THE DATA DELUGE: CHALLENGES AND OPPORTUNITIES OF UNLIMITED DATA IN STATISTICAL SIGNAL PROCESSING

Michael L. Seltzer[1] and Lei Zhang[2]

1. Speech Technology Group
Microsoft Research
Redmond, WA USA

2. Web Search & Mining Group
Microsoft Research Asia
Beijing, China

ABSTRACT

Recently, there has been a dramatic increase of the amount of audio, video, and images created and shared on the internet by users around the world. Much of this content is publicly available and free of cost. When viewed through the lens of pattern classification, this content can be seen as a virtually unlimited supply of training data for various statistical modeling and labeling tasks such as speech recognition and computer vision. In order to effectively exploit this data resource, significant research challenges must be addressed. In this paper, we present three significant challenges that must be solved to harness the potential of this "data deluge". We then describe recent work in spoken language processing and image processing that has begun to address these challenges in order to tackle large-scale classification tasks. By bringing together the work of these two communities, we hope to stimulate the cross-pollination of ideas and methods among different signal processing communities.

Index Terms— data deluge, web-scale data, pattern recognition, multimedia search

1. INTRODUCTION

Over the last several years, there has been a dramatic increase in the amount of audio and video on the internet. Users around the world are creating and sharing content at an astounding pace. By 2006, the number of podcasts managed by the podcast aggregator Feedburner exceeded the number of radio stations in world [4]. In 2007, approximately 11 million pictures were being uploaded to Flickr each day [23]. Users are swimming in a deluge of audio, video, images, and text. The availability of this virtually unlimited supply of multimedia content represents a tremendous opportunity to revolutionize the way in which systems which rely on training data, e.g. statistical pattern classifiers, are designed, built, and used.

Turning this unlimited supply of content into useful training data presents significant research challenges. The data are often unlabeled or labeled with simplistic keywords. The cost of manually labeling the data is prohibitive. In addition, quality of the data and the accuracy of the labeling can vary significantly across different sources and the sheer quantity of available data can overburden available computational resources. Thus, in order to fully exploit the potential of the multimedia data deluge, new algorithmic approaches to training, data selection, and labeling are required.

In this paper, we describe three significant challenges of the data deluge. We then describe recent work in spoken language processing and image processing that has begun to address these challenges to tackle large-scale statistical modeling and labeling problems. Finally, we discuss some open issues and the outlook for the future.

2. THREE CHALLENGES OF THE DATA DELUGE

2.1. So much data, so few labels

While a virtually unlimited supply of data on the web exists, the vast majority of it is either unlabeled or labeled in a manner not suitable for the task. Two main approaches to handling unlabeled data have been proposed in the literature. The first is called *active learning* [9], also known as selective sampling. In this task, the unlabeled data is analyzed automatically to determine which data should be selected for manual labeling. Usually the goal of active learning is to find the data that your current system does a poor job of classifying. An alternative approach to handling unlabeled data is *semi-supervised* learning, e.g. [32], where a large set of unlabeled data is pooled with a usually much smaller set of labeled data. The labeled data is used to infer the labels for the rest of the data. These two methods are complementary and can easily be combined to generate labels for all of the training data.

2.2. Finding the relevant (and correctly labeled) needles in the haystack

Because of the enormous scope of content on the web, simply finding the content that is most relevant to your task is a challenging problem. Once the relevant data is found, it is critically important to verify in some automated way that the labels, if present, are correct. Obviously, data that is mismatched to your task domain or has erroneous labels will degrade the performance of your system. Current data-finding methods focus on automatic ways to construct web queries to find the most relevant data. The problem of noisy labels can be addressed either through training algorithms that are robust to labeling errors [22] or algorithms specifically designed to find and reject mislabeled data [26].

2.3. Making media as searchable as text

To make content-based search of audio, image, or video as simple and effective as text-based search today, a classification or labeling scheme must be able to operate on web scale data

efficiently and robustly and represent the content in a manner suitable for indexing, ranking and search. We note that by coming up with effective solutions for this problem, the difficulties of the previous two problems will be reduced.

3. APPROACHES TO THE DATA DELUGE IN SPOKEN LANGUAGE PROCESSING

3.1. Model training with unlabeled or noisy data

In spoken language processing, active learning has been proposed a method for training models at multiple points in the processing chain, i.e. the acoustic and language models of the recognizer itself, and the spoken language understanding (SLU) component that tries to extract the semantic concept or intent from the recognizer's hypothesized word string.

Initial experiments in active learning for acoustic modeling were performed by Kamm and Meyer [24]. They proposed a word-level confidence measure to decide which segments of audio to manually transcribe [25]. Experiments on small data sets showed that performance comparable to or better than a system trained on a fully labeled training set could be achieved.

More recently, utterance-level confidence measures were used for active learning in acoustic modeling by Riccardi and Hakkani-Tur [12]. Experiments were performed on utterances from a spoken dialog system and consistent improvements over random sampling were observed. However, here too, the sample size was relatively small, with a training set of about 25 hours. They also showed that active learning was shown to be faster than random sampling at learning new words and new n-grams in language modeling [12].

For many speech recognition tasks, just recognizing the correct words in the transcript is not enough. The underlying *intent* of the user must be determined. In [10], Tur et al. proposed a strategy for unlabeled data that combined active learning and semi-supervised learning for a boosting classifier used for call routing that classified utterances into one of 49 call types. Two methods of active learning that are robust to noisy labels were proposed by Raymond and Riccardi [7]. These methods showed significant improvement over conventional active learning on two SLU tasks.

Active learning can be augmented by semi-supervised learning. In speech recognition, this is typically performed by decoding the unlabeled data using a model trained from labeled data [13]. This data is pooled with original set of labeled data to generate a new model. Thus, active learning can be used to selectively transcribe a subset of the unlabeled data and semi-supervised learning can be used to label the rest. Note that this is closely related to unsupervised model adaptation using traditional techniques except that in this case, the model is adapted to unlabeled training data not unlabeled test data, as in model adaptation schemes.

Motivated by the success of game-based labeling systems [21], Paek et al. recently proposed a game-based approach to language model data collection [27]. They proposed a game to elicit alternative wordings and paraphrases for business listings in a directory assistance application. On a limited test set, they showed performance that was statistically the same as that obtained by transcribing a set of received phone calls manually. Such approaches are appealing as they can be deployed by any user with minimal cost. However, the key to the success of these methods is finding a game that will both suitably label the data you have and be appealing enough to attract a significant body of users.

3.2. Finding the data you want

Because copious amounts of text data on the web is easily available and by definition already labeled, the primary challenge in exploiting the web data for language modeling is actually finding text that is well matched to the domain and topic of interest. This is a "needle in the haystack" problem as the data that is relevant to a particular task represents only a tiny fraction of available web data. Early work in using web data to build language models was done by Berger and Miller [2] and Zhu and Rosenfeld [31]. Both of these used the web to estimate n-gram counts for a LM for the news domain.

More recently, interest has shifted to using the web to find text that can be used to build models of a specific topic and/or style of speech. Sethy et al. proposed a mechanism for query generation based on the relative entropy of n-gram histories between a small in-domain topic LM and a larger generic background LM [3]. N-gram histories with large relative entropy were used as the basis for web queries. The documents returned from a web search with these queries were then used to augment the original language model. This resulted in a 14% relative reduction in WER on a task in the medical domain.

Bulkyo et al. devised a method to mine the web for text that matches the conversational style of phone calls or meetings [15]. Frequently occurring n-grams in the initial language model were used as web queries and the text from the resulting web pages were used to augment the existing language model training data. A small topic-specific language model was used to capture topic-dependent words and n-grams. This work was shown to be effective for building language models in both English and Mandarin.

3.3. Indexing and browsing spoken documents

The growth of podcasts has led to the need for a method to efficiently search the content of podcasts. Currently most podcasts are tagged with manually labeled metadata tags created by the podcast creator. However, it has been proposed that more accurate search of audio could be obtained via speech recognition, e.g. [20]. However, until recently, the number of audio or video documents in a collection was relatively small so the emphasis was on improving recognition accuracy rather than on efficient and robust representations of the audio. Recently, Chelba et al. proposed that a method for representing the word lattice of recognition hypotheses that is suitable for indexing and relevance ranking by a search engine [6]. Yu et al. extended this work and achieve additional reductions in size of the index, increasing search efficiency [34].

4. APPROACHES TO THE DATA DELUGE IN IMAGE PROCESSING

While the challenges of the data deluge across different media types are similar at a high level, the approaches taken can be significantly different. We now highlight some recent work in image processing that addresses the challenges of web-scale data.

4.1 Web as a repository of training data

The explosive growth of multimedia data and the phenomenal success in web search have had a great impact on research in computer vision and multimedia. In the past several years, we have seen exciting progress and new potential for leveraging the web as a repository of training data.

In *AnnoSearch*, 2.4 million Web images were crawled and a system was built to index the images as a knowledge base for image annotation [30][29]. Rather than training a limited number of visual concept models using traditional supervised learning techniques as in most previous work, Wang et al. reformulated image annotation as a novel two-step process: given an input image, first searching for a group of similar images in a large scale Web image database, and then mining key phrases extracted from the descriptions of the images. A method for high-dimensional indexing based on clustering was used to find a group of visually similar images efficiently [30]. As the result of the proposed search-based annotation, annotating with an unlimited vocabulary becomes possible.

In [8], Wang et al. further studied how to reduce the so-called *semantic gap* to improve the search for web images. Following on the success of text search on the web, a method was proposed to learn a new ranking-based distance measure in the visual space to approximate the distance measure in the textual space. The learned distance measure was used in the 2.4 million image database and demonstrated improved performance in both image retrieval and annotation.

Motivated by billions of images freely available online, Torralba et al. built a large dataset of 80 million images collected from the web for non-parametric object and scene recognition [1]. The images in the dataset are stored as 32×32 pixel color images. Each image is loosely labeled with one of the 75,062 non-abstract nouns listed in the Wordnet lexical database, which is believed to cover all visual object classes. Given an input image, the system searches in the 80 million image database to find its K nearest neighbors and use the associated labels to infer the semantic classes of the input image for object and scene recognition.

Mobile phones with embedded cameras are popular nowadays and have huge growth potential. Most current services for information acquisition on mobile devices use text-based input. Nevertheless, sometimes it is difficult for users to describe their information needs in words. Instead of current flat query modes, camera phones can support much richer queries, not only text but also images.

In Photo2Search, Fan et al. designed a web service to support users to search for relevant information on the Web via photos of what they see, for example a picture of a restaurant or a hotel [28]. They first collected millions of images with latitude/longitude metadata and indexed the images using local features. When a query image is input, the system extracts its local features and searches for a number of nearest neighbors in the database. After checking the geometric relationship of features, the system gets a list of images matched with the query image. The information associated with the corresponding location will be returned to users.

Similarly, in IM G2GPS, Hays et al. leveraged a dataset of over 6 million GPS-tagged images from the internet and proposed an algorithm for estimating a distribution over geographic locations from a single image using a purely data-driven scene matching approach [18]. They showed that geo-location estimates can provide the basis for numerous other image understanding tasks such as population density estimation, land cover estimation or urban/rural classification.

4.2 Removing label noise from web image collections

The success of text-based image search has shown the power of text information associated with web images. However, using the text surrounding an image on a webpage to label the image can be problematic because the text does not necessarily describe the content of the images. Thus, how to cleverly choose suitable web images and filter noisy labels is a crucial problem that needs to be addressed in different applications.

In Wang's AnnoSearch, the 2.4 million images were collected from several online photo forums because the images are of high quality and have rich descriptions, such as title, category and comments provided by photographers. To extract meaningful key phrases from this text, a text mining technology called search result clustering is employed to find dominant concepts and remove noisy labels in the search result [14]. In the work of visual distance learning described in [8], Latent Dirichlet Allocation [11] was used to measure the similarity between two short text descriptions associated with Web images, which has proven useful as a measure of semantic distance.

Label noise was also addressed in Torralba's work with 80 million tiny images mentioned in Section 4.1. First, Wordnet was used to provide a comprehensive list of 75,062 visual classes by extracting all non-abstract nouns. Then, seven independent image search engines were employed to download all images for all of these non-abstract noun queries. Running over 8 months, this method gathered about 80 million images in total. For each downloaded image, the query word is treated as its label as the resulting images are the most relevant images ranked by image search engines. Such a fully automatic process will inevitably introduce noisy labels. Therefore, in the recognition step (after the system returned K nearest neighbors of a query image), a voting scheme was proposed based on the Wordnet semantic hierarchy to output class labels at appropriate semantic levels.

4.3 Indexing challenges

Most existing Content-Based Image Retrieval (CBIR) systems suffer from a scalability problem and cannot scale to millions or billions of images because it is difficult to build an effective index for high dimensional image features. Motivated by the success of web search engines, many researchers have tried to map image retrieval problems to text retrieval problems, hoping to utilize the existing text-based indexing and ranking schemes. The basic idea is to map image features to words [19]. Typically images are first represented by local features, and then by clustering, each local feature is mapped to a discrete keyword. With this representation, comparing two images become matching words in them, and therefore, a text-based search engine can be utilized to reduce the computational and memory cost.

5. OPEN QUESTIONS AND NEXT STEPS

While the work in speech and image processing described in this paper shows that significant progress has been made in addressing the challenges of the data deluge, it is clear that many open questions remain. Interestingly, while the two fields explored in this paper share many challenges, they also have made progress in complementary areas.

In speech recognition, there has not been any work on acoustic modeling or spoken language understanding on web scale data sets. Current state of the art systems are built from a few thousand hours of data, but to date, no one has been able to make use of an orders of magnitude more data. Whether or not these large data systems will benefit from active and semi-supervised learning in the same way as the smaller systems remains an open question. In

order to most effectively utilize this abundance of data, perhaps new core modeling approaches will be required. For example, some researchers have proposed revisiting the use of discrete non-parametric distributions for efficiently building models with huge amounts of data [5][16].

On the other hand, image processing researchers have been able to performs experiments with web-scale data sets but have been struggling with the issues of representation [33]. That is, how can a content-based feature representation be mapped to a searchable text representation that captures the semantic content of the image while being suitable for indexing, ranking and search?

It is also interesting to consider the data deluge that we have described on a web scale is also occurring at a personal level. These days, a user can easily have a personal archive of tens of thousands of images, and hundreds of hours of video. Current research projects such as the SenseCam [17] that record the audio and video of practically every minute of daily life will increase this amount of personal data several orders of magnitude. Capturing this data is useless without an effective means of indexing and search.

Finally, while we have focused this paper on the algorithmic challenges of the data deluge, we have not addressed the corresponding computational challenges. It is apparent that processing data on a web scale requires significant processing resources. If the barrier to entry for conducting research with web scale data is prohibitively high, the pace of progress in this area will be slowed down.

Overall, we believe the deluge of audio, video, and image content on the web represents a watershed moment for multimedia signal processing and pattern recognition. By effectively addressing the challenges of dealing with web scale data, we can potentially discover solutions to challenging problems previously believed unsolvable.

7. REFERENCES

[1] A. A. Torralba, R. Fergus, W. -T. Freeman, "80 million tiny images: a large dataset for non-parametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[2] A. Berger and R. Miller, "Just-in-time language modeling", in *Proc. ICASSP 1998,* vol. 2, pp. 705-708, Seattle, WA, May 1998.

[3] A. Sethy, P. G. Georgiou, and S. Narayanan, "Building topic specific language models from webdata using competitive models," in *Proc. Interspeech 2005*, pp. 1293-1296. Lisbon, Portugal, Sept. 2005.

[4] B. Charny, "FeedBurner: podcasts outnumber radio stations," *eWeek.com,* Apr. 17, 2006

[5] B. Mak, S.-K. Au Yeung, Y.-P. Lai, M. Siu, "High density discrete HMM with the use of scalar quantization indexing," in *Proc. Interspeech 2005,* pp. 2121-2124, Lisbon, Portugal, Sept. 2008

[6] C. Chelba and A. Acero, "Position specific posterior lattices for indexing speech," in *Proc. ACL 2005,* Ann Arbor, MI, June 2005.

[7] C. Raymond and G. Riccardi, "Learning with noisy supervision for spoken language understanding," in *Proc. ICASSP 2008*, Las Vegas, NV, April 2008.

[8] C. Wang, L. Zhang, H.-J.-J.Zhang, "Learning to Reduce the Semantic Gap in Web Image Retrieval and Annotation,," *ACM SIGIR Conference* (*SIGIR*), Singapore, 2008

[9] D. Cohn, L. Atlas, and R. Ladner, "Improving generalization with active learning" *Machine Learning,* vol. 15, no. 2, pp. 201-221. May 1994.

[10] D. Hakkani-Tur, G. Riccardi, and G. Tur, "An active approach to spoken language processing," *ACM Trans. on Speech and Lang. Proc.*, vol. 3, no. 3, pp. 1-31, Oct. 2006.

[11] D.M. Blei, D. M., A.Y. Ng, A. Y., and M.I. Jordan, M. I. "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, Mar. 2003.

[12] G. Riccardi and D. Hakkani-Tur, "Active Learning: Theory and Applications to Automatic Speech Recognition," *IEEE Trans. on Speech and Audio Proc.*, vol. 13, no. 4, pp. 504-511, July 2005.

[13] G. Tur, D. Hakkani-Tur, and R. Shapire, "Combining active and semi-supervised learning for spoken language understanding," *Speech Communication*, vol. 45, no. 2pp. 171-186, 2005.

[14] H.-J. Zeng, Q.C. He, Z. Chen, and W.--Y. Ma, "Learning To Cluster Web Search Results,". *ACM SIGIR Conference* (*SIGIR*), 2004.

[15] I. Bulyko, M. Ostendorf, M. Sui, T. Ng, A. Stolcke, and O. Cetin, "Web resources for language modeling in conversational speech recognition," *ACM Trans. on Speech and Lang. Proc.*, vol. 5, no. 1, art. 1, Dec. 2007.

[16] J. G. Droppo, M. L. Seltzer, A. Acero, and Y.-H. Chiu, "Towards a non-parametric acoustic model: an acoustic decision tree for observation probability calculation," in *Proc Interspeech 2008*, Brisbane, Australia, Sept. 2008.

[17] J. Gemmell, L. Williams, K. Wood, R. Lueder, and G. Bell, "Passive capture and ensuing issues for a personal lifetime store," in *Proc. ACM Workshop on Continuous Archival and Retrieval of Personal Experiences*, New York, NY, Oct. 2004.

[18] J. Hays, A. A. Efros. "IMG2GPS: estimating geographic information from a single image," *IEEE International Conference. on Computer Vision and Pattern Recognition* (*CVPR*), Anchorage, 2008.

[19] J. Sivic, J. and A. Zisserman, A. "Video Google: A Text Retrieval Approach to Object Matching in Videos," *International Conference on Computer Vision* (*ICCV*), 2003.

[20] J.-M. Van Thong, P. J. Moreno, B. Logan, B. Fidler, K. Maffey, and M. Moores, "Speechbot: an experimental speech-based search engine for multimedia content on the web," *IEEE Trans. on Multimedia*, vol. 4, no. 1, pp. 88-96, Mar. 2002.

[21] L. von Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. CHI 2004*, pp. 319-326, Vienna, Austria, Apr. 2004.

[22] M. Kearns, "Efficient noise-tolerant learning from statistical queries" *Journal of the ACM*, vol. 45, no. 6, pp. 983-1006, Nov. 1998.

[23] P. Saugvignon, "Everything you wanted to know about Flickr but were afraid to ask," The Juicy Cow Blog. Retrieved from http://www.thejuicycow.com/2007/08/29/everything-you-wanted-to-know-about-flickr-but-were-afraid-to-ask/

[24] T. M. Kamm and G. G. L. Meyer, "Selective sampling of training data for speech recognition," in *Proc. HLT 2002,* San Diego, CA, Mar. 2002.

[25] T. M. Kamm and G. G. L. Meyer, "Word-selective training for speech recognition," in *Proc. ASRU 2003,* pp. 55-60, Dec. 2003.

[26] T. Nakagawa and Y. Matsumoto, "Detecting errors in corpora using support vector machines," in *Proc. Int. Conf. on Comp. Ling.*, vol. 1, pp. 1-7, Taipei, Taiwan, Aug. 2002.

[27] T. Paek, Y.-C. Ju, and C. Meek, "People watcher: a game for eliciting human-transcribed data for automated directory assistance," in *Proc. Interspeech 2007,* pp. 1322-1325, Antwerp, Belguim, Aug. 2005.

[28] X. Fan, X. Xie, Z. Li, M. Li, and W.–Y. Ma, "Photo-to-Search: Using Multimodal Queries to Search the Web from Mobile Devices," *7th ACM SIGMM Workshop on Multimedia Information Retrieval* (*MIR*), 2005.

[29] X. Li, L. Chen, L. Zhang, F. Lin, W.-Y. Ma,: "Image Annotation by Large-Scale Content-based Image Retrieval,," *Proc. of ACM Int. Conf. on Multimedia*, Santa Barbara, 2006

[30] X. Wang, L. Zhang, F. Jing, and W.-Y. Ma,: "AnnoSearch: Image Auto-Annotation by Search,," *Proc of the International Conference on Computer Vision and Pattern Recognition* (*CVPR*), New York, 2006

[31] X. Zhu and R. Rosenfeld, "Improving trigram language modeling with the world wide web," in *Proc. ICASSP 2001,* vol. 1, pp. 533-536, Salt Lake City, UT, May 2001.

[32] X. Zhu, "Semi-supervised learning literature survey," Univ. of Wisconsin, Madison, WI, Tech. Rep. CSTR 1530, 2008.

[33] Z. Li, X. Xie, L. Zhang, and W.-Y. Ma, "Searching one billion web images by content: challenges and opportunities," in *Proc. MCAM*, Weihai, China, July 2007.

[34] Z.-Y. Zhou, P. Yu, C. Chelba, and F. Seide, "Towards spoken-document retrieval for the internet: lattice indexing for lage-scale web-search architectures," in *Proc. HLT 2006,* pp. 415-422, New York, NY, June 2006.