Predicting Domain Adaptivity: Redo or Recycle?

Ting Yao, Chong-Wah Ngo, Shiai Zhu
Dept of Computer Science, City University of Hong Kong, Kowloon, Hong Kong
tingyao.ustc@gmail.com; cwngo@cs.cityu.edu.hk; shiaizhu2@student.cityu.edu.hk

ABSTRACT

Over the years, the academic researchers have contributed various visual concept classifiers. Nevertheless, given a new dataset, most researchers still prefer to develop large number of classifiers from scratch despite expensive labeling efforts and limited computing resources. A valid question is why not multimedia community "embrace the green" and recycle off-the-shelf classifiers for new dataset. The difficulty originates from the domain gap that there are many different factors that govern the development of a classifier and eventually drive its performance to emphasize certain aspects of dataset. Reapplying a classifier to an unseen dataset may end up GIGO (garbage in, garbage out) and the performance could be much worse than re-developing a new classifier with very few training examples. In this paper, we explore different parameters, including shift of data distribution, visual and context diversities, that may hinder or otherwise encourage the recycling of old classifiers for new dataset. Particularly, we give empirical insights of when to recycle available resources, and when to redo from scratch by completely forgetting the past and train a brand new classifier. Based on these analysis, we further propose an approach for predicting the negative transfer of a concept classifier to a different domain given the observed parameters. Experimental results show that the prediction accuracy of over 75% can be achieved when transferring concept classifiers learnt from LSCOM (news video domain), ImageNet (Web image domain) and Flickr-SF (weakly tagged Web image domain), respectively, to TRECVID 2011 dataset (Web video domain).

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Algorithms, Experimentation, Performance.

Keywords

Cross-domain concept learning, domain adaptation

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MM'12, October 29–November 2, 2012, Nara, Japan.
Copyright 2012 ACM 978-1-4503-1089-5/12/10 ...$10.00.

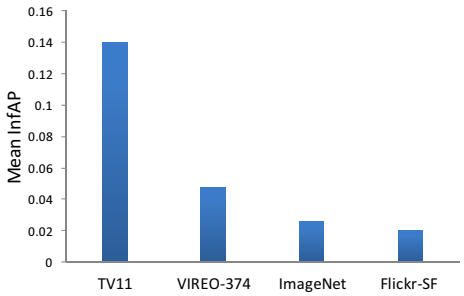

Figure 1: Mean InfAP of different classifiers on TRECVID 2011 test set.

1. INTRODUCTION

Visual concept detection has received intensive research attention since the early of year 2000 [9], and is probably the "hottest" topic with the most number of published papers in the area of multimedia. To date in the literature, there are various datasets (e.g., ImageNet) that include expert labeled training data freely available for developing concept classifiers. Some datasets (e.g., VIREO-374 [6]) even give away classifiers for free that could be directly employed for multimedia tasks. On the other hand, the popularity of social media has also generated massive amount of weakly tagged images and videos which could be leveraged for concept classifier learning. Nevertheless, despite the community generosity in sharing and contributing either the expert or weakly labeled data, given a new dataset most researchers will still perform "intensive manual labeling" and then train classifiers based on the examples collected from the dataset only. Many previous experiences have also shown that by doing so can generate satisfactory MAP (mean average precision) performance, though the learnt knowledge (classifiers) may not be useful at all when being applied in another dataset. Figure 1 shows the result of TRECVID 2011 SIN (Semantic Indexing) task [8] by using the training data from (TV11) and outside (VIREO-374 [6], ImageNet [2] and Flickr-SF [12]) of the dataset itself. VIREO-374 are classifiers learnt from expert labeled examples in LSCOM [7], ImageNet includes classifiers trained from Web images which are labeled through Amazon MTurk, while Flickr-SF includes classifiers learnt directly from loosely-tagged images. As shown in the figure, there is a huge performance difference in terms of MAP between re-training new classifiers and re-cycling off-the-shelf classifiers.

In the literature, there have been several techniques, in-

cluding Adaptive SVMs (A-SVMs) [10] and Domain Transfer SVM (DTSVM) [3], being proposed for addressing the challenge of cross-domain learning (or transfer learning). The problem is defined as, given very few training examples from a target domain (new dataset), adapting a source classifier so that the decision boundary is adjusted to fit the target domain. These techniques nevertheless mostly show marginal performance improvement after domain adaptation, due to a mixed of positive and negative transfers among different types of concepts, resulting in difficulty in interpreting the merit of transfer learning. After all, there is no clear picture of when a classifier should be re-developed (because of negative transfer) or adapted from a different domain (because of positive transfer).

This paper explores different workable parameters, in a quantitative way, which might impact the necessity of applying transfer learning. These parameters include the degree of shift between the data distributions of source and target domains, number of positive examples available in the target domain, concept category, visual diversity of a concept, and context variation between domains. Some of these parameters are partially utilized for transfer learning in the literature, for example in [3] by aligning data distributions through multi-kernel learning, and in [4] by visual similarity for sampling suitable training examples from source domain to enrich the limited examples in target domain for training classifiers. However, to the best of our knowledge, there are no rigorous studies yet on how the joint exploration of these parameters for either transfer learning or predicting domain adaptivity.

The main contribution of this paper is offering of a throughout analysis for different factors that could lead to domain gap with evidential support from empirical findings of several large video (VIREO-374 and TRECVID 20111) and image datasets (ImageNet and Flickr-SF). Especially, the analysis results in joint leveraging of various parameters for developing an effective technique for predicting negative (and positive) transfer of concept classifiers.

2. THE ORIGIN OF DOMAIN GAP?

We first briefly describe the following five parameters which could possibly be the factors that affect the effectiveness of transfer learning.

Data distribution mismatch (DM) is generally regarded as a crucial factor. When the training and testing data share similar data distribution, the performances of classifiers are expected to be high. Thus, the approach such as DTSVM [3] aims to narrow the distribution mismatch by learning multiple kernels. Maximum Mean Discrepancy (MMD) is a measure [1] that often employed for estimating data distribution between two domains based on the distance between sample means in a Reproducing Kernel Hilbert Space.

Number of positive training samples (TS) in the target domain gives a clue of whether a classifier should be re-developed. When very few training examples are available, a learnt classifier will have larger variance and thus higher prediction error. By incorporating the data from source domain, the variance could be reduced but at the risk of increasing bias, especially when the source domain data is drawn from a different distribution. Does the bias-variance tradeoff make the performance of classifier recycling unpredictable? Is there a threshold that when the number of training examples exceed a limit, one can give up recycle and redo from scratch?

Sample diversity (SD) of a concept is characterized by the coverage of visual appearance. A source classifier trained from diverse visual samples that cover different aspects of the concept is expected to be more easily transferred to a new domain than a classifier that is trained with narrow coverage of appearance.

Concept category (CC) could range from object, person, scene to event, and so on. Generally speaking, certain category of concepts (e.g., scene) have a relatively homogeneous visual representation, while appearance of certain category (e.g., event) is domain dependent and thus likely to produce larger gap.

The use of category to speculate transfer capability of a concept is however complicated by **semantic context (SC)**. For example, the concept *desert* in LSCOM often appears under the context of Iraq war. When training *desert* classifier, contextually related concepts such as *smoke* and *military vehicles* are also learnt. Applying such context dependent concepts to a new domain (e.g., Web video) where the learnt context does not exist is likely to result in negative transfer.

3. EMPIRICAL INSIGHTS

In this section, we provide empirical insights on the impact of these five parameters towards domain gap. We use A-SVM [10] for the experiments due to its simplicity and capability of generating compatible performance as more computationally expensive approach such as DTSVM [3]. Given few training examples from target domain, A-SVM basically seeks for additional support vectors from the new data to adjust the original decision boundary of a classifier. It optimizes the tradeoff that the new boundary should be close to the original one while being able to correctly classify the new training examples.

All the source classifiers are learnt with SVMs using five visual features: grid-based color moment and wavelet texture, three version of bag-of-visual-words (BoW) based on 1×1, 2×2 and 3×1 spatial layout. BoW is generated from SIFT of local interest points extracted by Difference-of-Gaussian (DoG) and Hessian Affine detectors. Late fusion is used to combine the five classifier scores.

3.1 Dataset

We use TRECVID 2011 (TV11) as the target (Web video) domain, while classifiers from VIREO-374 [6] (news video), ImageNet [2] (Web images with clean labels) and Flickr-SF [12] (Web images with loosely-tagged labels) as the source domain classifiers. TV11 contains 137,327 video shots as testing samples. Among the four datasets, there are 21 concepts which share common definition and each of them has a minimum of 100 positive training examples. These concepts are used in the experiments.

3.2 Cross-over Point Analysis

We first perform "cross-over point" analysis on the three parameters: TS, SD and CC. The analysis aims to reveal the trend on how many positive examples are required from target domain such that re-training of classifier is preferable than recycling. The cross-over point refers to a break even point where "redo" starts to surplus the performance of "recycle". Figure 2 shows the results when only TS is

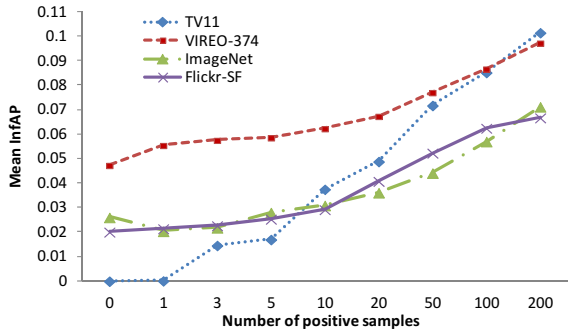

Figure 2: Comparison between recycling existing classifiers (VIREO-374, ImageNet, Flickr-SF) and redeveloping a new classifier (TV11) with the increase of positive training examples from the target domain.

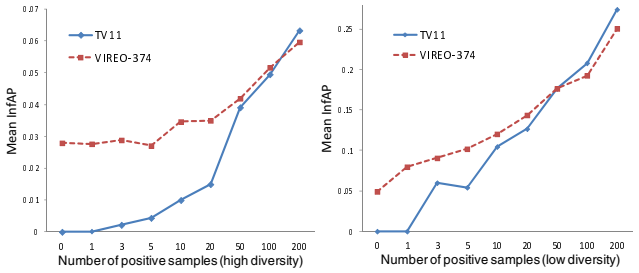

Figure 3: Comparison between concepts with high and low diversity.

considered – the cross-over points are observed on all three source domains. For the transfer from VIREO-374 to TV11, the cross-over point on average is happened around when there are 100 positive samples. This value ranges from 50 to 200 dependent on concepts. For image-to-video transfer, the cross-over points happen as earlier as before 10 positive examples, and there is no apparent difference between whether the classifiers are trained from clean (ImageNet) or noisy labels (Flickr-SF). The results give a clue that transferring from image-to-video is much harder than video-to-video.

We further experiment the fluctuation of cross-over points for concepts with high and low visual diversity. We estimate the SD of a concept directly from the training samples that are used for learning the source classifier. We employ the method in [2] by averaging the training samples and then measuring the diversity based on the size of lossless JPG file size on the average image. Figure 3 shows the results for five most diverse and homogeneous concepts on VIREO-374. Basically cross-over point for diverse concepts (130) are observed much later than of less diverse (50). We repeat the same experiment for CC on concept categories scene and event using VIREO-374. As observed, event concepts introduced a larger gap where cross-over point is observed at 60 earlier than scene at 100.

3.3 Correlation Analysis

In this section, we further conduct experiment to measure the correlation of DM, SD and SC with cross-over point. For DM, we use MMD to estimate the data distribution mismatch on the late fusion of five visual features. For SC, two context graphs are constructed for source and target domains respectively. The graph is composed of 283 concepts as nodes and pairwise concept correlation as edge weights. For VIREO-374 and TV11, the concept correlation is derived from the ground-truth of training dataset. For ImageNet, the concept correlation is measured by WordNet semantic similarity based on WUP [11], while for Flickr-SF the correlation is calculated by using Flickr Context Similarity (FCS) [5]. With the context graph, each concept in a domain is represented as a vector of 283 dimension, indicating the relationship of this concept to 283 concepts. An entry in the vector is set as the correlation value between two concepts, and zero otherwise if there is no edge directly connecting them in the graph. In our implementation, only top-5 concepts with the highest correlation are considered. Cosine similarity is then employed to measure SC relatedness between two concepts.

Table 1: Correlation of Data distribution mismatch (DM), Sample Diversity (SD), Semantic Context (SC) with cross-over point.

factor	Flickr-SF	ImageNet	VIREO-374
DM	-0.127	-0.182	-0.183
SD	-0.173	-0.198	-0.177
SC	0.162	0.147	0.138

Table 1 summaries the result. DM and SD are consistently observed as having negative correlation with cross-over point for the three source domains. In other words, when the degree of mismatch (diversity) is wider (smaller), small amount of training examples from target domain is already enough to re-train a new classifier that bypasses the performance of recycling old classifier. Similar observation is also for SC, where there is a consistent positive correlation with cross-over point for all three domains. This implies that the more contextually related two concepts, the more likely that one can benefit from recycling a classifier.

4. PREDICTING NEGATIVE TRANSFER

The empirical insights from the previous section basically hint that the correlation between the five parameters and the domain gap is not random. This gives rise to the possibility that the prediction of negative transfer is possible if any of the parameters are available for use. Here, we perform the prediction by joint utilization of DM, TS, SD and SC. Specifically, we train a SVM with these parameters as input and a binary decision as output. The input feature is a vector of nine dimensions, where the first six dimensions correspond to the normalized MMD values of five visual features and their average fusion, and the next three dimensions are the values for diversity, context relatedness and the number of given positive examples from target domain.

4.1 Experiment Settings

Similar to Section 3, we use TV11 dataset as target domain and the remaining datasets as source domain. We generate the training data for prediction directly from source domains. For example, the performance of classifiers from source A is simulated on source B. Based on the simulation similar to Section 3, we manually label whether applying a classifier from A to B will result in negative transfer, by each time varying the number of provided positive training examples given by B in the range of 0, 1, 3, 5, 10, 20, 50, 100, 200, 500 and 1000. The learnt prediction classifier is then

Table 2: Prediction accuracy on testing data. The row indicates source domain (A), and the column indicates the simulated domain (B). See main text for experiment setting.

	Flickr-SF	ImageNet	VIREO-374
Flickr-SF	-	0.694	0.804
ImageNet	0.740	-	0.802
VIREO-374	0.777	0.773	-

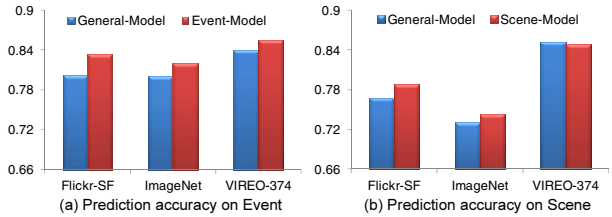

Figure 4: Comparison of category-specific models (Event-Model and Scene-Model) and a model learned using training examples from all categories (General-Model).

experimented against the testing data from TV11. Similar to training data generation, the testing data is generated assuming the given number of positive examples is in the range from 0 to 1000. In total, there are 693 samples being generated from TV11 for testing. The ground-truth of testing samples is based on the simulation result in Section 3.

Note that when applying the prediction classifier to TV11, we basically do not assume any prior knowledge in target domain. For example, the MMD of a concept is measured on the training data (with labels) of source domain and test data (without labels) of target domain. Basically, we randomly sample 100 examples from target domain for the estimation of MMD. For SD, the diversity value are obtained from the source domain. For SC, the context graphs are constructed from training set (for VIREO-374), WordNet (for ImageNet), and FCS (for Flickr-SF).

4.2 Results

Table 2 shows the results of predicting negative transfer on TV11. The results are encouraging in the way that, even when using the training data from Web image domain, the prediction can reach an accuracy of approximately 75%. To verify the performance, we use a rule-based classifier as baseline. The classifier naively predicts negative transfer if the number of training samples (TS) is more than a given threshold. By repeating the experiment with different thresholds (from 0 to 1000), on average our approach shows 28% (VIREO-374), 19% (ImageNet) and 22% (Flickr-SF) of improvement compared to baseline. Figure 4 further shows the performance for category-specific prediction. Basically models that are learnt specifically for concepts from event and scene categories offer higher prediction accuracy.

With the prediction of negative transfer, adaptive transferring of source classifiers become possible. Table 3 compares the performance of adaptive transfer to directly using source classifiers (Do-Nothing), re-training all classifiers (All-Redo) and recycling all classifiers with A-SVMs (All-Recycle). The performances of All-Redo and All-Recycle are both better than that of Do-Nothing on all three source

Table 3: Comparison of fixed and adaptive transferring of source classifiers. The performance is averaged over different numbers of positive examples provided by target domain.

	Flickr-SF	ImageNet	VIREO-374
Do-Nothing	0.021	0.027	0.048
All-Redo	0.054	0.054	0.054
All-Recycle	0.054	0.043	0.074
Adaptive-Transfer	**0.057**	**0.056**	**0.076**

datasets. The performance of All-Recycle is better than All-Redo when transferring from video-to-video. While for image-to-video transfer, All-Redo is superior to All-Recycle due to the domain gap. For all the three source domains, adaptive transfer shows the best performance.

5. CONCLUSION

Basing on the speculation of five parameters, this paper provides empirical analysis to verify the "contribution" of these parameters towards domain gap based on cross-over point and correlation analysis. The analysis eventually leads to the development of a prediction classifier that attains encouraging performance when transferring classifiers from news video (VIREO-374) and Web image (ImageNet and Flickr-SF) domain to Web video (TRECVID 2011) domain. By jointly applying the prediction technique with transfer learning technique (A-SVMs in our case), improvements are consistently observed compared to blind re-developing and blind recycling of classifiers.

6. ACKNOWLEDGMENTS

The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119709).

7. REFERENCES

[1] K. M. Borgwardt, A. Gretton, M. J. Rasch, H.-P. Kriegel, B. Schölkopf, and A. J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22:49–57, 2006.

[2] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[3] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank. Domain transfer svm for video concept detection. In *CVPR*, 2009.

[4] W. Jiang, E. Zavesky, S.-F. Chang, and A. Loui. Cross-domain learning methods for high-level visual concept classification. In *ICIP*, 2008.

[5] Y.-G. Jiang, C.-W. Ngo, and S.-F. Chang. Semantic context transfer across heterogeneous sources for domain adaptive video search. In *ACM MM*, 2009.

[6] Y.-G. Jiang, C.-W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *CIVR*, 2007.

[7] M. R. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE Multimedia*, 2006.

[8] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trevid. In *ACM MIR*, 2006.

[9] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on PAMI*, 2000.

[10] J. Yang, R. Yan, and A. G. Hauptmann. Cross-domain video concept detection using adaptive svms. In *MM*, 2007.

[11] W. Zhibiao and M. Palmer. Verb semantic and lexical selection. In *ACL*, 1994.

[12] S. A. Zhu, C.-W. Ngo, and Y.-G. Jiang. Sampling and ontologically pooling web images for visual concept learning. *IEEE Trans. on Multimedia*, 2012.