Multi-modal Language Models for Lecture Video Retrieval

Anonymous submission

ABSTRACT

We propose Multi-modal Language Models (MLMs), which adapt latent variable models for text document analysis to modeling co-occurrence relationships in multi-modal data. In this paper, we focus on the application of MLMs to indexing slide and spoken text associated with lecture videos, and subsequently employ a multi-modal probabilistic ranking function for lecture video retrieval. The MLM achieves highly competitive results against well established retrieval methods such as the Vector Space Model and Probabilistic Latent Semantic Analysis. Retrieval performance with MLMs is also shown to improve with the quality of the available extracted spoken text.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms

Keywords

Multi-modal retrieval, latent variable modeling, multi-modal probabilistic ranking

1. INTRODUCTION

The continued growth in user generated video on the internet has exacerbated the need for tools to facilitate content search and management. A quickly growing sector of internet-distributed content is *expository* or "how-to" video, a genre which includes lecture videos from online courses, presentations from conferences and seminars, and more general demonstration and tutorial videos. These videos typically include multi-modal data. For example, a video of a speaker delivering a presentation has both the speech and the slide modalities.

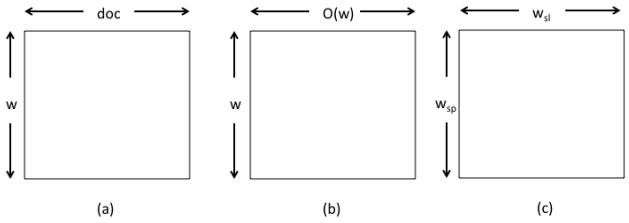

Figure 1: The left panel shows the traditional word-document matrix that is modeled using PLSA and LDA. The middle panel shows the matrix processed in word-topic modeling [5], where the co-occurrence of a word w and its neighborhood words $O(w)$ is modeled. The right panel shows the multi-modal modification we propose in which the co-occurrence of slide words and spoken words are modeled.

Lecture videos are relatively amenable to content-based indexing. Slides provide the videos with both a temporal and topical structure, and systems can exploit the slide text to enable search functionality inside the videos [1, 6, 13].

Retrieval systems have also extracted text from the spoken audio in lecture videos [8] using automatic speech recognition (ASR). Errors in ASR are commonplace due to poor audio recording quality or acoustic mismatch and can diminish the spoken text's utility for video retrieval. In some cases, spoken text is available as manually created closed captions (CC). While slides typically contain sparse and discriminative words, speech is usually voluminous, improvised, and comprised largely of generic terms.

In this work, we model the co-occurrence of words used in videos as spoken and slide text. The motivation is to effectively combine the sparse and discriminative slide text with the voluminous and generic spoken text to achieve better lecture video retrieval. Our contributions are:

- We propose Multi-modal Language Models (MLMs) to represent the co-occurrences of multi-modal data using latent variable modeling.
- We propose a multi-modal probabilistic ranking function, for use with the MLMs for lecture video retrieval.
- We introduce the Google I/O Dataset, a new dataset for studying multi-modal lecture video retrieval.

2. RELATED WORK

The Vector Space Model (VSM) for information retrieval treats documents as "bags" of words [10] and drives state-of-the-art text search systems including Lucene[1]. Representing documents as vectors leads naturally to modeling the document corpus as a matrix. Figure 1(a) depicts such a matrix in which the rows and columns are indexed by the words and documents respectively. To model inter-word relationships, the word-document matrix in Figure 1(a) can be processed to build latent variable models. Both Probabilistic Latent Semantic Analysis (PLSA) [7] and Latent Dirichlet allocation (LDA) [4] model documents using a distribution over latent variables (topics) to capture relations between co-occurring words. In [5], Chen proposed Word Topic Models (WTMs) which explore word co-occurrences locally within a document, as shown in the word-word matrix of Figure 1(b). The WTMs implement latent variable models on a finer sub-document level, thus demonstrating improved performance over conventional PLSA. While [5] constructs latent topic models for (unimodal) text documents, our MLMs are learned from multi-modal text as in Figure 1(c) and detailed below.

Researchers have in turn applied latent variable models to multi-modal domains. Many prior works explore the relation between images' visual features and their text annotations using variants of LDA and PLSA. Barnard et al. [2] developed multi-modal LDA to jointly model a common underlying topic distribution on image region descriptors and annotation words (i.e., tags). Blei and Jordan proposed Correspondence LDA (Corr-LDA) [3], which models a process that first generates region descriptors followed by generation of words (each word is linked to one of the image regions). [11] develops a less constricted multi-modal LDA model allowing for different latent variable distributions in each modality and using regression to more flexibly capture inter-modality relationships.

[9] proposes multi-layer PLSA to model visual features and tags. The multi-layer PLSA introduces two layers of latent variables (one being common to the two modalities) into the joint model, and does not require that tags associated with images necessarily describe the visual content. Rasiwasia et al. [12] use canonical correlation analysis (CCA) to model multi-modal data by jointly performing dimension reduction across the two modalities of words and pictures. The intermediate subspace search of CCA is suited for a scenario when there is no natural correspondence between representations in different modalities.

In this work, we build generative models for individual spoken words and slide words, which we call Multi-modal Language Models (MLMs). The MLMs are also learned using Expectation Maximization (EM). In contrast to conventional PLSA or LDA methods which implement latent variable models on the documents of the corpus, the training of MLMs is based on multi-modal word-word co-occurrences. This more direct formulation greatly simplifies both model training and retrieval.

3. MULTI-MODAL LANGUAGE MODELS

Multi-modal Language Models (MLMs) are a latent variable model learned from the *multi-modal* data matrix of Figure 1(c). The matrix entries indicate the number of times a spoken word and a slide word co-occur in the same video's transcripts. A set of latent variables models the essential relationships between slide words and spoken words. Following the graphical model in Figure 2, the joint probability of a slide word w_{sl} and a spoken word w_{sp} is given by:

$$p(w_{sp}, w_{sl}) = \sum_z p(w_{sp}|z)p(z|w_{sl})p(w_{sl}) \ . \quad (1)$$

We use EM to maximize the likelihood of the data co-occurrence matrix in Figure 1(c). Following the derivations in [7], the EM iterations are described below.

E-Step: Update

$$p(z|w_{sp}, w_{sl}) = \frac{p(w_{sp}|z)p(z|w_{sl})}{\sum_z p(w_{sp}|z)p(z|w_{sl})} \ , \quad (2)$$

M-Step: Update

$$p(w_{sp}|z) = \frac{\sum_{w_{sl}} \#(w_{sl}, w_{sp})p(z|w_{sl}, w_{sp})}{\sum_{w_{sl}, w_{sp}} \#(w_{sl}, w_{sp})p(z|w_{sl}, w_{sp})} \ , \quad (3)$$

$$p(z|w_{sl}) = \frac{\sum_{w_{sp}} \#(w_{sl}, w_{sp})p(z|w_{sl}, w_{sp})}{\#(w_{sl})} \ . \quad (4)$$

The result of EM training is the MLM for the slide and spoken words in the corpus following (1).

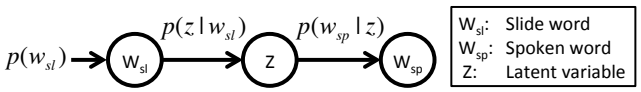

Figure 2: The graphical model of our proposed MLM. The relation between each slide word and spoken word is described by a set of latent variables. The model parameters $p(z|w_{sl})$ and $p(w_{sp}|z)$ can be estimated by the EM algorithm, while the prior of $p(w_{sl})$ is directly obtained from the corpus.

4. RETRIEVAL RANKING FUNCTION

Using the MLM, we propose a probabilistic multi-modal ranking function for lecture video retrieval. Each lecture video in the corpus is represented by its associated slide text transcript T_{sl} and spoken text transcript T_{sp}. We can utilize $p(z|w_{sp}, w_{sl})$ from (2) to compute the video specific latent variable distribution,

$$p(z|(T_{sl}, T_{sp})) = \sum_{(w_{sl}, w_{sp}) \in (T_{sl}, T_{sp})} \alpha((w_{sp}, w_{sl}), (T_{sp}, T_{sl})) \cdot p(z|w_{sp}, w_{sl}) \ , \quad (5)$$

where $\alpha((w_{sp}, w_{sl}), (T_{sp}, T_{sl}))$ represents the co-occurrence frequency of the word pair (w_{sl}, w_{sp}) observed in the specific video's transcripts (T_{sp}, T_{sl}).

For retrieval, denote the user query text as T_q. We estimate the query-video relevance by the conditional probability of T_q given (T_{sl}, T_{sp}) from the video:

$$p(T_q|(T_{sl}, T_{sp})) = \prod_{w_q \in T_q} \sum_z p(w_q|z)p(z|(T_{sl}, T_{sp})) \ . \quad (6)$$

[1] http://lucene.apache.org

Here, $p(z|(T_{sl}, T_{sp}))$ is calculated as in (5), but $p(w_q|z)$ is unknown. This reflects the intention gap between users' query language model for expressing information needs and the system's document model. It is reasonable to assume that the query words obey a similar distribution to the slide words or spoken words in the video corpus. Therefore, (6) can be rewritten, replacing $p(w_q|z)$ with $p(w_{sl}|z)$ and $p(w_{sp}|z)$ respectively:

$$p_{sl}(T_q|(T_{sl}, T_{sp})) = \prod_{w_{sl} \in T_q} \sum_z p(w_{sl}|z) p(z|(T_{sl}, T_{sp})) \,, \quad (7)$$

$$p_{sp}(T_q|(T_{sl}, T_{sp})) = \prod_{w_{sp} \in T_q} \sum_z p(w_{sp}|z) p(z|(T_{sl}, T_{sp})) \,. \quad (8)$$

We multiply $p_{sl}(T_q|(T_{sl}, T_{sp}))$ and $p_{sp}(T_q|(T_{sl}, T_{sp}))$ to define the final score for ranking videos given the query T_q:

$$\hat{p}(T_q|(T_{sl}, T_{sp})) = p_{sl}(T_q|(T_{sl}, T_{sp})) p_{sp}(T_q|(T_{sl}, T_{sp})) \,. \quad (9)$$

A notable implementation detail is that the conditional probability $p(z|(T_{sl}, T_{sp}))$ in (7) and (8) is query-independent. Therefore, it can be pre-computed once per video and stored to accelerate processing at query time.

5. EXPERIMENTS

5.1 The Google I/O Dataset [2]

To validate our multi-modal video retrieval scheme, we assembled a dataset of 209 presentation videos from Google I/O conferences in the years 2010-2012. The lengths of the videos range from 40 to 60 minutes. By crawling the conference web sites, we collected the following data for each presentation video:

- Slide text from PPT, PDF, HTML5, etc. (PPT)
- Closed-caption speech transcripts (CC)
- OCR extracted slide text from video frames (OCR)
- ASR speech transcripts from YouTube (ASR)

For automatic slide text extraction (OCR), we first match slide frames in the videos to the PPT slides using the system in [-][3]. We then use Microsoft Office document OCR to extract slide text from the matching video frames. The ASR transcript is downloaded from YouTube.

Automatically extracted OCR and ASR data are noisy versions of the actual slide and spoken text. To filter recognition errors, we discard OCR and ASR transcript words that appear only once in the corpus and are not in the english dictionary. We empirically verified that this filtering procedure does not hurt performance, while significantly reducing computation time. The lexicon from the closed caption transcripts contains 22,786 unique words, while the lexicon for the PPT transcripts contains 17,013 words. After filtering, we retain 29,279 and 36,118 words for the ASR and OCR lexicons, respectively.

Based on the talks' descriptions on the conference web sites, 275 queries were manually generated to simulate user queries. The queries are technical terms such as "listview android widget" and "NFC reader/writer API". Manual ground truth relevance judgments were compiled for all 275 queries across all 209 videos by one of the authors. We use mean average precision (mAP) [10] as the evaluation metric throughout our experiments.

[2]The Google I/O dataset will be released upon acceptance of this paper.

[3]Anonymized for double blind submission/review.

5.2 Baseline methods

We compare using MLMs for lecture video retrieval with two well established methods: VSM and PLSA. The Lucene documentation [4] or [10] describe VSM retrieval. For retrieval using PLSA, we used the following ranking function which is similar to (6):

$$p(T_q|D) = \prod_{w_q \in T_q} \sum_z p(w_q|z) p(z|D) \,, \quad (10)$$

where D denotes a video. This ranking function performed better than the folded-in latent space retrieval described in [7]. For each text modality we trained a 200 dimensional PLSA model.

For multi-modal video retrieval, we evaluate both early and late fusion strategies for both VSM and PLSA. For early fusion, the available slide and spoken text is concatenated to represent each video prior to indexing. For late fusion, retrieval scores are first computed independently for slide and spoken data, and then fused in a weighted sum:

$$S_{\text{late fusion}} = \lambda S_{sl} + (1 - \lambda) S_{sp} \,. \quad (11)$$

S_{sl} and S_{sp} represent the retrieval scores of slide and spoken text respectively, and $\lambda \in [0, 1]$ is optimized via two-fold cross validation.

5.3 Retrieval with error-free text data

The first set of retrieval experiments uses CC spoken text and PPT slide text, i.e., model training and video retrieval use error-free slide and spoken text transcripts. Table 1 shows the mAPs for the MLM trained with 200 latent variables, compared to the early and late fusion results from VSM, PLSA. The MLM significantly outperforms the second best performing method of VSM late fusion, with statistical significance at 99% confidence interval according to the paired t-test.

Table 1: Multi-modal retrieval performance using PPT and CC text on the Google I/O corpus.

	mAP@N		
	N=5	N=10	N=209
VSM early fusion	0.863	0.829	0.777
VSM late fusion	0.869	0.845	0.790
PLSA early fusion	0.858	0.830	0.767
PLSA late fusion	0.806	0.793	0.732
MLM (ours)	**0.902**	**0.875**	**0.830**

5.4 Retrieval with noisy text data

We repeat the experiments using the automatically extracted text from OCR and ASR to represent each video.

[4]`http://lucene.apache.org/core/4_0_0/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html`

Here, noisy OCR and ASR text data are employed to train the MLMs and the retrieval mAPs are tabulated in Table 2. While there is no performance gain as in the noise-free case (PDF and CC), the MLM achieves statistically indistinguishable performance to the competitive VSM late fusion baseline (paired t-test: t = 0.672, p=0.502).

Table 2: Multi-modal retrieval performance using OCR and ASR text on the Google I/O corpus.

	mAP@N		
	N=5	N=10	N=209
VSM early fusion	0.807	0.780	0.723
VSM late fusion	**0.829**	**0.811**	**0.747**
PLSA early fusion	0.807	0.785	0.708
PLSA late fusion	0.751	0.731	0.647
MLM (ours)	0.822	0.805	0.734

5.5 Impact of noise on MLM performance

The results above show that the use of noisy slide and spoken text can diminish the effectiveness of MLMs for lecture video retrieval. To study the degradation of retrieval performance due to noise in each data modality, we performed experiments using VSM on unimodal data (recall that we have 4 types of unimodal data: PPT, CC, OCR, ASR). The full unimodal VSM results are omitted here for brevity, but retrieval using OCR text performs on par with retrieval using PPT text (mAP@5 for OCR retrieval is 0.01 lower than PPT), whereas retrieval using ASR text is significantly worse than using CC text (mAP@5 for ASR retrieval is 0.206 lower than CC). We thus hypothesize that poor quality ASR degrades retrieval performance of MLMs trained using OCR-ASR word pairs.

To assess this hypothesis, we divide the 275 queries into two sets, according to whether the query's retrieval Average Precision at the top-5 candidates (AP@5, using noisy text data $T_{sl} = T_{OCR}$ and $T_{sp} = T_{ASR}$) for MLM outperforms that of the VSM late fusion baseline. Set 1 consists of those queries for which the MLM outperforms VSM late fusion, while set 2 contains queries for which MLM underperforms VSM late fusion. We next examine unimodal VSM retrieval performance in terms of mAP@5 on the two query sets, using CC and using ASR. On set 1 queries, the mAP@5 for VSM using CC is 0.149 higher than VSM using ASR. However, on set 2 queries, the mAP@5 difference between VSM using CC and using ASR is 0.227. In contrast, the analogous mAP@5 difference for PPT compared to OCR is 0.044 and -0.005 for set 1 and set 2, respectively. We thus observe a far greater performance gap between the sets using spoken text rather than slide text for retrieval. Additionally, ASR text quality is relatively low on the set 2 queries for which MLM retrieval performance is worse. Therefore, when the quality of automatically extracted ASR text is better, we anticipate the MLM will add greater value for retrieval.

6. CONCLUSIONS

We have proposed Multi-modal Language Models and a probabilistic ranking function for multi-modal video retrieval. We introduce a new dataset, the Google I/O dataset, that contains multi-modal lecture videos and text queries with ground truth relevance judgements. When using error-free PPT and CC transcripts for multi-modal retrieval, MLM significantly outperforms several baseline schemes using well established methods of VSM and PLSA. When only the automatically extracted OCR and ASR noisy text are available, our model shows similar performance to the best performing benchmark method, where the degradation of our model reflects the noise in the ASR data.

7. REFERENCES

[1] J. Adcock, M. Cooper, L. Denoue, H. Pirsiavash, and L. A. Rowe. Talkminer: A lecture webcast search engine. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 241–250, 2010.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *The Journal of Machine Learning Research*, 3:1107–1135, March 2003.

[3] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, SIGIR '03, pages 127–134, 2003.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, March 2003.

[5] B. Chen. Word topic models for spoken document retrieval and transcription. *ACM Transactions on Asian Language Information Processing*, 8(1):2:1–2:27, March 2009.

[6] Q. Fan, K. Barnard, A. Amir, and A. Efrat. Robust spatiotemporal matching of electronic slides to presentation videos. *IEEE Transactions on Image Processing*, 20:2315–2328, 2011.

[7] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1/2):177–196, January 2001.

[8] T. Kawahara, Y. Nemoto, and Y. Akita. Automatic lecture transcription by exploiting presentation slide information for language model adaptation. In *IEEE ICASSP*, 2008.

[9] R. Lienhart, S. Romberg, and E. Hörster. Multilayer plsa for multimodal image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, pages 9:1–9:8, 2009.

[10] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

[11] D. Putthividhya, H.T. Attias, and S.S. Nagarajan. Topic-regression multi-modal latent dirichlet allocation for image and video annotation. In *IEEE Computer Vision and Pattern Recognition*, 2010.

[12] N. Rasiwasia, J. C. Pereira, E. Coviello, G. Doyle, G. R. G. Lanckriet, R. Levy, and N. Vasconcelos. A new approach to cross-modal multimedia retrieval. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 251–260. ACM, 2010.

[13] A. Vinciarelli and J. Odobez. Application of information retrieval technologies to presentation slides. *IEEE Transactions on Multimedia*, 8(5):981–995, 2006.