Active Learning from Noisy Tagged Images

M. Ehsan Abbasnejad
ehsan.abbasnejad@adelaide.edu.au
Anthony Dick
anthony.dick@adelaide.edu.au
Qinfeng Shi
qinfeng.shi@ieee.org
Anton van den Hengel
anton.vandenhengel@adelaide.edu.au

Australian Institute for Machine Learning (AIML),
The University of Adelaide,
Adelaide, South Australia,
5005 Australia

Abstract

Learning successful image classification models requires large quantities of labelled examples that are generally hard to obtain. On the other hand, the web provides an abundance of loosely labelled images, i.e. tagged in websites such as Flickr. Although these images are cheap and massively available, their tags are typically noisy and unreliable. In an attempt to use such images for training a classifier, we propose a simple probabilistic model to learn a latent *semantic* space from which deep vector representations of the images and tags are generated. This latent space is subsequently used in an active learning framework based on adaptive submodular optimisation that selects informative images to be labelled. Afterwards, we update the classifier according to the importance of each labelled image to best capture the information they provide. Through this simple approach, we are able to train a classifier that performs well using a fraction of the effort that is typically required for image labelling and classifier training.

1 Introduction

In recent years, machine learning and computer vision communities have witnessed a great success in large-scale image classification. In public challenges such as ImageNet [4, 22], millions of *labelled* images are used to train algorithms such as deep Neural Networks that perform as well as humans. However, the success of such algorithms greatly depends on the availability of such training examples that are expensive and labour intensive. Furthermore, these labels are specific to the domain for which they were collected for.

One way to minimise the manual labelling effort is to use abundantly available images from online services such as Google Image Search, Flickr or Instagram that can be collected cheaply and massively. These images are not completely unlabelled: they are weakly identified by the user tags. However, these tags or search terms are often unreliable and noisy. People may use various words to refer to the same concept or conversely, the same tag may refer to various concepts. For example, images tagged as "tank" may refer to an armed vehicle, clothes, water tank, aquarium, etc. Thus, in this paper we ask *how to effectively utilise the tagged images from the web to learn a classifier with minimum manual labelling effort?*

To that end, we tackle three sub-problems: (1) how to disambiguate tags and relate them to the visual content of the image? (2) how to select images so that human-provided labels yield maximum information for the classifier and (3) how to update a classifier so that with minimum labels, the best classifier can be learnt? For (1), we use the deep representation of images and tags and find the semantic space in which these representations have the highest correlation. This approach is as if the image and tags are conditionally generated from this semantic space and in turn ensures semantically related images are clustered in a space with lower dimensions. For (2), we find "uncertain" labels in the semantic space to form an initial "belief" over decision boundary for an *active* classifier. To update this belief, we iteratively query for manual labelling of the most "informative points" and refine the decision boundary. We compare various measures for choosing these informative points. For (3), we formulate a probabilistic framework that uses *adaptive submodularity* [7, 14] in iterative label queries and online update of the classifier. This online update ensures more informative points contribute more to the change in the decision boundary.

Ultimately, the objective is to use a constant number of queries for manual labelling to train the best possible classifier. This active classifier should perform as "similar" as possible to a classifier trained on a complete labelled set if we could afford to do so. Under submodularity conditions, the greedy algorithm that our iterative approach represents is able to find the near-optimal subset of images for labelling incurring least difference to the optimal classifier. As such, we believe our approach based on theoretically sound motivations, lays the foundation for better algorithms that require minimal labelling effort to achieve good performance using noisy tagged images.

Our contributions in this paper are: (1) we provide an efficient general framework for image label gathering that requires fraction of conventional manual labelling effort; (2) we devise a probabilistic framework using deep Neural Networks on both images and tags to provide a semantic text-image embedding; (3) we provide simple alternatives to mutual information that are much more efficient for this purpose. We empirically examine our approach on various datasets including the one that we have collected using the ambiguous tag "tank" from Flickr. We believe our simple approach opens new avenues for research in using publicly available images to building new datasets, particularly for specialised domains.

2 Related work

Learning from a combination of correctly labelled and tagged images has been previously investigated in computer vision. For instance, web images can be used to augment the already manually labelled training sets for better performance (e.g. [1, 8, 19]) in line with semi-supervised learning ideas. Similarly, [28] proposed ConceptLearner which relies on the tags with the same word to be treated as positive examples, and a randomly selected subset of other tags as negative, to train an SVM model. Image-Text embedding (with noise-free labels) have been successfully used for zero-shot learning. For instance, DeViSE proposed by [6] and its extension [21] use deep features in a hinge loss formulation to find a semantic embedding.

A distinctive feature of our approach is that we can actively update our model by querying correct labels. Active learning is a well-studied field of research in machine learning (interested readers can refer to [24] for an introduction). Submodular optimisation as a method for subset selection has been naturally used for active learning (e.g. [6, 7, 15]). We use active learning with noisy tags in a semantic space that has fewer parameters to learn.

3 Active Learning from Tags

Assume we are given images $\mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathcal{X}$ with their corresponding set of tags, $\mathbf{z}_1, \ldots, \mathbf{z}_n$ where $\mathbf{z}_i \in \{0,1\}^q$ is a vector of binary variables for image i, with each variable representing the presence of one of q tags. The objective of our approach is to utilise the noisy tags to select a subset of images to be labelled to produce a set $D \subset \mathcal{D}$ with $|D| \ll n$ which does not require all the images to be labelled. The size of this set $|D|$ is pre-determined and we denote by C. The performance of a classifier trained on D has to be as similar as possible to a classifier trained on the fully labelled set $\mathcal{D}$ (if we had access to it). Formally, we have

$$D = \arg\min_{D' \subset \mathcal{D}} \quad g_\ell(\mathcal{D}) - g_\ell(D'),$$

where g_ℓ is some *gain function* defined on a given set, D' is a subset of $\mathcal{D}$ and *oracle* function $\ell : \mathcal{X} \to \{1, \ldots, m\}$ provides the correct label (it can ask from an expert for a label). Intuitively, we select a subset of images to be labelled that provides the maximum gain. This gain is closest in value to the gain we could have obtained if we had all the images labelled.

3.1 Image-Tag Embedding

Our proposed approach is summarised in Figure 1. In the first three steps, the common semantic embedding is found. This semantic embedding is where the deep representation of the images from a Convolutional Neural Networks (CNNs) (e.g. AlexNet [17], VGG [26] or ResNet [11]) fine-tuned for the given tags (to learn a discriminative feature) and the deep tag representation from the Word2Vec [20], have the highest correlation. In this semantic space, the concepts are clustered (e.g. military, clothes and aquarium), although represent uncertain decision boundary between potential labels. In the subsequent three steps of Figure 1, we actively update our classifier network where the informative samples are identified and their corresponding label is queried. The number of points queried is C.

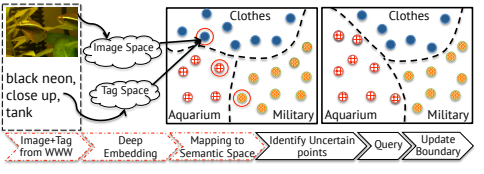

Figure 1: The sequence of steps in our approach: first three blocks (in dashed red) are done once and the last three are performed iteratively. The semantic space is found using the deep representations where semantic concepts are clustered (clothes, military and aquarium in this example). Subsequently, the algorithm queries the uncertain points (in red circles) at each iteration and updates the decision boundary of the classifier (black dashed lines).

Let's denote by $\Phi_{\mathbf{x}_i} \in \mathbb{R}^{d_1}$ image $\mathbf{x}_i$'s and by $\Psi_{\mathbf{z}_i} \in \mathbb{R}^{d_2}$ tag $\mathbf{z}_i$'s representations obtained from respective networks (we drop their corresponding parameters for brevity). Our goal is then to find a common space from which tag and image vectors are generated. We consider this common latent space as the d-dimensional latent space found using Variational Auto-encoder (VAE) [13] that encodes the image representation to the latent space and reconstructs samples to represent tags. In other words, our approach learns to encode the image to the latent representation and decode the latent representation to a tag.

Let θ_i represent the embedding of the i's image in the latent concept space. The set of points obtained from this image encoding yields uncertain *semantic concepts* (potential labels for images). Subsequently, semantic labels are clustered with the following likelihood,

$$p(\theta_{1\ldots n}|m) = \prod_{k=1}^{m} \prod_{i=1}^{n} \mathbb{I}[\theta_i \in \mathcal{C}_k] \times \exp(-\|\mathbf{c}_k - \theta_i\|^2) \tag{1}$$

where m is the number of semantic concepts, $\mathcal{C}_k$ is the set of encoded images belonging to semantic concept k and $\mathbf{c}_k$ representing its centroid. We maximise this likelihood using EM to find the uncertain *semantic* label $\hat{y}_k \in \{1,\dots,m\}$ for each θ_i. These semantic labels are used to train a classifier (i.e. a neural network with softmax output). We use this classifier trained on these labels to form an initial belief for the true labels that will be queried.

3.2 Iterative Labelling

Once we found an initial belief over the decision boundary (discrimination between various potential labels) and the uncertain semantic labels, we match these semantic labels with the correct labels by an oracle's (e.g. human) intervention and refinement. This is done by selecting the points in the semantic space that improve the given classifier the most. We turn to adaptive submodularity where we can use the prior belief we obtained from the tag-image embedding to find the subset of images for which the correct labels provide best classifier performance. This classifier is *active* as it queries and improves by oracle-provided labels.

Adaptive submodular optimisation is used for subset selection where there is inherent uncertainty (label uncertainty in our approach). It is shown that a greedy algorithm that selects each element sequentially based on the distribution of the labels, produces the near-optimal subset. Upon selecting a new instance, we update the belief over the decision boundary for subsequent iteration (shown in *Classifier Update* section). Formally, at each step t, we pick an instance that maximises *conditional expected marginal benefit*:

$$\Delta(\Phi_{\mathbf{x}_i}|D_{t-1},\theta_i) = \mathbb{E}_{p(y_i|D_{t-1})}\left[g(D_{t-1}\cup\Phi_{\mathbf{x}_i}) - g(D_{t-1})\right]$$

where g is the gain function, y_i is the correct label and with abuse of notation, D_{t-1} is the set of image embeddings from previous iteration. The semantic labels found in the previous section are used as the prior belief and will be updated with the correct labels using the oracle's feedback. A common candidate for the gain function is the entropy of the predictions (due to conjugate duality of the partition function in softmax and the entropy). At each iteration we select a point with the following criteria:

$$\Phi_{\mathbf{x}_i}{}^* = \arg\max_{\Phi_{\mathbf{x}_i}\in D'} \Delta(\Phi_{\mathbf{x}_i}|D_{t-1},\theta_i) = \mathbb{E}_{p(y_i|\Phi_{\mathbf{x}_i},D_{t-1})}\left[H(y_1,\dots,y_m|D_{t-1}\cup\Phi_{\mathbf{x}_i},\theta_i) - H(y_1,\dots,y_m|D_{t-1},\theta_i)\right]. \quad (2)$$

A greedy algorithm selects $\Phi_{\mathbf{x}_i}{}^*$ at each step with maximum entropy. This greedy approach is very similar to that of [9], although that was not formalised as a submodular problem. However, one drawback of this approach is that it is computationally inefficient. This is because we need to update the classifier with negative of the loss at that point for the given label and evaluate the expected entropy. Thus, effectively evaluating the impact of including a point in the training examples to update the model. Instead we consider three simpler methods for selecting the optimal point as approximations for $\Delta(\Phi_{\mathbf{x}_i}|D_{t-1},\theta_x)$ that estimate the differential entropy in Equation 2:

- **Case 1: Entropy:** Using the entropy of the prediction $H(y_i|\Phi_{\mathbf{x}_i},\theta_i,D_{t-1})$ where the entropy of a label is used as a surrogate for the joint probability of labels.

- **Case 2: Marginal:** We can use the difference of 2 labels with highest probability to estimate the entropy [24]. The intuition behind this is that, a high entropy prediction has a small difference between the estimated probabilities in the joint, i.e. $p(\dot{y}|\Phi_{\mathbf{x}_i},D_{t-1},\theta_x) - p(\dot{\dot{y}}|\Phi_{\mathbf{x}_i},D_{t-1},\theta_x)$ where $\dot{y} = \arg\max_y p(y|\Phi_{\mathbf{x}_i},D_{t-1})$ and $\dot{\dot{y}}$ is the second best. Therefore, this measure is a suitable surrogate for Equation 2.

- **Case 3: Gradient:** We directly approximate a lower bound on the entropy with the softmax classifier. We estimate an update for the weight vector and *pessimistically* select the point to be queried for its correct label. We rewrite Equation 2, ${\Phi_{\mathbf{x}_i}}^* = \arg\max_{\mathbf{x}_i \in D'} \mathbb{E}_{p(y_i|\Phi_{\mathbf{x}_i}, D_{t-1})} \left[-\frac{\alpha_i v_i}{\langle \alpha, \mathbf{v} \rangle} \log \left(\frac{\alpha_i v_i}{\langle \alpha, \mathbf{v} \rangle} \right) \right]$ where $v_j = \exp(\mathbf{w}_{\mathbf{y_j}} \theta_i)$, $\mathbf{v} = [v_1, \ldots, v_m]$ and $\alpha_j = \exp(\delta^\top_{\mathbf{x}_i, y_j} \Phi_{\mathbf{x}_i})$, $\alpha = [\alpha_1, \ldots, \alpha_m]$. From Cauchy-Schwarz inequality we know $\langle \alpha, \mathbf{v} \rangle \leq \|\alpha\| \|\mathbf{v}\|$, thus we have,

$$\mathbb{E}_{p(y_i|\Phi_{\mathbf{x}_i}, D_{t-1})} \left[-\frac{\alpha_i v_i}{\|\alpha\| \|\mathbf{v}\|} \log \left(\frac{\alpha_i v_i}{\langle \alpha, \mathbf{v} \rangle} \right) \right] \geq \mathbb{E}_{p(y_i|\Phi_{\mathbf{x}_i}, D_{t-1})} \left[-\frac{\alpha_i v_i}{\|\alpha\| \|\mathbf{v}\|} \log \left(\frac{\alpha_i v_i}{\|\alpha\| \|\mathbf{v}\|} \right) \right].$$

  for which the optimal value is at $\frac{\alpha_i}{\|\alpha\|} \propto \frac{\|\mathbf{v}\|}{v_i}$. For fixed value of α_i, picking v_i that is inversely proportional will maximise the lower bound on the entropy measure in Equation 2. Therefore at each iteration, we pick ${\Phi_{\mathbf{x}_i}}^*$ with minimum $\max_i \frac{\|\mathbf{v}\|}{v_i}$. This value is easily computed for each image from the predictions from the previous iteration.

In the subsequent section, we use these correct labels to update the classifier so that important points have a larger impact on the decision boundary.

3.3 Classifier Update

At each iteration of the algorithm we query the oracle for new correct labels and update the classifier. Inspired by the passive-aggressive online algorithm [3], we update the decision boundary such that it changes more by the points for which classification is harder. Hence, $p(\mathbf{w}^{\text{new}}_{y_i}, y_i | \Phi_{\mathbf{x}_i}, D_{t-1}, \mathbf{w}_{y_i}, \theta_i) = p(\mathbf{w}^{\text{new}}_{y_i} | \mathbf{w}_{y_i}, D_{t-1}) p(y_i | \mathbf{w}_{y_i}, \Phi_{\mathbf{x}_i}, \theta_i)$ with the convention that $p(\mathbf{w}|D_0, \theta_i) = p(\mathbf{w}|\theta_i, \hat{\mathbf{Y}})$ is the prior and $\hat{\mathbf{Y}}$ denotes the potential uncertain labels for all the images. We formulate the problem as

$$\max_{\mathbf{w}^{\text{new}}_{y_i}} p(\mathbf{w}^{\text{new}}_{y_i} | \mathbf{w}_{y_i}, D_{t-1}), \quad \text{s.t. } p(y_i | \mathbf{w}^{\text{new}}_{y_i}, \Phi_{\mathbf{x}_i}, \theta_i) = 1 - \varepsilon,$$

for a positive value $\varepsilon \geq 0$. We consider a normal distribution centred at the previous value of the weight parameter and enforce the prediction value after update to be close to one. Then the Lagrangian is

$$\frac{1}{2} \|\mathbf{w}^{\text{new}}_{y_i} - \mathbf{w}_{y_i}\|^2 + \gamma_i \left(1 - p(y_i | \mathbf{w}^{\text{new}}_{y_i}, \Phi_{\mathbf{x}_i}, \theta_i) - \varepsilon\right). \tag{3}$$

For the softmax classifier, we have a typical gradient update for weights $\mathbf{w}^{\text{new}}_{y_i}$. In this formalisation though, we compute the learning rate in a closed form allowing for each correct image label to have a different effect on the decision boundary proportionate to its loss:

$$\gamma_i = \frac{1 - p(y_i | \mathbf{w}^{\text{new}}_{y_i}, \Phi_{\mathbf{x}_i}, \theta_x) - \varepsilon}{\|\nabla_{\mathbf{w}^{\text{new}}_{y_i}} p(y_i | \mathbf{w}^{\text{new}}_{y_i}, \Phi_{\mathbf{x}_i}, \theta_x)\|^2}. \tag{4}$$

which is derived from solving for $\mathbf{w}^{\text{new}}_{y_i}$ in Equation 3 and then replacing back into the equation to obtain γ_i.

The proposed approach is detailed in Algorithm 1. Note that since the semantic space has lower-dimensions, there are fewer parameters to learn and therefore is more efficient to train. In addition, even though we have used approximations, theoretical guarantees for the adaptive submodular functions still hold, albeit with looser bounds based on the approximation quality. These theoretical guarantees ensure the convergence indicating the samples selected using this greedy approach and the subsequent classifier is near-optimal.

Algorithm 1 Active Submodular Image-Tag Learning

Require: $\mathcal{D}, m \geq 2$
1: Find the projection of images/tags latent space and the embedding of images in this space θ_i
2: Find $\hat{y}_1, \ldots, \hat{y}_m$ from cluster assignments of $\theta_i, \forall i$
3: $\mathbf{w} = \arg\max_{\mathbf{w}'} \log\left(p(\mathbf{w}') \prod_{i=1}^{n} p(\hat{y}_i | \Phi_{\mathbf{x}_i}, \theta_i, \mathbf{w}')\right)$ // train softmax as a prior using semantic labels
4: $D = \emptyset$
5: **while** $|D| < C$ **do** // for some constant $C \ll n$
6: $\quad \Phi_{\mathbf{x}_i}{}^* = \arg\max_{\Phi_{\mathbf{x}_i} \in D'} \Delta(\Phi_{\mathbf{x}_i} | D_{i-1}, \theta_i)$
7: $\quad$ Ask for label y_i from the oracle
8: $\quad$ Update weights with γ_i given in Eq. 4 using y_i
9: $\quad D_i = D_{i-1} \cup \Phi_{\mathbf{x}_i}{}^*$
10: **end while**
11: **return** $\mathbf{w}, D$ // classifier weights and the labeled set

4 Experiments

In this section we examine how well our approach works for real-world problems, namely image classification on the ImageNet dataset, Flickr images, Cifar10 and Caltech 101. In particular, we are interested in investigating how well the clusters found in the semantic space from deep image and tag can help train a base classifier. Subsequently, we use the active learning discussed in this paper to improve the initial classifier from the uncertain labels. We employ three deep learning architectures (AlexNet [12], VGG [26] and Residual Network [11])) and fine-tuned them in an auto-encoder to reconstruct the tag vector representation obtained from Google's word2vec[1]. As such, we learn the latent space from which both images and tags are generated. This latent space characterises our semantic space. Subsequently, we perform EM on the likelihood term in Equation 1 to obtain the semantic labels $\hat{y}$ in the semantic space. Using these semantic labels and mapped images in the latent space, we train a simple one layer softmax classifier. We report our results on various semantic latent dimensions so that we can investigate its effect on the performance.

Once the initial decision boundary is determined, we perform iterative labelling where the images with highest gain (as in Equation 2 and Case 1-3) are selected to be labelled by an oracle and update our softmax classifier.

We compare the results of the proposed active learning approach with the following baselines: (a) random selection of images to be actively labelled (as done in a typical gradient descent update), (b) images with highest prediction variance and (c) least confident predictions. For (b), we treat the prediction of the softmax for each correct label y_i as parameters of a categorical distribution where the variance is defined as $p(y_i|\mathbf{x}_i)(1 - p(y_i|\mathbf{x}_i))$. The intuition is that we ask for the labels of points for which the classifier is uncertain as measured by the variance (as opposed to entropy). in addition, we compared our approach with (c) least confident predictions corresponding to selecting images whose prediction label has the smallest probability. This criterion results in picking the points closest to the margin for the softmax. We have compared our approach to the exact expected entropy which performs worse (both in terms of accuracy and time) than our approximations in case 1-3 and hence was not included in the paper (as in Algorithm 1). Moreover, we compared our approach with the "full classifier" for which we used the whole correctly labelled training set to train a softmax classifier. For the full classifier, we have determined its learning rate using cross-validation.

[1] https://code.google.com/archive/p/word2vec/

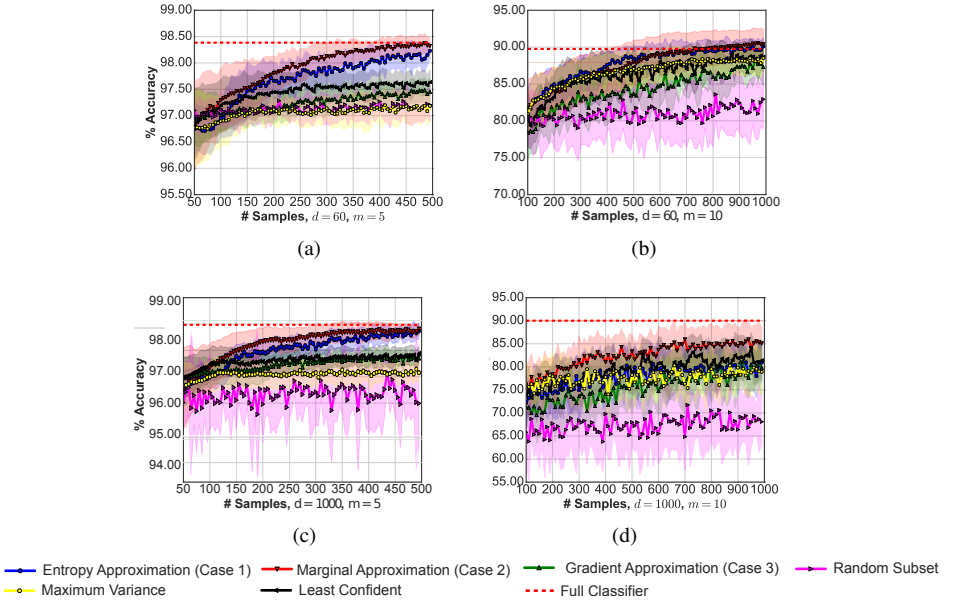

Figure 2: Results from running our approach on ImageNet dataset. Through active labelling using a fraction of images, we are able to perform comparably to the full classifier trained on the fully labelled dataset (in dashed line). As shown, lower dimension of the semantic space shows better improvement with each new observation. Our approaches are also consistently better than choosing a random subset.

Labelled Images: We use images from three popular image datasets, namely ImageNet [23], CIFAR-10 [16], Caltech101 [18], and treat their class labels as tags to evaluate the performance of our approach in this section. For ImageNet, we choose 5 and 10 breeds of dogs respectively with words such as "Maltese dog, Dalmatian, German shepherd, Siberian husky, St Bernard, Samoyed, Border collie, bull mastiff, chow, Afghan hound" that contain more than a single word. This helps us determine if the image-tag mappings in the semantic space are effective. Once we obtain the semantic concepts $\hat{y}$ and train our softmax model, subsequently we perform active learning. We use AlexNet to obtain the image representation. Subsequently, we find the semantic space where the tags and images space are highly correlated. This relation can also be found as a transformation of image vector to the tag vector as done in regression . As such, we first compared the mapping found using our approach with linear ridge regression (LRR) as shown in Table 1. By training a classifier using all the labelled images, we found our approach outperforms LRR. Since the embedding dimension is manually determined, it is more flexible. In addition, our approach better discovers the non-linear image-tag relation.

Method	Parameters	$k = 5$	$k = 10$
LRR	$d = 300, \lambda = 0$	85.5 ± 0.82	84.3 ± 0.1
	$d = 300, \lambda = 0.01$	90.53 ± 0.63	84.7 ± 0.12
	$d = 300, \lambda = 0.1$	95.46 ± 0.33	86.76 ± 0.1
	$d = 300, \lambda = 0.15$	95.4 ± 0.53	85.2 ± 0.13
Ours	$d = 60$	96.38 ± 0.63	90.57 ± 0.15
	$d = 120$	96.84 ± 0.34	88.92 ± 0.23
	$d = 1000$	98.85 ± 0.29	86.5 ± 0.25

Table 1: Ours vs. LRR embeddings for ImageNet

Using the clusters in the latent space, we obtain the semantic concepts $\hat{y}$. Using these semantic concepts we train an active classifier. Each experiment is carried out 5 times and their mean accuracy and standard error (as a shade with their corresponding colour) is summarised

#	Entropy	Marginal	Gradient	Least Confident	Variance	Random
CIFAR-10 with noisy tags, $m = 10, d = 500$						
150	89.61 ± 0.59	87.17 ± 0.57	89.66 ± 0.61	87.19 ± 0.80	88.04 ± 0.64	88.21 ± 0.78
600	90.01 ± 0.24	88.75 ± 0.40	90.07 ± 0.31	88.62 ± 0.45	89.65 ± 0.49	89.25 ± 0.62
900	90.19 ± 0.28	89.01 ± 0.31	90.25 ± 0.24	88.71 ± 0.32	90.21 ± 0.21	89.68 ± 0.55
Caltech-101 with noise, $m = 101, d = 500$						
505	88.11 ± 0.77	87.19 ± 0.67	87.36 ± 0.76	86.82 ± 0.73	87.10 ± 0.79	84.23 ± 0.72
2020	89.42 ± 0.47	88.63 ± 0.81	89.26 ± 0.59	89.35 ± 0.80	88.51 ± 0.32	86.07 ± 0.27
4040	89.62 ± 0.36	89.42 ± 0.24	89.40 ± 0.42	89.28 ± 0.33	88.65 ± 0.23	85.76 ± 0.97
5555	89.57 ± 0.35	89.37 ± 0.73	89.48 ± 0.32	88.94 ± 0.49	88.92 ± 0.70	87.59 ± 1.19

Table 2: Accuracy of our approach on CIFAR-10 and Caltech 101 with noisy tags

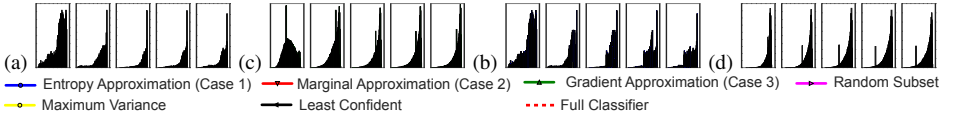

Figure 3: Distribution of the prediction entropy for Flickr dataset for $d = 120, m = 10$ with x-axis representing the entropy and the y-axis its frequency for the whole Flickr dataset. Each histogram represents the prediction entropy after adding 200 observations (1000 in total). It decreases with more observations. We compared (a) Entropy approximation, (b) Marginal Approximation, (c) Gradient Approximation and (d) Random subset.

in Figure 2. As shown, the embedding and the concept discovery step successfully initialises a classifier that performs well compared to the classifier that was training on correct labels. Subsequently, in the active learning step, the performance of the classifier improves as true labels are obtained. We observe that when the dimension of the data is smaller, as expected with fewer parameters to fit, new observations improve the performance of the algorithm faster. It is interesting to note that when the number of classes is smaller (i.e. 5 classes in our experiments) random subset selection performs comparably to the other approaches. However, as the dimensions or the number of classes increase, smarter approaches outperform random selection.

We performed similar experiment on CIFAR-10 using Residual network (ResNet-127) [10] and Caltech-101 using VGG-16 [25]. To replicate the noisy behaviour of tags in the embedding space we add Gaussian noise (mean zero and variance 1) to the vector representation of the tags (class labels in this case). This resembles the real-world situation where people often use various words to convey a concept rather than a unified term. We add this noise to 10% of training instances. This is equal to adding random related tags for each image and perform the mapping afterwards. The results of running this approach on the noisy classes are presented in Table 2. As shown, random selection performs almost the same albeit less poorly. Similarly, Least confident gain that evaluates the instances based on the distance to the margin is under-performing. As observed in Table 2, the Entropy and Gradient approximation outperform other counterparts. As seen in early stages where the number of labelled examples is smaller, Gradient approximation performs better and as the number of labelled instances increase the Entropy improves. we believe this is because the Gradient approximation is a better estimate of the joint entropy we in fact need.

Tagged Images from *Flickr*: As another test of our approach to train a classifier with minimum training examples, we use images available in Flickr Create Commons 100 Million Dataset[2] [27]. It contains 100 million multimedia urls from which 99.2% are images that are uploaded to Flickr between 2004 to 2014 published under Creative Commons license. For each image, user tags are also provided. We have selected a subset of images whose tags contain

[2] https://bit.ly/yfcc100md

the word *tank*[3]. Tank is a word with several meanings: military tank, water tank, tank top, helium tank, fish tank, etc. This dataset (called Tank dataset) contains 59303 images of which we use 80% for training and validation and 20% for testing. There are a total of 2299 distinct tags (tags occurring less than 50 times are omitted) . The histogram of the most frequent tags is shown in Figure 4. As can be seen, this dataset has a very diverse range of tags and images despite its limited scope. This diversity requires an approach like ours to iteratively and actively select the ones that are more informative and can lead to training of a better classifier. Similar to the ImageNet experiment, once we find the semantic space, we cluster the images to obtain semantic labels. The most frequent words in each cluster is shown in Table 3. Note the clusters are based on images but because images and words are correlated in this concept space, we are able to group the words too. The tag clusters found using this method is very different from using approaches such as Latent Dirichlet Allocation (LDA) [2] where only the frequency of words are considered for finding semantic relations between words.

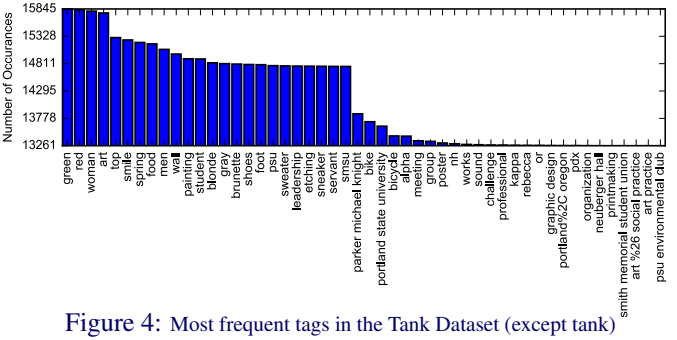

Figure 4: Most frequent tags in the Tank Dataset (except tank)

As seen in Table 3 using both tags and images, we can group semantically related image-tags in a cluster. Since we don't have true labels here, we obtain the labels actively from the prediction of VGG-16 (i.e. we use VGG-16 as an oracle for our active learner). We choose the highest predicted class from the set of labels for all the images in the dataset. Figure 5 shows sample images for each semantic label that the algorithm is querying in the 200th iteration using Gradient Approximation Criterion (Case 3) for $m = 5, d = 60$. Querying the labels will improve the prediction of the classifier.

Figure 3 shows the distribution of the average prediction entropy for each image in Flickr dataset. As can be seen, the entropy increases at first due to new observations that defy the classifier's belief over labels and then starts decreasing. Surprisingly, marginal approximation causes the entropy to increase more rapidly than other criteria. We believe this is due to its selection criterion, that selects the samples with largest difference in prediction values. The points selected this way, change the softmax more drastically leading to model uncertainty increase. Interestingly, it seems Gradient Approximation criterion (case 3) performs the best in reducing the entropy in this dataset.

Figure 5: Images from the cluster of semantic space with each column representing a concept cluster. As can be seen there are varied and yet sensible categories for $m = 5, d = 60$: water tank, Tank mountains, fish tank, army tank, and tank-tops.

Using 5 classes related to tank and only 1000 labeled examples we are able to achieve VGG's accuracy in predicting the labels for test images in this dataset.

[3] We will release the dataset and the code.

5 Discussion

In our approach, we used three measures to estimate the average joint entropy as the gain for querying the labels from the oracle. Each measure may perform better in a particular circumstance. The time complexity of each approximation after computing the predictive probabilities is $O(m)$ for all the measures. As such, they are equally fast.

military, mountain, bike, group, wall, men, shoes, california, food, smile
woman, green, red, art, top, blonde, brunette, shoes, foot, psu, sweater
england, poland, usa, photo, ww2, museum, fernando, stankuns, brasil, europe
tanker, tank, boat, aircraft, airplane, soldier, military aircraft, armour, military trucks, artillery
food, small, student, aqua, boat, sea, beach

Table 3: Frequent tags in the semantic space.

When the dimension of the latent space is smaller, the Marginal approximation (Case 2) performs better than the others. In larger dimensions, however, Entropy and Gradient approximation (Case 1, 3) perform better in exploiting the regions where the labelled samples increase the test accuracy better. This is because in Marginal approximation the difference of two top predictions decreases as the number of labels increase. Interestingly, when the number of classes and the latent space both increase, Entropy approximation performs better than the rest (in terms of accuracy).

On the other hand, when the number of labelled examples we intent to query is small, with smaller latent dimensions gradient and Marginal approximation perform better while for larger ones Gradient and Entropy are preferred. As the number of classes increases, the Entropy approximation is preferred as it leads to better and faster accuracy improvement.

For noisy semantic spaces, that is, when the number of noisy tags increase and involve various unrelated words, Entropy and Gradient approximation work the best. It seems, when the number of labelled instances is small Gradient approximation outperforms other approaches and as the number of labelled examples increases marginal improvement of Entropy approximation is better. We conjecture this is due to the derivation of the Gradient approximation that lower bounds the objective. As such, this lower bound is less susceptible to noise.

6 Conclusion

In this paper we proposed a simple probabilistic model for joint semantic learning of images and tags. We argued that using a latent variable model of the semantic space where the deep representation of images and tags are generated from, is able to capture clusters of images that are related. We subsequently used adaptive submodular optimisation to obtain correct labels for informative images with which we update our classifier. Empirically we showed this simple model is able to learn a classifier with a fraction of the labelling effort that is typically required. In particular, we experimented with a dataset built from the ambiguous word tank and utilised the co-occurrences of the words that refer to the same concept. Using our approach we will be able to build datasets with arbitrary granularity of labels. Based on the principles of adaptive submodularity we have theoretical guarantees that simple greedy algorithm is capable of selecting a subset of samples to be labelled that are near-optimal.

One advantage of taking vector representation for labels and discovering semantic space between tags and images is that we can easily generalise our algorithm for "zero-shot" learning where not all labels are observed during training.

References

[1] Alessandro Bergamo and Lorenzo Torresani. Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In J. Lafferty, C. Williams, J. Shawe-taylor, R.s. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 181–189. 2010.

[2] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003. ISSN 1532-4435.

[3] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, December 2006. ISSN 1532-4435.

[4] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[5] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Marc Aurelio Ranzato, and Tomas Mikolov. Devise: A deep visual-semantic embedding model. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2121–2129. Curran Associates, Inc., 2013.

[6] Akshay Gadde, Aamir Anis, and Antonio Ortega. Active semi-supervised learning using sampling theory for graph signals. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 492–501, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2956-9.

[7] Daniel Golovin and Andreas Krause. Adaptive submodularity: Theory and applications in active learning and stochastic optimization. *J. Artif. Int. Res.*, 42(1):427–486, September 2011. ISSN 1076-9757.

[8] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, chapter Improving Image-Sentence Embeddings Using Large Weakly Annotated Photo Collections, pages 529–545. Springer International Publishing, Cham, 2014.

[9] Yuhong Guo and Russ Greiner. Optimistic active learning using mutual information. In *Proceedings of the 20th International Joint Conference on Artifical Intelligence*, IJCAI'07, pages 823–829, San Francisco, CA, USA, 2007. Morgan Kaufmann Publishers Inc.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015.

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[13] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. 2014.

[14] Andreas Krause and Daniel Golovin. Submodular function maximization. In *Tractability: Practical Approaches to Hard Problems*. Cambridge University Press, February 2014.

[15] Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *J. Mach. Learn. Res.*, 9:235–284, 2008. ISSN 1532-4435.

[16] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[17] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In P. Bartlett, F.c.n. Pereira, C.j.c. Burges, L. Bottou, and K.q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1106–1114. 2012.

[18] R. Fergus L. Fei-Fei and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. *IEEE. CVPR 2004, Workshop on Generative-Model Based Vision*, 2004.

[19] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):985–1002, June 2008. ISSN 0162-8828.

[20] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[21] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S. Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. In *International Conference on Learning Representations (ICLR)*, 2014.

[22] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[23] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[24] Burr Settles. Active learning literature survey. Technical report, University of Wisconsin-Madison, 2010.

[25] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[26] K. Simonyan and A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv e-prints*, September 2014.

[27] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Commun. ACM*, 59(2):64–73, January 2016. ISSN 0001-0782.

[28] Bolei Zhou, Vignesh Jagadeesh, and Robinson Piramuthu. ConceptLearner: Discovering Visual Concepts from Weakly Labeled Image Collections. *Computer Vision and Pattern Recognition (CVPR)*, 2015.