Semi Supervised Learning in Wild Faces and Videos

David Rim
david.rim@polymtl.ca
Kamrul Hassan
md-kamrul.hasan@polymtl.ca
Chris Pal
christopher.pal@polymtl.ca

Ecole-Polytechnique Montreal
Montreal, Quebec, CA

We propose an approach for improving unconstrained face recognition based on leveraging weakly labeled web videos. It is easy to obtain videos that are likely to contain a face of interest from sites such as YouTube through issuing queries with a person's name; however, many examples of faces not belonging to the person of interest will be present. We propose a new technique capable of learning using weakly or noisly labeled faces obtained in this setting augmented with a margin-like property based on a null category region model following the work of Lawrence and Jordan [4]. However, we add implement this with softer constraints in the context of noisy, rather than missing, labels.

We create a subset of the LFW dataset [3] using the 50 subjects having the most labeled examples, yielding 2. We combine this set with 4000 negatives drawn from the remaining subjects having as least 3 examples, a subset of 6733 examples from LFW. We create a large dataset of facial images sampled from Youtube videos designed to mirror this dataset, by issuing queries using the subject name, and extracting facial images from the set of resulting videos.

We use the simple approach of using the search queries as a weak label. We then use simple descriptor based methods shown to have good performance in the unconstrained face verification task and combine this feature representation with a novel probablistic model with a low-density separation approach and show that unconstrained facial recognition in both static and video images can be significantly improved using unlabeled data.

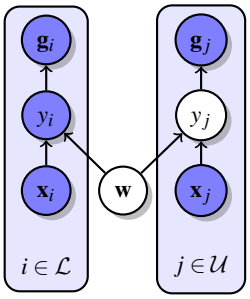

Figure 1: Graphical model of semi-supervised learning as presented in [4], $\mathcal{U}$ is the index set of unlabeled examples, and $\mathcal{L}$ is the index set of the labeled examples.

In our model, the dataset $\mathcal{D}$ is composed of sets of variables $\mathcal{D} = \{(\mathbf{x}_n, y_n, \mathbf{g}_n)\}_{n=1,2,\ldots,N}$, with $\mathbf{x} \in \mathbb{R}^d$, $y \in \{-1,1\}$, and $\mathbf{g} \in \{-1,1\}$, where the conditional dependencies are specified as in the model shown in Figure (1). That is, the LFW images have labels, y, and the task is to output a classifier parameterized by the random vector $\mathbf{w}$, which, when combined with an observed $\mathbf{x}$, maps to the correct subject y. Since not every example has a corresponding label, we operate under the general setting of binary classification in the presence of missing values. Let $\mathcal{L}$ be an index set for the labeled data $\mathcal{D}_i$, and $\mathcal{U}$ be an index set for the unlabeled data $\mathcal{D} \setminus \mathcal{L}$ The dependencies shown in the model above allows us to use the additional observed values $\mathbf{g}$ to train a discriminative classifier using the unlabeled data.

We note that because of the d-separation of $\mathbf{w}$ and $\mathbf{g}_i$ by y_i, it is unnecessary to learn a maximum likelihood solution for $\mathbf{w}$ if the labels y_i are observed.

However, if the label, y_j, is missing, then observing $\mathbf{g}_j$ makes $\mathbf{w}$ and $\mathbf{x}_j$ conditionally dependent, having the common observed ancestor, seen in Figure (1), so that unlabeled data are useful. If the relationship between y_j and g_j is tightly coupled, then g_j may be thought of as a noisy label. The general idea is that g_j is a variable which is far easier to obtain than the true label, y_j. In this case, the search query provides a weak or noisy label g_j for each facial image contained in the video. That is, facial images retrieved from videos returned from issuing "George W Bush" as a query to YouTube can be weakly labeled as "George W Bush." We treat the recognition problem as a one-vs-all multi-class classification task, and desire to learn a $\mathbf{w}$ for each subject individually.

We learn the above model using a variational approach, in which the objective is to maximize

$$\log p(\mathcal{D}) = \int_{\mathbf{Z}} q(\mathbf{Z}) \log \frac{p(\mathcal{D},\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} + KL(q||p) = \mathcal{L}(q) + KL(q||p), \quad (1)$$

where Z are the latent variables, y_j. As is typically the case in variational inference [1], we let $q(\mathbf{Z})$ factorize with respect to each y_j for $j \in U$, and minimize the KL divergence $KL(q||p)$.

We treat the distribution $p(g_j|y_j)$ as a multinomial, letting

$$p(g_j = 1|y_j) = \begin{cases} \mu^+ & \text{if } y_j = 1, \\ \mu^- & \text{if } y_j = -1, \\ \mu^0 = 1 - \mu^+ - \mu^- & \text{otherwise, if } y_j = 0, \end{cases} \quad (2)$$

and assume symmetry, $p(g_j = -1|y_j = -1) = p(g_j = 1|y_j = 1)$. We model $p(y_n|\mathbf{x}_n)$ as a soft max function of $\mathbf{w}$.

Furthermore, we model the prior distribution of $\mathbf{w}$ with a zero-mean Gaussian, as in the Relevance Vector Machine (RVM) [5], except that instead of the the gamma hyperprior, as used in Bishop and Tipping's variational approach [2], we use an Exponential hyperprior. This can be shown to result in an Exponential prior, which approximates an L_1 penalty which yields sparse $\mathbf{w}$.

The latent variables y_j are learned using EM based on our parameterization, and the weights w are maximized according to the model, yielding an RVM-like formulation of a semi-supervised learning algorithm which could also be used transductively.

We show the overall effectiveness of using weak labels derived from video data combined with existing facial images to aid learning tasks. Noisy labels are readily available in video for faces using this method. Adding a null-category with soft constraints produces better results than using the labeled data alone. In all cases, the use of unlabeled data is able to improve the final classifier. The model is fully probabilistic and gives greater flexibility including the use of non-standard priors, most specifically sparsity-inducing priors. Going beyond experiments with a known noise level, using a realistic dataset an estimate of the noise yielded improvement as well. On a realistic task of interest, improving a static face classifier using readily available video images, the method also produces good results.

[1] H. Attias. A variational Bayesian framework for graphical models. *Advances in neural information processing systems*, 12(1-2): 209–215, 2000.

[2] C.M. Bishop and M.E. Tipping. Variational relevance vector machines. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 46–53, 2000.

[3] G. B Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. *UMass, Amherst, TR 07*, 49: 1, 2007. URL `http://vis-www.cs.umass.edu/papers/eccv2008-lfw.pdf`.

[4] N.D. Lawrence and M.I. Jordan. Semi-supervised learning via Gaussian processes. *NIPS*, 17:753–760, 2005.

[5] ME Tipping. The Relevance Vector Machine. *Advances in Neural Information Processing Systems*, 12, 2000.