Exploring SVM for Image Annotation in Presence of Confusing Labels

Yashaswi Verma
http://researchweb.iiit.ac.in/~yashaswi.verma/
C. V. Jawahar
http://www.iiit.ac.in/~jawahar/

CVIT
IIIT-Hyderabad
Hyderabad, India
http://cvit.iiit.ac.in

We address the problem of automatic image annotation in large vocabulary datasets. In such datasets, there exist three practical issues: (a) **Incomplete-labeling**: The training samples are not exhaustively tagged with *all* relevant labels from vocabulary. This is because while building a dataset, human annotators find some labels as "obvious" and miss them in the ground-truth. E.g., an image tagged with *"car"* might not be tagged with *"vehicle"*. (b) **Label-ambiguity**: There are some labels that convey same semantic meaning and thus can be used interchangeably, due to which usually only one of them is assigned by annotator. E.g., an image tagged with *"flowers"* may not be tagged with *"blooms"*, as both convey the same meaning. (c) **Structural-overlap**: There are some labels that, in spite of being different, *share* structural properties. E.g., though *"tiger"* and *"lion"* are two different labels, structurally they are very similar.

Most of the earlier works in this domain have focused on nearest-neighbour based models. A recent work [4] also tries to integrate label information in the nearest-neighbour set-up. In this work, first we demonstrate that even the conventional SVM outperforms several benchmark models. Then we propose a new loss function based on the hinge-loss in order to make SVM tolerant against the three issues discussed above. For this, we introduce a tolerance-parameter "t" that adjusts both the margin as well the gradient update-rule for each sample separately. We call this model as *Suppor Vector Machine with Variable Tolerance* (or *SVM-VT*). Specifically, we formalize the SVM-VT model as that of solving the following optimization problem:

$$\min_{\mathbf{w}} \frac{\lambda}{2}||\mathbf{w}||^2 + \frac{1}{m}\sum_{j=1}^{m}[1 - y_j t_j(\mathbf{w}\cdot\mathbf{x}_j)]_+, \tag{1}$$

where the additional parameter $t_j \in [0,1]$ controls the tolerance against the errors made in the classification of sample $\mathbf{x}_j$. The hyperplane $\mathbf{w}$ is learnt such that it is more strict towards correctly classifying samples with high value of t_j and any such error leads to a large shift in the hyperplane.

We propose a heuristic approach for determining the t-value for each sample given a label. For a label l, let S^+ and $\bar{S}^+$ be the sets of its positive and negative examples respectively. We consider three factors to determine the semantic relatedness of each sample $\mathbf{x}_j \in \bar{S}^+$ with that label:
(a) *Reverse nearest-neighbours based score*: For a fixed value of K (= 5), let p_k be the number of samples in S^+ that have $\mathbf{x}_j$ as their k^{th} nearest neighbour. Then we define

$$score_1(\mathbf{x}_j|l) = \frac{\sum_{k=1}^{K}(\frac{p_k}{k})}{\sum_{k=1}^{K} p_k + \varepsilon} \tag{2}$$

(b) *Visual similarity based score*: We compute the visual similarity score $sim(\cdot)$ (scaled into range $[0,1]$) of $\mathbf{x}_j$ with its nearest neighbour $\mathbf{x}_i^* \in S^+$ using JEC [3] method and define

$$score_2(\mathbf{x}_j|l) = sim(\mathbf{x}_j, \mathbf{x}_i^*) \tag{3}$$

(c) *Label cooccurrence based score*: For a label l, let $\mathbf{y} \in \{0,1\}^m$ be such that its i^{th} entry is 1 if the i^{th} training image is tagged with l, and 0 otherwise. We compute cooccurrence score $co\_occur(l_i, l_j)$ between two labels l_i and l_j using cosine similarity between their corresponding vectors $\mathbf{y}_i$ and $\mathbf{y}_j$. Let $\mathbf{x}_j$ be tagged with labels L_j, then we define

$$score_3(\mathbf{x}_j|l) = \max_{l_j \in L_j} co\_occur(l, l_j) \tag{4}$$

From these, we define tolerance parameter for sample $\mathbf{x}_j$ given label l as

$$t_j = 1 - \frac{1}{3}(score_1(\mathbf{x}_j|l) + score_2(\mathbf{x}_j|l) + score_3(\mathbf{x}_j|l)) \tag{5}$$

We compute t_j only for samples in $\bar{S}^+$, and take $t_j = 1$ for all positive samples assuming that they are correctly annotated. From equation 5, it can be seen that for some negative sample, smaller tolerance value corresponds to higher chance of it being related to a given label and vice-versa. Figure 1 shows negative samples (along with their ground-truth labels) with least t-scores for two labels each from three benchmark datasets.

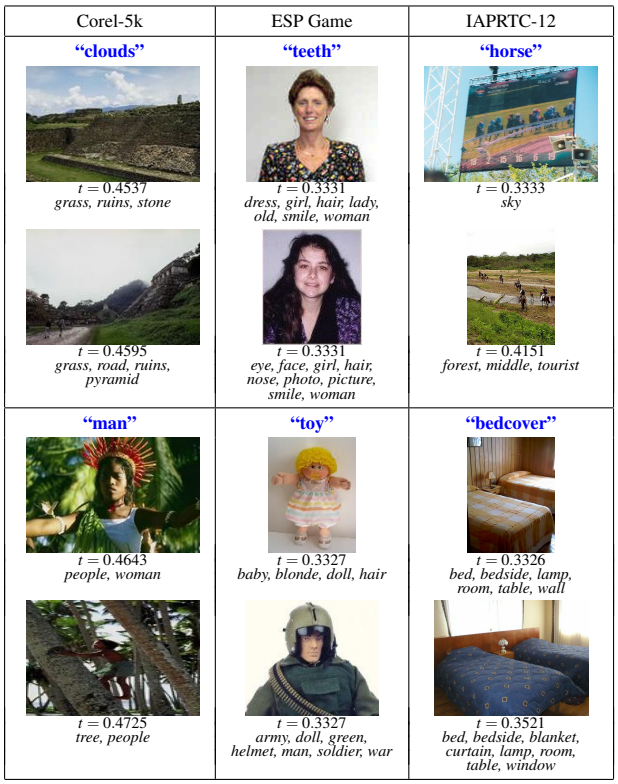

Figure 1: For example labels (in blue) from the three datasets, the top "negative" samples that have least t-scores with corresponding ground-truth labels (smaller t-score implies higher semantic relevance).

Dataset →	Corel-5k	ESP Game	IAPRTC-12
Method ↓	P/R/F1/N+	P/R/F1/N+	P/R/F1/N+
MBRM[1]	0.24/0.25/0.245/122	0.18/0.19/0.185/209	0.24/0.23/0.235/233
JEC[3]	0.27/0.32/0.293/139	0.22/0.25/0.234/224	0.28/0.29/0.285/250
TagProp-ML[2]	0.31/0.37/0.337/146	**0.49**/0.20/0.284/213	**0.48**/0.25/0.329/227
TagProp-σML[2]	**0.33/0.42/0.370/160**	0.39/**0.27/0.319/239**	0.46/**0.35/0.398/266**
KSVM	0.29/0.43/0.346/174	0.30/0.28/0.290/256	0.43/0.27/0.332/266
KSVM-VT (Ours)	**0.32/0.42/0.363/179**	**0.33/0.32/0.325/259**	**0.47/0.29/0.359/268**

Table 1: Performance comparison among different methods. The prefix 'K' corresponds to kernelization using chi-squared kernel.

We evaluate the performance using average precision per label (P), average recall per label (R), average F1 score, and number of labels with positive recall (N+). Table 1 shows the annotation performance of different methods. It can be seen that our method shows promising results on the task of image annotation on three challenging datasets, and establishes a baseline for such models in this domain.

[1] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR*, 2004.
[2] M. Guillaumin, T. Mensink, J. Verbeek, and C. Schmid. Tagprop: Discriminative metric learning in nearest neighbour models for image auto-annotation. In *ICCV*, 2009.
[3] A. Makadia, V. Pavlovic, and S. Kumar. Baselines for image annotation. In *IJCV*, 2010.
[4] Y. Verma and C. V. Jawahar. Image annotation using metric learning in semantic neighbourhoods. In *ECCV*, 2012.