) or Pereyra et al. (2017) that uniformly increase the uncertainty in the predictions.

Learning the semantic space. As mentioned before, computation of the neighborhood $\mathcal{N}(\mathbf{x})$ and the similarities $\mathcal{K}_{ij}$ assumes access to a projection $r(\cdot)$ that maps inputs to a semantic space. Unless strong priors exist like known taxonomies exist in the input space, there is no obvious choice for this projection. As such, we propose learning it alongside the task – specifically, we initialize $r(\cdot)$ with some domain specific neural network (*e.g.* pre-final layer of a CNN on ImageNet (Deng et al., 2009) for natural images) and then finetune it jointly with the model.

In practice, if the network for learning $r(\cdot)$ has sufficient capacity, it can project all points in the dataset to a unique dimension of its own s.t. $\mathcal{K}_{ij} = 0 \; \forall i, \forall j, i \neq j$, reducing our objective to MLE, (1). Further, looking at the derivative of the objective w.r.t. $\mathcal{K}_{ij}$

$$\frac{\partial \mathcal{L}_i}{\partial \mathcal{K}_{ij}} = \frac{\lambda \ell(\tilde{y}_i, y_j)}{|\mathcal{N}(\mathbf{x})|}$$

we can see that the objective constantly works towards reducing the similarity $\mathcal{K}_{ij}$ as $\ell(\cdot, \cdot)$ is always non-negative. To constrain the network from pushing similar data points apart, we regularize by penalizing the model for deviating too much from the initial structure as –

$$\frac{\mu}{|\mathcal{N}(\mathbf{x}_i)|} \sum_{j \in \mathcal{N}(\mathbf{x}_i)} (\mathcal{K}_{ij} - 1)^2 \tag{3}$$

where μ is a hyper-parameter that controls the strength of this penalty. Note that this penalty implies that our initial choice for $r(\cdot)$ is already reasonable and only requires minor adjustments for it to be task-specific.

2.2. Generalization to Sequence Prediction

Consider sequential output tasks where an input $\mathbf{x}$ is mapped to a sequence $\mathbf{y} = \{y_1, \ldots, y_T\}$. The standard objective for seqeunce modeling is to maximize log-likelihood of the output token at time t given previous tokens and the input as $\sum_t \log \Pr(y_t | y_{t-1}, \ldots y_1, \mathbf{x})$. We can trivially extend our objective in Eq. (2) by weighing each term inside the summation with $\mathcal{K}_{ij}$. However, for grounded sequence generation tasks like captioning, it is often the case that only a portion of the neighbor's output is applicable to a given input; for instance, a pair of images may both contain a dog, but only one also has a cat. In such cases, only specific phrases or words may be reasonably borrowed between images (*i.e.* "big dog"). To incorporate this notion of 'partial' supervision, we extend our objective to leverage attentional models like that of Lu et al. (2017).

We now briefly explain this attentional model and refer the reader to Lu et al. (2017) for a more detailed discussion. Consider a set of visual features $V = \{v_1, \ldots, v_k\}$ that each encode different regions of the image and a global feature v_g given by their average. As shown in Fig. 3, the model takes in this global image feature, w_t an embedding of the previous word y_t and the spatial features V to compute the attention vector, $\boldsymbol{\alpha}_t \in \mathbb{R}^{k+1}$ that weighs the importance of each of the spatial regions and the history encoding h_t to compute the posterior $\log \Pr(y_t | y_{t-1}, \ldots, \mathbf{x})$. Unlike standard attention-based architectures (Xu et al., 2015) that only attend to image regions, Lu et al. (2017) extend it to incorporate both image and language components. Interpreting the sum of visual attention weights, denoted by $\alpha_t \in [0, 1]$ as the importance of the image for the generation of the next word allows us to incorporate the notion of only copying partial sequences by weighing each term in the factored log-likelihood by this visual importance weight. Specifically, let $\alpha_{j,t}$ be the *visual* importance of the neighbor $\mathbf{x}_j$ for predicting the word $y_{j,t}$

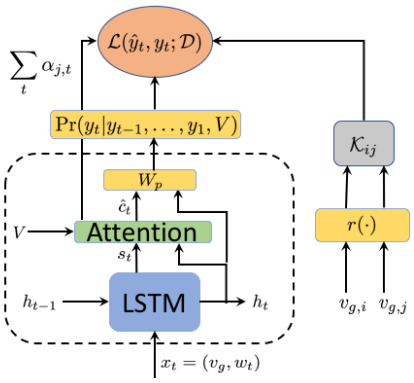


Figure 3. An diagram of our approach for sequential prediction task. Only the relevant segments (via a soft attention mechanism) from the neighbors output are used. See Sec. 2.2 for details.

given $\mathbf{x}_i$. Then, the second term in (2) can be modified as:

$$-\frac{\lambda}{\mathcal{N}(\mathbf{x})} \sum_{j \in \mathcal{N}(\mathbf{x})} \mathcal{K}_{i,j} \sum_{t \in [T]} \alpha_{j,t} \log \Pr(y_{j,t}|y_{j,t-1}, \ldots, y_1, \mathbf{x}_i) \tag{4}$$

Likewise, the regularization constraint from (3) is updated to ensure that α's do not go to zero:

$$\frac{\mu}{T\mathcal{N}(\mathbf{x})} \sum_{j \in \mathcal{N}(\mathbf{x})} \mathcal{K}_{ij} \left(\sum_{t \in [T]} \alpha_{j,t} - 1 \right)^2 . \tag{5}$$

2.3. Implementation Details.

We now discuss some subtle but important implementation details training models using our objective.

Each mini-batch consists of B examples and their corresponding N neighbors (including itself) – resulting in a total of $B \times N$ samples that need to be processed for an *effective* batch-size of B. The procedure for sampling neighbors has to be "aware" of the constantly changing representation space $r(\cdot)$ and so, is also updated in tandem. We call this *adaptive* updation of the neighborhood of each image. As it is expensive to compute similarities and sample these mini-batches, we make the following practical choices – First, the parameters of $r(\cdot)$ are updated using a *much* smaller learning rate ($10\times$) compared to the model itself. Second, the similarities between data points and thus the neighborhood is only updated every few iterations.

Finally, our method is fairly robust to the setting of λ and μ, the repulsive and attractive terms in our objective. In general, as the number of neighbors participating in the objective increases smaller values of λ suffice. Further, large values of μ enforce the neighborhood to strongly respect the initial structure and so, depending on the quality of the initialization of this representation space, μ can be varied. A more detailed analysis of the these hyper-parameters is provided in the supplement.

3. Related Work

Multi-label Classification with missing labels. (Verma & Jawahar, 2013) extend the work of Bucak et al. (2011) by incorporating taxonomy of the label space into their cost-senstive ranking formulation. Extending these methods to sequential outputs is challenging as there is no documented or natural way of constructing such a taxonomy for say, English sentences. Instead, our approach jointly learns dependencies between data points and does not require similarity or taxonomy information as input. Further, Yu et al. (2014) propose imposing a low-rank constraint on the weights to be able to capture correlation between labels. However, such norm-based regularization schemes have been showed to be ineffective for deep networks (Zhang et al., 2017). In a similar vein, Lin et al. (2014b) constrain the posterior to be of low rank. However, such an approach is not feasible when the output is a sequence.

Semi-supervised learning. The use of homophily for semi-supervised learning has been well-studied (c.f. Zhu (2005); Zhu et al. (2003) for a survey). A dominant approach is the use of relationship graphs as regularization – these methods assume access to a relationship graph (that is not available in our setting) as part of the input, or that it can be easily gleaned through measuring similarity in the input space; the underlying assumption being that the network structure is independent of the labels given the input. Weston et al. (2012) extends this line of work by embedding inputs using a deep neural network. Perozzi et al. (2014) and Yang et al. (2016) extends this to infer graph context to aid in classification. In general, the goal of this line of work is to enable a consensus in the label assignment of unlabeled examples, by leveraging the neighborhood structure. However, in our setting all examples are labeled, albeit incompletely (missing labels). The central hypothesis of our work is that there exists an embedding space in which neighborhood structure is evident where neighboring data points can be used to supervise the learning of a multi-modal output space. Thus, the model jointly optimizes for both – uncovering this underlying semantic structure in the data while also, learning a multi-modal input-output mapping.

Entropy Regularization and Label smoothing Pereyra et al. (2017) propose regularizing the model with a negative entropy term and in a similar vein, Szegedy et al. (2015) propose a simple label smoothing strategy where an arbitrarily small probability mass is re-distributed to classes other than the observed ground-truth uniformly. Further, Chorowski & Jaitly (2017) draw from Szegedy et al.

(2015) and propose a neighborhood smoothing scheme for a n-gram language model where probability mass is distributed based on observed n-grams in the dataset and not uniformly like the previous work. In contrast, the goal of our formulation is to not just increase or decrease the prediction entropy but to *shape* the conditional $\Pr(y|\mathbf{x})$ to reflect multi-modal input-output mappings by placing meaningful beliefs on the output space. In that spirit, our work shares motivation with the line of work on producing diverse structured outputs – (Batra et al., 2012; Guzman-Rivera et al., 2012; Prasad et al., 2014; Guzman-Rivera et al., 2014; Lee et al., 2016).

Non-MLE based Image-captioning. Interestingly, Dai et al. (2017); Shetty et al. (2017) obtain diverse outputs (relative to MLE) using adversarial training without making any explicit assumptions about the multi-modal nature of the task. As Shetty et al. (2017) requires a multi-modal dataset, we instead compare to Dai et al. (2017) and show that we outperform when only having access to limited uni-modal data. Similarly, Jain et al. (2017) use variational auto-encoders for the task of producing visually grounded questions and report diverse outputs. However, as observed with adversarial training, without multiple output annotations the latent variable does not contribute in capturing the multi-modal output space. Finally, we also compare to Rennie et al. (2017) that directly optimizes for the task-specific metric (like CIDEr (Vedantam et al., 2015) or SPICE (Anderson et al., 2016) for image-captioning) using policy gradients and show that we outperform their approach in our setting.

Nearest-neighbor based captioning. While Devlin et al. (2015) explore a nearest-neighbor based approach to captioning, Chen et al. (2017) build on it to propose a modified objective that weighs each word based on its occurrence in nearest neighbor images. Similarly, Mun et al. (2017) propose an attention scheme that factors in the *consensus* caption. Unlike our approach, both these methods push the model towards producing more generic descriptions that are applicable to multiple similar images.

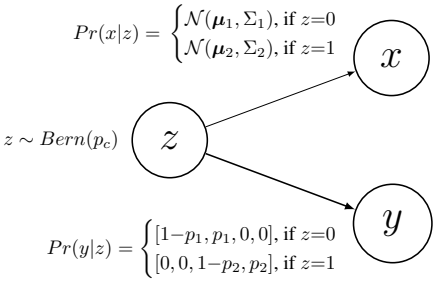
Similarity based on outputs. Inan et al. (2017) propose a re-use of the word-embeddings by augmenting the cross-entropy loss with a KL divergence term between the predictions and the normalized vector of the dot-products between the target-word embedding and the entire vocabulary. This term encourages the model to place belief on completions that are not necessarily observed in the dataset. While the high-level goals of both our objective and this work are similar, this approach relies only on distances in the output space and does not factor that the language generated can be conditioned on inputs in a different perceptual modality.

4. Experiments

We first explore the properties of our objective in a controlled toy setting and evaluate the performance w.r.t. the true data distribution. For completeness, we then show results on the multi-label with missing labels task on standard multi-label attribute datasets and finally, discuss the performance of our method on two visually-grounded language generation tasks – captioning and question generation.

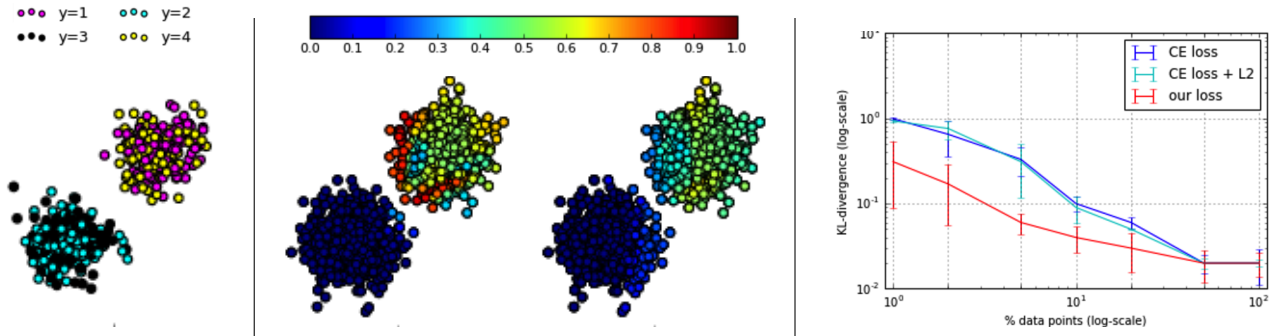
4.1. Synthetic Experiments

We consider a 4-label classification problem, where the dataset $\mathcal{D} = \{(\mathbf{x}, y)\}$ is generated according to the graphical model shown below.



Specifically, each data point $\mathbf{x}$ is sampled from one of two Gaussians – $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ or $\mathcal{N}(\boldsymbol{\mu}_1, \Sigma_2)$ depending on the state of the latent variable z. Each Gaussian is associated with two of the four labels ($z_0 \rightarrow \{y_1, y_2\}$ and $z_1 \rightarrow \{y_3, y_4\}$). However, for each data point $\mathbf{x}$ we observe only one of the two possible labels. Using terminology from Section 2, the true multi-modal mapping f maps the input in each cluster to two labels; however, we only observe one due to a stochastic label sampling function g. Fig. 4(a) shows a plot of a dataset generated through this process (means of $(-1, -1)$ and $(1, 1)$ with a diagonal variance of 0.2). To simulate the data-sparse regimes typical of real-world tasks, we transform the data to a much higher dimensional space (*e.g.* 2^{13} in our experiments) through a randomly initialized deep neural network with ReLU activations that doubles the input dimensionality. For a trained classifier to perform well, it needs to discover the 2D data manifold that reveals the underlying neighborhood structure on which labels are based; and not simply overfit to local statistics in the high dimensional space. Since the underlying data generation process is known, we can examine the hypothesis using KL divergence between the true posterior and the predicted distribution of trained models.

Implementation Details. In our experiments, we use a dataset of size 2048; 512 for training and the rest to for evaluation. We use a two-layered neural network with 32 neurons in each layer and train it via SGD with a learning rate of $4e{-}5$ and a momentum of 0.9. Further, we also compare with training using cross-entropy (CE) in conjunction



(a) Synthetic Dataset (b) $P(y{=}1|x)$: CE+L2 (c) $P(y{=}1|x)$: Ours (d) Sample Efficiency

Figure 4. (a) Our toy experiment uses synthetic data with uniform label mixing within each cluster. (b) We find that cross-entropy (CE) training (even with L2 regularization) results in overfitting – for instance, some regions of the rightmost cluster are very confidently predicted as class 1 despite the true distribution being unbiased. (c) In contrast, our approach accurately predicts equal likelihood within clusters. (d) Compared to training with CE loss, our approach accurately matches the true distribution as seen by the significantly lower KL-divergence values w.r.t. the true-distribution while utilizing an order of magnitude fewer samples.

with simple L2 regularization to show that our objective goes beyond such simple regularization schemes.

In these experiments, we find evidence that our method

1. **Induces smoothness in the conditional distributions.** Fig. 4 shows a setting where both CE and its L2 regularized version obtain similar test losses but differ drastically in the label assignments compared to our objective. Specifically, Fig. 4(b) shows the conditional probability $P(y = 1|x)$ for test points from a CE+L2 trained model. Even for the L2 regularized model, there is significant overfitting with some regions of the rightmost Gaussian confidently predicted as class 1. In contrast, our approach shown in Fig. 4(c) results in near uniform probability between classes 1 and 2 within the cluster.
2. **Acts as a regularizer.** Since our objective enforces that neighboring data points have similar output distributions, over-fitting by making overly confident predictions is strongly penalized. This is evidenced by the low KL-divergence w.r.t. to underlying data distribution that our objective achieves (see Fig. 4(d)).
3. **Improves sample efficiency.** As shown in Fig. 4(d), even with fewer samples as compared to Maximum Likelihood training (with and without L2 regularization), our model is able to more accurately model the true data-distribution as evidenced by the significantly lower KL divergence w.r.t. to the data-generating model.

Further details are provided in the supplement.

4.2. Attribute Prediction.

Datasets and Models. We now replicate the synthetic setup in a multi-label image attribute prediction setting on two real world datasets – Animals with Attributes (AWA; Xian et al. (2017)) and Caltech UCSD Birds 200-2011 (CUB; Wah et al. (2011)). Specifically, we randomly sub-sample from the set of all positive attributes for an image and evaluate the performance of models based on their ability to recover all annotated attributes for each image.

	Method	# observed labels	Average Precision@k		
			@1	@5	@10
AWA	CE+L2	1	46.82	51.03	54.18
	Ours	1	**49.23**	**53.46**	**57.12**
	CE+L2	20%	52.74	57.48	63.71
	Ours	20%	**56.79**	**61.54**	**66.28**
CUB	CE+L2	1	27.10	31.40	35.62
	Ours	1	**29.32**	**33.19**	**38.83**
	CE+L2	20%	32.42	35.94	39.21
	Ours	20%	**35.64**	**38.20**	**43.01**

Table 1. In the multi-label classification with missing labels setting, we observe that our proposed method outperforms standard cross-entropy training with $\sim 3\%$ improvements when k=10 on both AWA and CUB datasets with both just one randomly sampled label and observing 20% of the annotations

While AWA contains $\sim 30K$ images across 50 categories and 85 attributes, CUB is much sparser with $\sim 11K$ images across 200 categories and 312 attributes. We report results under two aggressive missing-label settings; sampling either only a single attribute or 20% of the annotated attributes for each image. For both L2 regularized cross-entropy (CE+L2) and our loss, we use the pre-final layer activations of Resnet-152 (He et al., 2016) as the image-representation and train a two-layered MLP, optimized using Adam (Kingma & Ba, 2015) with a learning rate of $1e{-}5$ and a batch size of 64. Early-stopping was used to pick the best set of parameters.

Results. Unlike the toy setting in Section 4.1, the data generating distribution is unknown and so, we cannot evaluate using KL-divergence w.r.t. it. As is common to the multi-class with missing labels setting, we use `Average Precision@k` – that measures the number of correct annotations present in the top-k ranked predictions. From

Table 1 it can be gleaned that our approach outperforms standard cross-entropy training in both settings (just one randomly sampled label and 20% of the annotated labels) and on both AWA and the much harder, CUB datasets. For instance, our method achieves an improvement of $\sim 3\%$ when evaluated on the top-10 retrieved predictions.

4.3. Visually Grounded Language Generation

Datasets. We report results on standard image-captioning datasets – Flickr-8k (Hodosh et al., 2013), Flickr-30k (Young et al., 2014) and COCO (Lin et al., 2014a). We use standard splits (Karpathy & Fei-Fei, 2015) of size 1000 to report results on the first two and a test split of size 5000 for the COCO dataset. Futher, to mimic the problem of missing labels we train only on a single (arbitrarily chosen) caption while evaluating on all 5 captions. This approach helps us to evaluate if our model learns to place beliefs on other unseen but 'correct' outputs while seeing only *uni-modal* training data. Owing to the small size of the PASCAL-50S dataset (Jas & Parikh, 2015), we only evaluate on it while using a model trained on COCO.

For VQG, We use three datasets built on a small subset of COCO, Flickr and Bing images (Mostafazadeh et al., 2016). We train on $\sim 2.5K$ images and report results on a test of size $\sim 1.5K$ for each dataset. To stay consistent with the captioning experients, we report retrieval numbers on a randomly chosen subset of 1000 images and their 5 corresponding questions from the test set.

Models. For both tasks, we train a model similar to Lu et al. (2017) that uses activations from an ImageNet pre-trained Resnet-152 (He et al., 2016) architecture as image-representations. For both captioning and VQG, the learnt LSTM model has one layer, 1024-dimensional hidden states and is optimized using Adam (Kingma & Ba, 2015) with a learning rate of $1e{-}4$. The similarities $\mathcal{K}_{ij}$ used to weigh the supervision from neighboring data-points are computed in a learnt space got by projecting the image-representations through a 2-layered MLP with 512 hidden units in each layer. As discussed in sec. 2.3, the learning rate for this transformation is is $10\times$ smaller compared to the LSTM parameters.

Ablations Recall that in Section 2.2 we described two variants of our approach for sequence prediction:

1. **Caption-Transfer-without-Attention**: where we transfer entire captions from neighboring images, and
2. **Caption-Transfer-with-Attention**: where we use attention models to selectively weigh relevant portions of the neighbor's caption.

Baselines. In addition to standard maximum likelihood training we also compare to two ablations of our method:

	Method	Oracle Metrics @20			distinct 4-grams	Recall$_5$ @100
		CIDEr	SPICE	METEOR		
Flickr-8k	MLE	0.5072	0.1564	0.1553	2205	0.71
	(Rennie et al., 2017)	**0.5272**	0.1509	0.1498	1834	0.74
	(Dai et al., 2017)	0.4982	0.1598	0.1420	1730	0.72
	Caption-Transfer-Without-Attention	0.5181	0.1561	0.1565	**2503**	0.89
	Caption-Transfer-With-Attention	0.5240	**0.1620**	**0.1614**	2498	**0.95**
Flickr-30k	MLE	0.6729	0.1642	0.1723	1920	1.30
	(Rennie et al., 2017)	0.6832	0.1520	0.1689	1824	1.36
	(Dai et al., 2017)	0.7120	0.1692	0.1752	1730	1.34
	Caption-Transfer-Without-Attention	0.7180	0.1721	0.1794	1822	1.52
	Caption-Transfer-With-Attention	**0.7246**	**0.1802**	**0.1843**	**2101**	**1.63**
COCO	MLE	0.8014	0.2132	0.2245	4218	1.45
	augment	0.8294	0.2147	0.2331	3766	1.49
	no-refine	0.8316	0.2182	0.2304	4117	1.63
	(Rennie et al., 2017)	0.8410	0.2013	0.2272	3988	1.53
	(Dai et al., 2017)	0.8120	0.2117	0.2340	4011	1.49
	Caption-Transfer-Without-Attention	0.8398	0.2165	0.2378	4128	1.78
	Caption-Transfer-With-Attention	**0.8422**	**0.2210**	**0.2405**	**4270**	**1.84**

Table 2. While reporting standard task-specific and diversity metrics, we observe that our methods outperform standard MLE and the baselines on all three datasets Flickr-8k, Flickr-30k and COCO on the retrieval task that measures multi-modal output mappings. Further, note that the task-specific metrics like CIDEr and SPICE are generally lower since we train using only one caption.

1. **augment** – captions of neighboring images are directly appended as ground truth to create a larger training set (corresponds to setting $\mathcal{K}_{ij} = 1$). Improvements on this setting indicate the advantage of both softly-enforcing the neighborhood as well as learning the representation space for computing similarities.
2. **no-refine** – Unlike the full setting, the representation is held fixed and is not refined from the generic representations to specialize for the task at hand. Improvements over this baseline denotes the advantages of jointly learning the representation space (and adapting the neighborhood) apart from transfering supervision.

Further, we compare to two other strong methods that do not employ MLE — (Dai et al., 2017) use adversarial training to distinguish between human and generated captions and (Rennie et al., 2017) directly optimize for a task-specific metric using policy gradients with a novel variance reduction baseline.

Evaluation Metrics. We evaluate the sequence generation models using the following metrics that each evaluate for certain desirable properties of the model –

1. *Oracle Metrics.* Each output is evaluated against the reference sequences using standard captioning metrics like CIDER or SPICE. Following previous works (Guzman-Rivera et al., 2014; Lee et al., 2016; Snell & Zemel, 2017) that consider ambiguous tasks, we evaluate the decoded lists using *oracle* metrics that report the best output in the list – mimicking an 'oracle' user that selects the most suitable option for a downstream task.
2. *Diversity Metrics.* Apart from producing high-quality captions, linguistic diversity in the decoded lists provides a good signal for the multi-modal nature of the learnt model. Similar to (Li et al., 2015), we measure di-

	Method	Oracle Metrics @20			distinct 4-grams	Recall_5 @100
		CIDEr	SPICE	METEOR		
Flickr	MLE	0.3510	0.1201	0.1273	1294	0.35
	(Rennie et al., 2017)	0.3822	0.1240	0.1298	1350	0.38
	(Dai et al., 2017)	0.3572	0.1286	0.1287	1258	0.34
	Caption-Transfer-Without-Attention	0.3720	0.1292	0.1392	1388	0.46
	Caption-Transfer-With-Attention	**0.3827**	**0.1445**	**0.1462**	**1395**	**0.51**
COCO	MLE	0.3233	0.1182	0.1245	967	0.33
	(Rennie et al., 2017)	0.3485	0.1224	0.1209	1104	0.35
	(Dai et al., 2017)	0.3296	0.1215	0.1241	**1270**	0.38
	Caption-Transfer-Without-Attention	0.3471	0.1262	0.1276	1192	0.43
	Caption-Transfer-With-Attention	**0.3506**	**0.1282**	**0.1304**	1220	**0.49**
Bing	MLE	0.2890	0.1202	0.1309	755	0.37
	(Rennie et al., 2017)	**0.3681**	0.1225	0.1287	910	0.38
	(Dai et al., 2017)	0.3224	0.1199	0.1256	937	0.42
	Caption-Transfer-Without-Attention	0.3355	0.1289	0.1296	984	0.49
	Caption-Transfer-With-Attention	0.3410	**0.1314**	**0.1336**	**1021**	**0.55**

Table 3. We find that our methods outperform the baselines on all three datasets Flickr, COCO and Bing datasets on the retrieval task that evaluates ability of the model to make multi-modal output maps. Similar to image-captioning, note that the task-specific metrics are generally lower as we train using only one question.

versity in the decoded lists by reporting the number of distinct n-grams (normalized by sentence length) present in the decoded lists.

3. *Retrieval Metrics.* A drawback of the first two metrics is that they also depend on the inference procedure used to decode the output lists (*e.g.* beam search). Therefore, we directly evaluate the beliefs placed by the model on different outputs in a retrieval setting where a pool of human-annotations are ranked based on their log-probability under the model for a given image. Then, we compute $\texttt{Recall}_\texttt{m}\texttt{@k}$ that evaluates for – *the average number of the* m *ground truth captions that were present in the top-k retrieved sequences*.

Results. For both captioning and VQG, we decode output lists for all methods using beam search with a beam size of 20. As can be seen from Table 2 and Table 3, both variants of our approach outperform cross-entropy training on all three standard image-captioning metrics. Excluding Flickr-30K for captioning and Bing for VQG, our approach performs the best in terms of output quality (as evidenced by higher oracle numbers). Further, our approach achives the best performance on both diversity and retrieval metrics indicative of the multi-modal mapping learnt by the model on both tasks of interest. Additionally, our methods outperform both *hard* and *no-refine* ablations of our method – we only show the performance for captioning on COCO owing to space constraints and provide the rest in the supplement.

5. Discussion

Sample Efficiency on Captioning. We perform sample-efficiency experiments similar to those in the toy-setting for captioning on the Flickr-8k dataset in two ways – 1) by gradually increasing the number of unique images used from 1000 to 8000 while still using only one caption annotation. 2) by gradually increasing the number of additional ground truth captions used from 1 to 5. As seen in Fig. 5 (left), our approach obtains a Recall_5@100 score of 1.58 using only $5K$ images which is very close to what is obtained using all $8k$ images (=1.60) in the first case. In the second case, we observe in Fig. 5 (right) that using only 3 captions per image (1.65) nearly obtains the same retrieval score as using all the 5 captions (=1.66). This demonstrates that our approach can lead to efficient learning of the true distribution even in data-sparse regimes.

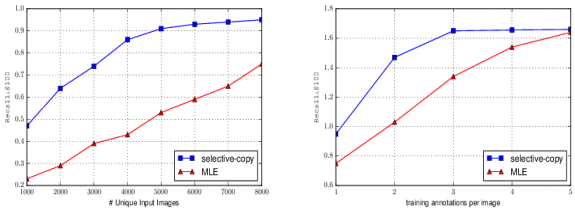


Figure 5. Our method is sample efficient compared to standard cross-entropy training. On the Flickr8k image-captioning task, our method performs comparably to CE with the full dataset while only using $\sim 3K$ images **(left)**. Similar trends exist as the number of annotations per image is increased **(right)**.

While the primary focus of our work is to learn multi-modal mappings even with access to uni-modal datasets, we observe that our proposed method performs competitively when trained and evaluated under standard captioning settings *i.e.* all 5 captions are used during training and *one* best output is evaluated (as against using oracle metrics). For instance, our method achieves a METEOR and CIDEr score of 0.28 and 1.14 respectively, slightly outperforming Lu et al. (2017) (0.27 and 1.09) on the COCO-captioning task. We observe similar trends on question-generation and include detailed results in the supplement.

6. Conclusion

In this work, we propose a novel objective that incorporates the inductive bias that the outputs of neighboring data points can be used to provide additional supervision especially when obtaining exhaustive annotations is expensive or worse, intractable. The proposed objective allows the model to place beliefs on multiple plausible outputs while still observing only one annotation per input. We first study the properties of our method on a synthetic dataset where the underlying data-distribution is known allowing us to control the difficulty of the experiments and directly evaluate the learnt posteriors. Further, we replicate this toy setting on a real-world multi-label prediction problem using standard attribute datasets. Finally, we show that our approach leads to better quality outputs with higher diversity on two well-established visually grounded language-generation tasks – captioning and question generation. We observe that our approach outperforms various ablations and baselines on both tasks on the various evaluation metrics used.

References

Anderson, P., Fernando, B., Johnson, M., and Gould, S. Spice: Semantic propositional image caption evaluation. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2016.

Batra, D., Yadollahpour, P., Guzman-Rivera, A., and Shakhnarovich, G. Diverse M-Best Solutions in Markov Random Fields. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2012.

Bucak, S. S., Jin, R., and Jain, A. K. Multi-label learning with incomplete class assignments. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Chen, M., Ding, G., Zhao, S., Chen, H., Liu, Q., and Han, J. Reference based lstm for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Chorowski, J. and Jaitly, N. Towards better decoding and language model integration in sequence to sequence models. In *Interspeech*, 2017.

Dai, B., Lin, D., Urtasun, R., and Fidler, S. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

Deng, J., Ding, N., Jia, Y., Frome, A., Murphy, K., Bengio, S., Li, Y., Neven, H., and Adam, H. Large-scale object classification using label relation graphs. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

Devlin, J., Gupta, S., Girshick, R., Mitchell, M., and Zitnick, C. L. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015.

Guzman-Rivera, A., Batra, D., and Kohli, P. Multiple Choice Learning: Learning to Produce Multiple Structured Outputs. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

Guzman-Rivera, A., Kohli, P., Batra, D., and Rutenbar, R. Efficiently enforcing diversity in multi-output structured prediction. In *Artificial Intelligence and Statistics*, pp. 284–292, 2014.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Hodosh, M., Young, P., and Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47: 853–899, 2013.

Inan, H., Khosravi, K., and Socher, R. Tying word vectors and word classifiers: A loss framework for language modeling. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Jain, U., Zhang, Z., and Schwing, A. Creativity: Generating diverse questions using variational autoencoders. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Jas, M. and Parikh, D. Image specificity. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Karpathy, A. and Fei-Fei, L. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Kingma, D. P. and Ba, J. ADAM: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

Lee, S., Purushwalkam, S., Cogswell, M., Ranjan, V., Crandall, D. J., and Batra, D. Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems (NIPS)*, 2016.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A diversity-promoting objective function for neural conversation models. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies (NAACL HLT)*, 2015.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollar, P., and Zitnick, C. L. Microsoft COCO: Common objects in context, 2014a.

Lin, V. X., Singh, S., He, L., Taskar, B., and Zettlemoyer, L. Multi-label learning with posterior regularization. In *NIPS Workshop on Modern ML and NLP*, 2014b.

Lu, J., Xiong, C., Parikh, D., and Socher, R. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Mostafazadeh, N., Misra, I., Devlin, J., Mitchell, M., He, X., and Vanderwende, L. Generating natural questions about an image. *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2016.

Mun, J., Cho, M., and Han, B. Text-guided attention model for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Pereyra, G., Tucker, G., Chorowski, J., Kaiser, Ł., and Hinton, G. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017.

Perozzi, B., Al-Rfou, R., and Skiena, S. Deepwalk: Online learning of social representations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2014.

Prasad, A., Jegelka, S., and Batra, D. Submodular meets structured: Finding diverse subsets in exponentially-large structured item sets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014.

Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., and Goel, V. Self-critical sequence training for image captioning. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Shetty, R., Rohrbach, M., Hendricks, L. A., Fritz, M., and Schiele, B. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*, 2017.

Snell, J. and Zemel, R. S. Stochastic segmentation trees for multiple ground truths. 2017.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. Rethinking the inception architecture for computer vision. In *arXiv preprint arXiv:1512.00567*, 2015.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Verma, Y. and Jawahar, C. Exploring svm for image annotation in presence of confusing labels. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2013.

Vijayakumar, A. K., Cogswell, M., Selvaraju, R. R., Sun, Q., Lee, S., Crandall, D., and Batra, D. Diverse beam search: Decoding diverse solutions from neural sequence models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2018.

Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. The caltech-ucsd birds-200-2011 dataset. 2011.

Weston, J., Ratle, F., Mobahi, H., and Collobert, R. Deep learning via semi-supervised embedding. In *Neural Networks: Tricks of the Trade*, pp. 639–655. Springer, 2012.

Xian, Y., Lampert, C. H., Schiele, B., and Akata, Z. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2015.

Yang, Z., Cohen, W., and Salakhutdinov, R. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2016.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Proceedings of the Annual Meeting on Association for Computational Linguistics (ACL)*, 2014.

Yu, H.-F., Jain, P., Kar, P., and Dhillon, I. Large-scale multi-label learning with missing labels. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2014.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning requires rethinking generalization. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Zhu, X. Semi-supervised learning literature survey. In *Technical Report 1530, University of Wisconsin, Madison*. University of Wisconsin, Madison, 2005.

Zhu, X., Ghahramani, Z., Lafferty, J., et al. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2003.