Evolutionary Data Purification for Social Media Classification

Stuart James and John Collomosse
Centre for Vision, Speech and Signal Processing (CVSSP)
University of Surrey, UK.
{s.james, j.collomosse}@surrey.ac.uk

***Abstract*—We present a novel algorithm for the semantic labeling of photographs shared via social media. Such imagery is diverse, exhibiting high intra-class variation that demands large training data volumes to learn representative classifiers. Unfortunately image annotation at scale is noisy resulting in errors in the training corpus that confound classifier accuracy. We show how evolutionary algorithms may be applied to select a 'purified' subset of the training corpus to optimize classifier performance. We demonstrate our approach over a variety of image descriptors (including deeply learned features) and support vector machines.**

I. Introduction

Billions of social media posts are shared world-wide each day, and much of this data is visual. Automated labeling (classification) of this unstructured, diverse visual data into high-level semantic concepts presents significant challenges beyond those posed by modern lab based datasets *e.g.* ImageNet [10]. In particular, the high intra-class variability of social media captured 'in the wild' requires a step-change in the volume of training data required to learn representative image classifiers. The issue is compounded by modern deep learning approaches, that whilst delivering significant improvements in annotation performance, are relatively data hungry. Unfortunately annotated social media for training is expensive to acquire at scale. This has motivated strategies such as crowd-sourced markup (*e.g.* gamified annotation) or the combination of auxiliary data with domain adaptation (*e.g.* incorporating results of a Google of FlickR image search) to bolster the training corpus. However both strategies result in noisy annotation in training that degrades performance of learned classifiers.

This paper presents a novel technique for filtering (or 'purifying') noisy annotated training data for the purpose of training supervised image classifiers. By selecting the best subset of available data for training, we show that significant performance gains can be achieved for social media image classification using contemporary deep-learning approaches. Whilst sparsification of data for supervised classification is not new to machine learning, our evolutionary strategy for exploring the combinatorial space of data selection is, to the best of our knowledge, novel and forms the main contribution of this paper. Specifically, we apply a Genetic Algorithm (GA) to iteratively optimize a binary vector describing the selection of training data, encouraging labeling consensus within the positive training examples for several high-level semantic concept groups. Our work uses deeply learned feature sets obtained from the fully connected layers of a convolutional neural network (CNN) with the decision layer provided by support vector machine (SVM) classifiers. We also evaluate over a variety of other contemporary features and SVM configurations to showcase the general applicability of our approach. Ultimately our evolutionary approach increases the practicality of semantic classification for high intra-class varying data, allowing for 'in the wild' social media labeling.

A. Related Work

Early image annotation work focused upon the extraction of specific binary scene semantics *e.g.* indoor/outdoor [28]. Generalization to multiple semantic classes was initially performed by jointly modeling the space of semantic labels and visual features. By introducing a set of latent variables to encode hidden states between labels and image descriptors, semantic labels can be inferred in an unsupervised manner [1], [12], [18], [25]. Extensions of such approaches to socially contributed image data (*e.g.* FlickR) have been explored through propagation of labels across an attribute graph [3], [23], [29] benefiting from contextual information *e.g.* derived from metadata tags. Most social media image classification has focused on use of FlickR data, benefiting from loose user-tagging of imagery offering valuable prior context. Little has been done to explicitly classify social media from untagged sources *e.g.* Facebook.

Supervised image classification has benefited extensively from gradient domain image descriptors (e.g. HOG, SIFT) and the past decade has seen various feature space encoding strategies, from basic vector quantization [8] to VLAD [15] and Fisher Vector encodings [27]. Significant performance advantages for semantic labeling have been obtained by combining these representations with machine learning classifiers particularly support vector machines (SVMs) [4], [5], [9]. Higher level concepts, such as image aesthetics [22], have been explored using similar pipelines trained using ratings sourced from FlickR. Marchesotti *et al.* expanded this early work to incorporate textual bi-grams from user comments combined with visual features, learning 'beautiful' and 'ugly' image attributes [21].

A common problem with the above methods is scalability to large numbers of concepts, and large-scale categorization techniques have been developed by Wang *et al.* [32], [33] and [31]. These partly leverage auxiliary search engines to retrieve related images from web-scale image sets, utilizing text word search to obtain a ranked list of candidate tags. Malisiewicz *et al.* [20] showed that training ensembles of classifiers (*e.g.* one per training data item) could yield further

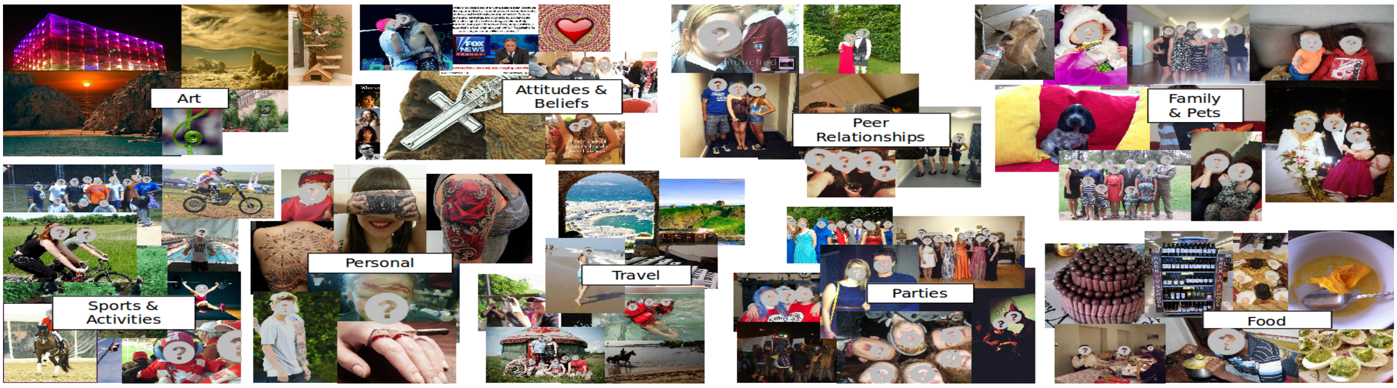

Fig. 1: Corpus of social media harvested from participant Facebook profiles, labeled to high-level semantic concepts identified through anthropological study. Figure showing representative examples of each concept and demonstrating a high degree of intra-class variability in both appearance and content.

performance benefits, though such 'exemplar SVM' approaches scale linearly with training corpus size making them impractical for datasets exhibiting high intra-class variation *e.g.* social media [35]. Juneja *et al.* [16] later explored localization of semantics concepts within images, utilizing exemplars SVM to improve performance. Most recently a shift toward deeply learned neural networks (*e.g.* CNNs) that simultaneously learn both the descriptors and classification stages of the pipeline have gained popularity due to the step-change in performance reported on standard image classification and object recognition benchmarks [17]. The power of deep learning frameworks typically lies within the learned features (*e.g.* within the fully-connected layers of a CNN) which may be decoupled post-training and fed into alternative classifiers. For example, utilizing a spatially localized CNN, Dixit *et al.* [11] built semantic Fisher vector representation combining the feature representation trained in a CNN and used SVM to classify images with context.

Our work also makes use of CNN derived features with an SVM classification back-end, but we apply evolutionary algorithms to identify an optimal subset of the noisy training data available to us. Although evolutionary algorithms have been used for feature space re-weighting and transfer learning in supervised classification [26], our work contrasts in the use of evolutionary optimization for training data selection; essential when dealing with high volume, noisily annotated datasets such as social media.

II. Social Media Classification

Image classification commonly focuses on recognition *e.g.* of dominant physical objects within images, and significant advances have been made due to the availability of large annotated datasets. However practical scenarios seeking to 'make sense' of social media imagery do not focus on object detection but rather the classification of imagery into fewer, higher level semantic groups [7] *e.g.* sports, family, friends, beliefs. Such imagery tends to depict mixtures of concepts within cluttered scenes and so annotation carries greater uncertainty. Traditionally to overcome intra-class ambiguity, approaches rely upon extensive user annotation drawing a labeling consensus from multiple annotators of the same image, so requiring extensive manual effort [34]. Alternatively machine learning algorithms can bootstrap a consensus across a large data corpus. Uncertainty within the annotated corpus is further compounded by the poor quality of such casually captured images. Data purification can overcome both issues through selection an optimal subset of training data simultaneously reflecting both the reliability of the label and the quality of the visual representation.

A. Dataset and Augmentation

In this work we study the challenging problem of Facebook image annotation. We select this platform due to the abundance of unlabeled social imagery, distinct from well-studied platforms such as FlickR for which images are associated with carefully curated keyword tags. By contrast we use purely visual data.

Twenty college-aged participants were recruited from the same geographic region, and consented to the harvesting of all photographic content within their private Facebook profiles. Building upon an anthropological study of this age group [7], a set of nine high-level semantic concept groups were identified reflecting common themes within posts, namely: Art, Attitude & Beliefs, Family & Pets, Food, Friends, Travel, Celebrations; Personal style and self-imagery (e. g. selfie) and Sports. Under controlled conditions we invited participants to manually annotate each other's photographs with these nine concept group labels over a total of 5k images. Note that multiple concepts may be annotated as present within an image (Fig.1).

Given the high within-class diversity of the dataset, the corpus is boosted using weakly labeled auxiliary content harvested from Google Image Search. An image trawler was implemented to identify additional images based on the above keyword concepts. Since only the top few results for a given keyword are typically relevant, we exploit the WordNet taxonomy [24], applying the *'Is-A'* relationship to construct a syn-set of related keywords. We harvest an additional 23k images evenly distributed across the nine concept groups using this method. These weakly labeled images are not pre-filtered in any way, yet the use of additional noisy annotated data is beneficial when training when applying our data purification approach (Subsec. II-C).

B. Visual Representation

The representation of visual data for supervised classification has transformed in recent years. Here we explore four different approaches to representing visual information across two methodologies; Shallow and Deep. Shallow representations are derived from prescriptive, hand crafted gradient-domain features, commonly applying the feature-space quantization strategies (*e.g*. Bag of Visual Words) to encode interest point descriptors. We explore SIFT [19] and PHOW-Colour [2] feature descriptors allowing us to demonstrate the benefit of additionally encoding colour information in the latter. In our work we use the defacto standard hard-assignment Bag of Visual Words (BoVW) pipeline [8] constructing a dictionary via K-Means (where K is the codebook size), and assigning descriptors to bins on a nearest-neighbor basis to form a frequency histogram describing the image. A recent trend in image classification is the resurgence of deep representations, a popular choice being the Convolutional Neural Network (CNN) [17] in which convolutional filter banks are optimized to learn an feature representation directly from training data. CNNs requires large training corpora to generalize well, and we explore CNNs trained both on ImageNet and on ImageNet and our 28k image corpus. Below we explain parameters and settings used within our experiments:

- **Method 1: CNN Features** – are extracted from the Fully Connected layers of the Neural Network, it has been shown [30] that using a SVM can improve performance vs using the end-layers of the Neural Network classifier. Therefore we extract descriptors from the ImageNet trained model, using the FP7 layer as in [6].
- **Method 2: Optimized CNN Features** – Training a CNN requires large data corpora, which ImageNet provides. It has been shown that fine tuning the model over the target domain can improve performance, *e.g*. [6] demonstrated that 3% improvement can be achieved through such a process on the VOC-2007 dataset. We therefore apply 100k iterations of training to optimize the ImageNet model over our 28k social image dataset.
- **Method 3: SIFT with BoVW** – SIFT descriptors encode local gradient information. We resized images, preserving aspect ratio, with width constrained to 300 pixels. Dense SIFT descriptors are extracted within 32×32 windows spaced at regular intervals (4 pixels), descriptors are the encoded into a BoVW representation using $K = 2000$.
- **Method 4: PHOW-Color with BoVW** – PHOW is a variant of SIFT at multiple scales. The descriptor is computed independently over each channel of the HSV colour space, and the results concatenated. We form a BoVW representation similarly using $K = 2000$.

In all of the above experimental configurations the resultant descriptors are normalized by $\ell 1$ norm.

C. Evolutionary Data Purification

We partition the social media dataset evenly into training, validation and test sets (approximately 1.7k images each). The auxiliary (Google Image) data is then used to augment the training data resulting in a training corpus of approximately 25k images. Due to the potential presence of multiple labels per images, we train a binary classifier (SVM) independently for each concept (sports, family, etc.). As later shown in Sec. III, it is possible to gain a substantial performance boost by training each classifier with a limited subset of the available data. Our problem is to select that optimal subset for each of the classifiers. We do so by iteratively turning on/off items of training data, and evaluating the trained model against the validation set. Clearly the space of training data configurations is very high dimensional and this optimality criterion makes the space also very turbulent. Stochastic searches that model evolutionary processes, such as Genetic Algorithms (GAs), are often cited among the best search strategies in such situations; large regions of problem space can be covered quickly, and local minima more likely to be avoided [13], [14].

We optimize the data selection for each classifier using a GA. The GA encodes the selection of training data as a binary genome applying principles of natural selection to iteratively 'breed' better solutions. A population of individuals, each corresponding to a data selection solution, are initially seeded at random. We represent the population as $Q \times N$ matrix $X(t)$ coding for a population of $Q = 50$ individuals at generation t, each of which has genome N bits in length, one per per data item. Note that the population does not change side over generations. Individuals are bred at together to produce successive generations $i = [2,250]$ of the population through processes modeling genetic cross-over, genome mutation, and fitness-proportionate reproduction (Fig. 2). Each of these processes are defined in the following paragraphs. We choose a maximum of 250 iterations, identified through observation of the fitness function indicating this to be where improvement plateaus (Fig.3). Although the trajectory in fig.3 indicates continual improvement, there is increasing risk of over fitting to the validation set as iterations of GA continue.

I **Fitness-proportionate selection** For each member of the population an SVM classifier is trained using only data items indicated by positive bits in $X_{i,1..N}(t)$. We opt for linear kernels within the SVM, as non-linear kernels require computationally expensive training as well as additional parameter optimization that prohibit practical application in a GA evaluation loop. The set of classifiers trained for each individual $\mathcal{M} = \{M_1,...,M_Q\}$ are then evaluated against the validation set, to yield a fitness score $F(i;t)$ for each individual as follows.

For each item in the validation set $\mathcal{V} = \{v_1,...\}$ we have also a binary ground-truth for the presence of the concept, forming binary vector $L = \{l_1,...\}$. We first normalize the response for classifier M_i by normalizing the distance of each validation datum v_j from decision boundary through a sigmoid $\mathcal{S}(v_j;M_i)$:

$$\mathcal{S}(v_j;M_i) = \frac{1}{1+\exp(-12\mathcal{N}(M_i(v_j))-0.5)} \qquad (1)$$

where $\mathcal{N}(.)$ is the normalization by the min and max limits of the model identified over the training data. Resulting in $[0 \rightarrow 1] \in \Re$ probability $\rho(v_j)$, of the concept being present.

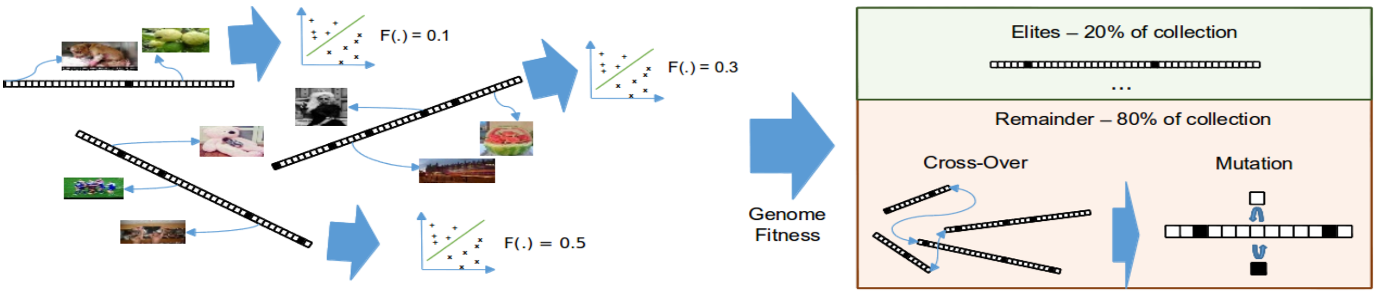

Fig. 2: Illustration one iteration of GA optimization. A new population of genomes are stochastically generated via fitness proportionate selection, dependent on the performance of a classifier (trained using the selected data subset) over a validation set. The fittest 20% of individuals are selected via elitism and given a free-pass to the next generation. The remainder are evolved through genetic cross-over and mutation.

A binary decision vector $D = \{d_1, ...\}$ for presence of the concept is obtain thresholding $d_j = \rho(v_j) \geqslant 0.5$. The fitness $F(i;t)$ of individual i is expressed as the precision over the validation set:

$$F(i;t) = 1 - \frac{\sum_{j=1}^{|V|} d_j \oplus l_j}{|\mathcal{V}|} \qquad (2)$$

To produce the generation X_{t+1}, the best performing 20% individuals are given a 'free-pass' (copied directly) to the next generation, implementing elitism. The remaining 80% of the next generation, are created through fitness proportionate selection. Individuals within X_t are sampled stochastically (with replacement) with a bias to $F(i,t)$, and pairs of parents are subjected to genetic cross-over producing a new offspring that is mutated and added to X_{t+1}.

II **Genome Cross-Over** occurs between the two selected parents in order to define a new individual for X_{t+1}. A split point $[1,N]$ is identified randomly within the genome and the two parents are spliced together at that point to a single new genome. This process results in a new combined configuration of training samples to be selected.

III **Genetic Mutation** The offspring resulting from cross-over is subjected to mutation to induce diversity in the population, countering the homogenizing effect of elitism. Each bit of the offspring's genome is visited and flipped with a 1% change, introducing (or removing) training samples that may have been in a dormant state in the initial configuration.

The computationally expensive step is the complete evaluation of the population X_t, motivating practical use of linear SVMs within the GA loop over potentially beneficial non-linear SVM kernels. This does not preclude the final data selection at $t = 250$ being used to train a non-linear *e.g.* RBF kernel SVM. In many cases there is a performance increase in doing so (Sec. III) but given the optimization of the fitness function is governed by a linear kernel this may not always be so. In rare cases the use a final RBF kernel results in a decrease in performance, detectable through use of the validation set. In the cases where there is a decrease in performance, we can revert to a linear SVM model. This effect is predominantly seen within 'Friends & Peer Relationship' concept that is challenging to distinguish from 'Family & Pets' and ' Parties & Celebrations' all of which exhibit similar visual cues.

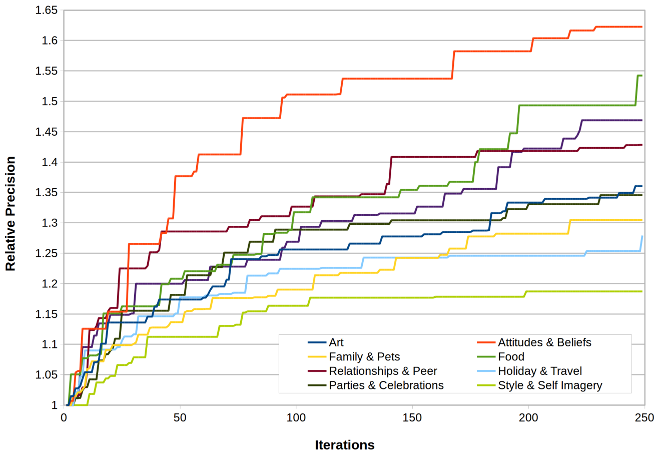

Fig. 3: Iterations of GA training showing the relative performance gain of the best performing genome over the validation data. Showing one fold for the 9 concepts for the Optimized CNN feature type (Method 2)

III. Results and Discussion

We evaluate classifier performance (with and without application of our data purification technique) using 5-fold cross-validation over our social media dataset. Recall that this data is split evenly into training, validation and test data yielding $\sim 1.7k$ images in each partition. The auxiliary data from Google was used only to bolster the training set (to $\sim 25k$ images). For each of the descriptors evaluated (Sub-sec. II-B) training is performed via this data. The validation set is used only for experimental configurations using data purification.

Table I shows the mean results over the folds. We split the results into two tables for deep and shallow feature types, for easier comparison. We additionally highlight the top performing configuration per concept group for each of the tables. We compare final SVM classifiers using both linear kernels and non-linear (RBF) kernels. As expected the results show CNN-

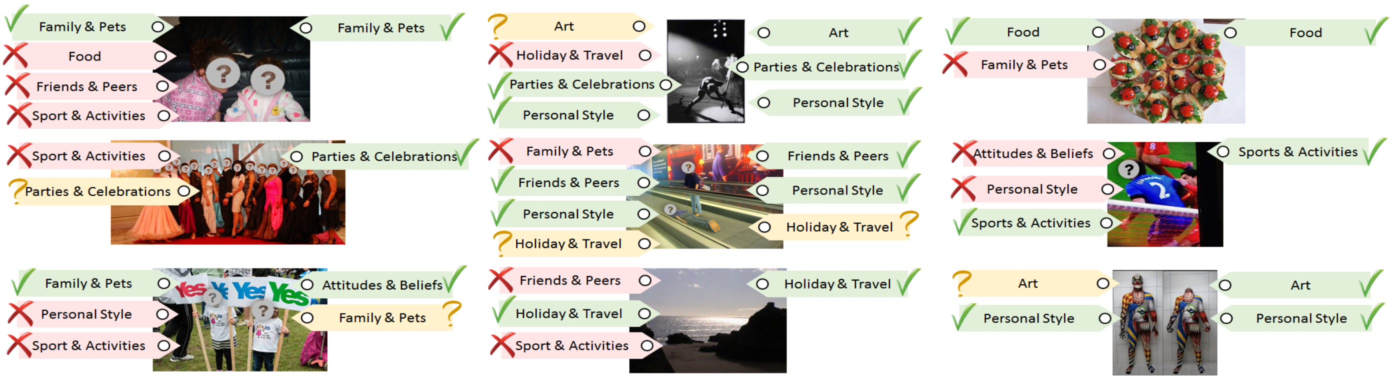

Fig. 4: Visual examples of label annotation using Method 2 where the initial classifier labels are displayed on the left of the image and the final labels after GA is performed are presented on the right.

Descriptor	**Method 1: CNN features**				**Method 2: Optimized CNN features**			
Phase	Initial		Post GA		Initial		Post GA	
Classifier Kernel	Linear	RBF	Linear	RBF	Linear	RBF	Linear	RBF
Art	13.8 ± 3.68	28.8 ± 1.97	29.9 ± 1.05	36.0 ± 4.60	23.9 ± 2.19	29.7 ± 1.70	**35.6 ± 1.51**	31.2 ± 1.85
Attitudes & Beliefs	4.96 ± 0.88	8.49 ± 0.47	10.6 ± 0.54	**16.9 ± 2.05**	7.46 ± 0.36	13.0 ± 2.50	16.5 ± 1.47	14.5 ± 3.25
Family & Pets	31.2 ± 2.41	29.1 ± 1.00	32.3 ± 1.07	34.2 ± 2.21	29.7 ± 1.47	40.9 ± 1.66	**46.8 ± 3.97**	43.1 ± 2.23
Food	7.64 ± 1.42	30.6 ± 1.61	27.7 ± 2.85	**55.1 ± 6.22**	15.9 ± 1.18	33.1 ± 3.53	34.6 ± 4.31	38.4 ± 2.80
Relationships & Peer	5.41 ± 1.88	11.4 ± 0.54	14.3 ± 0.63	11.7 ± 4.50	11.1 ± 1.29	16.1 ± 1.40	**20.2 ± 1.44**	16.8 ± 1.76
Holidays & Travel	14.7 ± 3.48	28.3 ± 0.72	26.9 ± 2.02	**39.8 ± 2.98**	22.1 ± 0.87	29.6 ± 0.91	32.6 ± 1.03	32.0 ± 1.58
Parties & Celebrations	8.73 ± 2.42	20.0 ± 1.05	21.3 ± 0.59	20.6 ± 3.25	17.1 ± 2.16	28.2 ± 3.22	**32.3 ± 2.13**	28.5 ± 3.74
Style & Self Imagery	21.4 ± 5.48	30.3 ± 0.57	33.3 ± 1.51	37.4 ± 2.10	28.3 ± 2.59	37.8 ± 2.49	**41.9 ± 3.42**	40.6 ± 2.49
Sports	12.2 ± 2.43	20.0 ± 1.36	23.0 ± 1.00	25.0 ± 4.32	18.5 ± 1.20	24.8 ± 2.10	**29.9 ± 2.11**	26.0 ± 2.49
Mean	12.2 ± 2.67	23.0 ± 1.03	24.4 ± 1.25	30.8 ± 3.58	19.4 ± 1.48	28.1 ± 2.17	**32.3 ± 2.38**	30.1 ± 2.40

Descriptor	**Method 3: PHOW-Color with BoVW**				**Method 4: SIFT with BoVW**			
Phase	Initial		Post GA		Initial		Post GA	
Classifier Kernel	Linear	RBF	Linear	RBF	Linear	RBF	Linear	RBF
Art	15.0 ± 1.70	18.1 ± 0.39	22.6 ± 0.85	**37.4 ± 12.3**	14.2 ± 0.77	18.4 ± 0.54	21.4 ± 1.19	26.6 ± 3.05
Attitudes & Beliefs	4.83 ± 0.52	16.7 ± 2.55	**36.8 ± 6.74**	20.0 ± 3.83	4.97 ± 0.69	14.3 ± 2.89	27.8 ± 10.4	29.4 ± 9.33
Family & Pets	16.1 ± 0.41	19.7 ± 0.29	24.2 ± 0.98	**25.0 ± 1.97**	16.8 ± 0.41	21.3 ± 1.91	24.4 ± 1.57	24.0 ± 4.52
Food	6.17 ± 0.51	23.8 ± 3.29	**44.0 ± 5.95**	27.6 ± 6.73	6.12 ± 0.73	18.9 ± 3.92	24.8 ± 7.73	22.8 ± 6.20
Relationships & Peer	6.39 ± 0.62	13.7 ± 3.88	**20.5 ± 2.75**	2.86 ± 6.39	6.28 ± 1.07	12.8 ± 1.45	16.7 ± 2.68	12.2; ± 21.3
Holidays & Travel	11.4 ± 1.26	26.8 ± 0.68	30.3 ± 1.56	**41.6 ± 1.69**	12.6 ± 2.00	26.4 ± 0.73	28.3 ± 1.81	40.4 ± 2.71
Parties & Celebrations	7.39 ± 0.24	16.3 ± 1.02	**22.1 ± 1.77**	17.3 ± 5.48	6.48 ± 0.57	16.9 ± 0.87	22.0 ± 4.31	16.9 ± 0.87
Style & Self Imagery	19.2 ± 0.99	23.1 ± 0.85	28.2 ± 0.62	**30.4 ± 2.32**	19.0 ± 1.00	23.4 ± 2.20	26.5 ± 2.29	29.8 ± 3.08
Sports	10.3 ± 1.89	16.6 ± 1.19	**22.0 ± 2.60**	16.6 ± 1.19	9.71 ± 1.30	0.00 ± 0.0	20.9 ± 2.27	0.00 ± 0.0
Mean	10.7 ± 0.96	19.4 ± 1.57	**27.8 ± 2.65**	26.6 ± 5.28	10.7 ± 0.96	16.9 ± 1.61	23.7 ± 3.81	24.8 ± 7.83

TABLE I: Average per-class precision over folds for the four types of feature (sec. II-B). Demonstrating the performance difference between linear and non-linear (RBF) kernels before and after the GA optimization process.

derived features to out-perform the traditional gradient domain descriptor/BoVW models with the best performing deep representation before GA purification improving by 44% on the shallow descriptors. We observe that without any additional processing the optimized (fine-tuned) CNN results yield a 22% and 58% improvement for RBF and Linear classifier kernels respectively. Within shallow representations generally PHOW outperforms SIFT with 14% in the case of RBF kernel. It is interesting to note categories Art and Attitudes & Beliefs, using the BoVW approaches out-perform those of the CNN. However, for Sports, the SIFT descriptor fails to discriminate.

In the best case, GA purification yields a mean precision over all concepts of 32% representing a 67% relative improvement over non-purified training data. Despite the intractability of including a non-linear kernel within the GA optimization loop, using the identified training data subset to learn a non-linear classifier as a final step always equals or betters the performance of the system. This is shown over all configurations except Optimized CNN where the RBF is only able to get a marginal improvement due to higher precision of results pre-purification. The Linear SVM over Optimized CNN features performs best of all, in terms of both precision and computational performance. The PHOW feature is generally more unstable across folds than CNN based features with a higher standard deviation. Fig. 4 visually illustrates the benefits of purification quantified in Table I. Green ticks and red crosses indicate the correctness of automatically annotated tags generated by classifiers trained with, and without, purified data.

For one class and one fold of the GA on a dual 3.4GHz hex-core Intel CPU takes 2.2 hours to complete (excluding feature extraction, which varies between descriptor choice). In total our experiments take approximately 20 hours to perform the GA purification for all data and over all classes. GA implementation can easily be optimized by distributing computation of the evaluation step *e.g.* via map-reduce. However our experiments focus on a single-thread CPU implementation to provide a solid benchmark. In the case of Fine-tuned CNN it takes an additional 12 hours to fine-tune the ImageNet model on a GeForce GTX 660 Ti GPU.

IV. Conclusion

Automated labeling of imagery on social networking sites is challenging due to high intra-class content diversity, poor image quality, and the high-level semantic nature of concepts typically desirable to label in such data. Consequently high volumes of training data are needed for supervised classification tasks, which often exhibits noisy annotation. We have presented a technique for enhancing the precision of supervised classifiers over such data, using an optimization strategy to select an optimal subset of noisy annotated data to train the classifier. Applying our GA purification approach to select this optimal subset we are able to improve precision relatively by 67% over classical approaches that use all training data. Future work could explore different feature modalities *e.g.* comment field text and the potential of applying data fusion within our framework to further enhance classifier performance.

Acknowledgement

This work was funded by EPSRC grant EP/L00383X.

References

[1] D. M. Blei and M. I. Jordan. Modeling annotated data. *Proceedings of the International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 127–134, 2003.

[2] A. Bosch, A. Zisserman, and X. Munoz. Image Classification using Random Forests and Ferns. *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1–8, Oct. 2007.

[3] X. Cai, F. Nie, W. Cai, and H. Huang. New graph structured sparsity model for multi-label image annotations. *Proceedings of the IEEE International Conference on Computer Vision*, 1:801–808, 2013.

[4] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):394–410, 2007.

[5] E. Chang, K. Goh, G. Sychay, and G. Wu. CBSA: Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(1):26–38, 2003.

[6] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the Devil in the Details: Delving Deep into Convolutional Nets. *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–11, 2014.

[7] J. Collomosse, S. James, A. Durrant, D. Trujillo-pisanty, W. Moncur, K. M. Orzech, and S. Martindale. Enhancing Digital Literacy by Multimodal Data Mining of the Digital Lifespan. In *Proceedings of Digital Economy (DE2014)*, 2014.

[8] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. pages 1–22, 2004.

[9] C. Cusano, M. Bicocca, and V. Bicocca. Image annotation using SVM. *Proceedings of SPIE*, (1):330–338, 2003.

[10] 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.

[11] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos. Scene Classification with Semantic Fisher Vectors. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[12] S. Feng, R. Manmatha, and V. Lavrenko. Multiple Bernoulli relevance models for image and video annotation. *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, 2, 2004.

[13] D. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley, Reading, MA, 1989. ISBN: 0-201-15767-5.

[14] J. Holland. *Adaptation in Natural and Artificial Systems. An introductory analysis with applications to biology, control, and artificial intelligence*. Univ. Michigan Press, 1st edition, 1975. ISBN: 0-472-08460-7.

[15] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *IEEE Conference on Computer Vision & Pattern Recognition*, jun 2010.

[16] M. Juneja and A. Vedaldi. Blocks that Shout : Distinctive Parts for Scene Classification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[17] a. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. *Proceedings of Neural Information Processing Systems (NIPS)*, pages 1097–1105, 2012.

[18] V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. *16th Conference on Advances in Neural Information Processing Systems (NIPS)*, pages 553–560, 2003.

[19] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision (IJCV)*, 60(2):91–110, Nov. 2004.

[20] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[21] L. Marchesotti and F. Perronnin. Learning beautiful (and ugly) attributes. *British Machine Vision Conference*, pages 1–11, 2013.

[22] L. Marchesotti, F. Perronnin, D. Larlus, and G. Csurka. Assessing the aesthetic quality of photographs using generic image descriptors. *Proceedings of the IEEE International Conference on Computer Vision*, pages 1784–1791, 2011.

[23] J. McAuley and J. Leskovec. Image labeling on a network: Using social-network metadata for image classification. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 7575 LNCS(PART 4):828–841, 2012.

[24] G. A. Miller. Wordnet: A lexical database for english. *COMMUNICATIONS OF THE ACM*, 38:39–41, 1995.

[25] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[26] M. Pei, E. Goodman, W. Punch, and Y. Ding. Genetic algorithms for classification and feature extraction. *Classification Society Conference*, pages 1–28, 1995.

[27] F. Perronnin and C. Dance. Fisher kernels on visual vocabularies for image categorization. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, June 2007.

[28] M. Szummer and R. W. Picard. Indoor-Outdoor Image Classi cation. *Neural Computation*, (445), 1998.

[29] J. Tang, R. Hong, S. Yan, T.-S. Chua, G.-J. Qi, and R. Jain. NN-sparse graph-based label propagation over noisily tagged web images. *ACM Transactions on Intelligent Systems and Technology*, 2(2):1–15, 2011.

[30] Y. Tang. Deep Learning using Linear Support Vector Machines. In *Procedings of the International Conference on Machine Learning (ICML)*, 2013.

[31] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958–1970, 2008.

[32] C. Wang, F. Jing, L. Zhang, and H. J. Zhang. Scalable search-based image annotation. *Multimedia Systems*, 14(4):205–220, 2008.

[33] C. H. Wang, F. Jing, L. Zhang, and H. J. Zhang. Image Annotation Refinement using Random Walk with Restarts. *In Proceedings of ACM Multimedia*, pages 2–5, 2006.

[34] J. Whitehill, T. fan Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 2035–2043. Curran Associates, Inc., 2009.

[35] Z. Zhang and P. Yang. An Ensemble of Classifiers with Genetic Algorithm. *Evaluation*, 9(1):18–24, 2008.