Polylingual Multimodal Learning

Aditya Mogadala

Institute AIFB, Karlsruhe Institute of Technology, Germany
{aditya.mogadala}@kit.edu

Abstract. The growth of multimedia content on the web raise diverse challenges. Over the last decades various approaches are designed to support search, recommendations, analytics and advertising majorly based on the textual content. Now, due to the overwhelming availability of media (images and videos) require advancements in the existing technologies to leverage multimedia information. Recent progress made in machine learning to foster continuous representations of the text and effectual object detection in videos and images provide new opportunities. In this aspect, my research work aims to leverage data generated from multimedia to support various applications by finding cross-modal semantic similarity. In particular, it aims to compare semantically similar content generated across different channels by jointly modeling two different modalities. Modeling one or more modalities together can be helpful to generate missing modalities and retrieve cross-modal content. My research also extends textual information in multiple languages to support the growth of polylingual content on the web.

1 Introduction & Motivation

The web contains different modalities. A modality represents information from multimedia items like image, video and text. Most of the time, one or more modalities are represented together to provide a multi-view experience.

Solving challenges posed by multimedia content provide different applications. Most of the earlier research considered multimedia items separately and designed approaches for the tasks like image and video retrieval [1, 2], image annotation [3], image segmentation [4], face detection [5], person identification and tracking [6] etc. In the recent years, multimedia and computer vision communities published considerable research in bridging the gap between modalities to facilitate cross-modal applications [7]. Their research aims to address the problems of automatic image tagging with class labels [3], usage of image queries for text retrieval [8], automatic generation of image and video descriptions [9–11]. Some of these approaches leverage more than one modality by jointly modeling them together. I divide these multimodal learning approaches into three different categories. The first set of approaches generate annotations or tags for images or videos. The second set of approaches provide descriptions (captions) to images and videos with larger phrases or sentences. While, the third set of approaches identifies images and text belonging to same semantic category with cross-modal
retrieval [12, 13]. But, most of the work pertaining to text along with images or videos is limited to English.

Natural language processing (NLP) and information retrieval (IR) communities have been working on different multilingual [14] and cross-lingual applications [15] over the past decades. While they only concentrate on text and diminish the importance of other modalities present in multimodal documents.

Given these limitations of earlier research conducted in natural language processing and computer vision communities supporting different applications. My research work focus on leveraging multilingual and cross-lingual approaches by using more than one modality. Also, it aims to extend multimodal learning to multiple languages. In particular, it aims to find similarities between text present in multiple languages and images or videos by jointly modeling them. Some of the challenges and issues in this research are itemized below.

– How to jointly model different multimedia items like images and videos along with text.
– How to handle variations of text present in different forms like keywords, phrases, noisy data (e.g. tweets, comments) and paragraphs.
– How to design language independent multimodal learning approaches.
– How to extract features which are scalable to multiple data-sets and are not domain dependent.

The remainder of this proposal is organized into the following sections. Background and related work are mentioned in section 2. The section 3 presents limitations of state of the art and provide novelty of my research. Approach along with datasets and evaluation metrics are mentioned in section 4. The section 5 details the work done and the work in progress. Conclusion is discussed in section 6.

2 Background & Related Work

My research work identifies its background from other learning approaches like multi-view and multi-task learning, structured prediction etc., below I list four closely related categories based on tasks.

2.1 Cross-lingual Semantic Similarity (Text only)

Matching semantically similar documents or words belonging to two different languages had been important task to support applications like cross-language information retrieval, machine translation and cross-language classification. In general, cross-language similar documents are retrieved using query translation [16], CL-LSI [17] or CL-KCCA [18]. Other approaches found relatedness with lexicons and semantic knowledge bases [19]. Fuzzy and rough set approaches [20] are also proposed to find cross-language semantic similarity. But lately, more interest is developed to learn latent spaces of two different languages in-order to predict word translations. Latent topics considered as concepts was used for
semantic word similarity tasks in monolingual [21] and multilingual [22, 23] settings. Cross-language latent topics concentrate on extracting concepts based on global co-occurrence of words in a training corpus with and without considering contextual information.

2.2 Cross-modal Semantic Category Similarity

Bridging different modalities is sometimes seen as a task of matching modalities to same semantic class or categories. There are several approaches proposed using joint dimensionality reduction approaches [12, 13] or formulating an optimization problem [24] where correlation between modalities is found by separating the classes in their respective feature spaces. Other approaches aim in learning heterogeneous features implicitly without any external representation. Joint representation of multiple media used by Zhai et al., [25] focus on learning which incorporates sparse and graph regularization.

2.3 Cross-modal Description with Text

Lately, considerable interest has been shown to automatically generate descriptions or captions for images and videos. These descriptions can be either belong to annotations or variable length phrases. Srivastava et al. [26] had learned a joint model of images and their textual annotations with deep boltzmann machines to generate one from other. Vincente et al., [27] automatically created a dataset containing 1 million images with associated visually relevant descriptions from Flickr\(^1\) by performing queries and filtering. Other approaches [9, 10] generated image descriptions with a constraint on text neural language model and images or used fixed templates. Some approaches extended the idea from still images to videos with deep recurrent neural networks [11] and unified frameworks [28]. Other approaches [29] mapped complex textual queries to retrieve videos by understanding visual concepts and semantic graph generated from sentential descriptions. Anna et al. [30] pursued a different use-case by creating a descriptive video service to help visually impaired users to follow a movie.

2.4 Cross-modal Description with Knowledge Bases

Multimedia search understands and organize images and videos in a better manner. Combining visual and semantic resources to analyze the image annotations can provide some world knowledge to visual information. Imagenet [31] is a hierarchical image database structured on lexical database Wordnet. Other approaches combined image tags from Flickr with a commonsense reasoning engine ConceptNet [32] to understand the general tagging behavior. Qi et al., [33] propagates semantic knowledge from text corpus to provide annotations to web images. Visual knowledge bases like NEIL [34] help to continuously build common sense relationships and labels instances of the given visual categories from Internet.

\(^1\) https://www.flickr.com/
Language and context information leveraged from structured knowledge bases was used by Alexander et al., [35] to suggest basic-level concept names depending on the context in which visual object occurs.

3 Limitations with State of the art

The main motivation of my research is to find semantic similarity between content generated across modalities. Though some of the approaches mentioned in the section 2 achieve this in various ways, there are still some limitations which need to be addressed. In this aspect, my research extends or improves multimodal learning for existing and newly created tasks. Below, I divide these tasks originating from two different perspectives.

3.1 Vision for Language

Most of the research conducted earlier is used to compare content across languages for various tasks like cross-language retrieval, cross-language semantic similarity or cross-language classification and mostly focused on only textual information. But due to growth of multimodal content, different language documents are frequently accompanied by images or videos. In my research, I aim to bridge languages with visual clues present in these multimodal documents. This approach creates less dependency on language specific tools and can be scalable to languages which lack resources or tools. Siberalla et al. [36] made a similar attempt to identify semantically similar words with the help of visual information. Though it was only limited to English vocabulary.

3.2 Language for Vision

As mentioned in section 2, there are many ways to link text with images or videos. Text can be generated as annotations, descriptions or labels of semantic categories. Approaches that annotate objects in an image or videos are either limited by domain or leverage information from visual knowledge bases. Most of the keywords which are used to annotate objects are present in English and are depending on word translations to extend them to other languages. Similarly, approaches that are developed for automatic image and video captioning with variable length descriptions are also limited to English. Possible explanation for this limitation is due to approaches that use predefined templates or depend on generation of grammar. Similar issue is been observed with approaches which considered cross-modal retrieval based on same semantic category labels of images and text. Most of the knowledge bases (KB) like DBpedia etc are cross-lingual, though approaches which leverage KB still work with annotations in English.

Observing the possibilities and to support the multilingual web, my research aim to extend multimodal learning beyond English language. This can trigger various applications that can improve multimedia search or cross-modal recommendations.
4 Approach

Variation in discrete and continuous representations of information like text and images or videos respectively require composite approaches to find semantic similarity. In this aspect, my research aims to learn correlations between media and textual content by learning joint space representation. As discussed earlier in the section 1, learning correlations between two different modalities can be divided based on tasks that support cross-modal retrieval and generation. Below, I formulate the problem for each of these tasks and explore possible approaches.

4.1 Polylingual Cross-Modal Semantic Category Retrieval

Multimodal documents are found on web in the form of pair-wise modalities. Sometimes, there can be multiple instances of modalities present in the documents. To reduce the complexity, I assume a multimodal document \( D_i = (Text, Media) \) to contain a single media item (image or video) embedded with a textual description. A collection \( C_j = \{D_1, D_2...D_i...D_n\} \) of these documents in different languages \( L = \{L_{C_1}, L_{C_2}...L_{C_j}...L_{C_m}\} \) are spread across web. Formally, my research question is to find a cross-modal semantically similar document across language collections \( L_{C_o} \) using unsupervised similarity measures on low-dimension correlation space representation. To achieve it, I propose following approach which learns correlated space between modalities in different languages for cross-modal retrieval.

**Correlated Centroid Space Unsupervised Retrieval (C\(^2\)SUR) [37]** In this approach, I find correlated low-dimension space of each text and media (Image) with kernel canonical correlation analysis (kCCA) modified with k-means clustering.

Let \( m_T = \{m_{T_1}...m_{T_k}\} \) and \( m_I = \{m_{I_1}...m_{I_k}\} \) denote the initial \( k \) centroids for the correlated text and image space respectively obtained with kCCA. Iterating over the samples of the training data, I perform assignment and update steps to obtain the final \( k \) centroids. The assignment step assigns each observed sample to its closest mean, while the update step calculates the new means that will be a centroid.

Correlated low-dimension space of text and image samples of the training data is given by \( CS_{TR_T} \) and \( CS_{TR_I} \) respectively. Choice of \( k \) is dependent on number of classes in the training data, while \( p \) represents the total training samples. \( S^{(t)}_T \) and \( S^{(t)}_I \) denote new samples of text and image modalities assigned to its closest mean. Algorithm 1 lists the procedure. Now the modified feature space is used for cross-modal retrieval with distance metrics like cosine etc.

**Experimental Data and Evaluation** To evaluate the approach in a polylingual scenario, I use the wiki dataset\(^2\) containing 2866 English texts and images

\(^2\) http://www.svcl.ucsd.edu/projects/crossmodal/
Algorithm 1 Correlated Centroid Space

Require: \( CS_{Tr} = \{x_1, \ldots, x_n\}, CS_{I} = \{x_1, \ldots, x_n\} \)

Ensure: \( p > 0 \) (Output: Final K-Centroids)

Assignment Step:
\[
S_{t,j}^{(t)} = x_j : \|x_j - m_{t,i}\| \leq \|x_j - m_{t,i}^*\| \forall i = 1 \ldots k
\]

Update Step:
\[
m_{t+1}^{(t+1)} = \frac{\sum_{j \in S_{t,j}^{(t)}} x_j}{|S_{t,j}^{(t)}|}, m_{t+1}^{(t+1)} = \frac{\sum_{j \in S_{t,j}^{(t)}} x_j}{|S_{t,j}^{(t)}|}
\]

created using Wikipedia’s featured articles is expanded to German and Spanish \(^3\) while keeping the original images for every language. Thus, the expanded dataset consists of text and image pairs in three different languages. Evaluation for cross-modal retrieval will be done with mean average precision (MAP) \([12, 13]\) and mean reciprocal rank (MRR) scores. Experiments are 10 fold cross-validated to reduce selection bias.

4.2 Polylingual Cross-Modal Description

Description of a given video or an image depends on the generation of text. To achieve it several approaches are designed with dependency on predefined textual templates and image annotations. Objects identified in an image is used to fill predefined templates to generate descriptions. Though these kind of approaches imposes limits, may still have the advantage that results are more likely to be syntactically correct. Also, it limits its generalization to many languages.

Few other approaches overcome these limitations by generating grammar, though they pose similar issues as earlier in a polylingual scenario. In this aspect, I find my research question of language independent multimodal learning approaches. Recently, for image descriptions two different approaches proposed using multimodal log-bilinear model (MLBL) \([9]\) and recurrent neural network (mRNN) \([10]\) which does not use language specific information and can show impressive results if applied to different languages. mRNN is feed with image and textual features generated with region convolution neural networks (RCCN) and continuous word representations respectively. Now to generate description, an idea similar to Long Short-Term Memory (LSTM) \([38]\) is used for a sequence model. LSTM is helpful to decode the vector into a natural language string. Similar approaches were extended to videos \([11]\).

In this aspect, my research aims to learn multilingual space for textual features along with image or videos to support polylingual multimodal learning. Multilingual space will help to produce descriptions for the languages that are represented in the space. Considerable research has been done in learning multilingual space in NLP community to support cross-language applications. Re-

\(^3\) \url{http://people.aifb.kit.edu/amo/data/Text-Ger-Spa.zip}
recently, multilingual models are developed for compositional distributional semantics [39].

My research aims to use multilingual space of continuous word representations combined with CNN as an input to modified mRNN to support polylingual cross-modal description. Generated descriptions are further verified for its correctness and readability with Knowledge bases (KB) concepts, entities and common sense facts.

Experimental Data and Evaluation There are several data sets available for English text and images or videos. Flickr8K, Flickr30K and COCO datasets contain images and descriptions, while ImageCLEF 5 in subtask-2 provide images with annotations to generate descriptions. Though there are few datasets for textual descriptions in multiple languages, IAPR TC-12 benchmark 6 provide each image with an associated text caption in three different languages (English, German and Spanish). For evaluation, The BLEU scores [40] are used to evaluate generated sentence by measuring the fraction of n-grams that appear in the ground truth.

5 Results and Work in Progress (WIP)

Below, I present the initial results obtained for polylingual cross-modal semantic category retrieval and discuss the work in progress (WIP) for polylingual cross-modal description.

5.1 Polylingual Cross-modal Semantic Category Retrieval (Results)

Table 1 shows the initial results obtained on text and image queries for English, German and Spanish on the Wiki dataset. I used polylingual topic models (PTM) [22] to extract textual features as a distribution of topics in multiple languages, while each image is represented as 128-dimension SIFT descriptor histogram. MAP scores for \( C^2 \text{SUR} \) for German and Spanish with different topic variations. For example, \( C^2 \text{SUR}-10 \) represents 10-topics. Please note, that the related work can only be applied to English text.

5.2 Polylingual Cross-Modal Description (WIP)

For polylingual cross-modal descriptions, significant contribution comes from building multilingual space of languages. Currently, I am working on building multilingual space of word embeddings for one or more languages using class aligned document corpora and sentence aligned parallel corpora. This is achieved using noise contrastive large-margin updates which ensure non-aligned parallel sentences and non-aligned classes documents observe a certain margin from each other.

---

4 http://mscoco.cloudapp.net/
5 http://www.imageclef.org/2015/annotation
6 http://www.imageclef.org/photodata
<table>
<thead>
<tr>
<th>(Language) System</th>
<th>Image Query</th>
<th>Text Query</th>
<th>Average (MAP)</th>
</tr>
</thead>
<tbody>
<tr>
<td>English</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SM [12]</td>
<td>0.225</td>
<td>0.223</td>
<td>0.224</td>
</tr>
<tr>
<td>Mean-CCA [13]</td>
<td>0.246 ± 0.005</td>
<td>0.194 ± 0.005</td>
<td>0.220 ± 0.005</td>
</tr>
<tr>
<td>SCDL [41]</td>
<td>0.252</td>
<td>0.198</td>
<td>0.225</td>
</tr>
<tr>
<td>SliM² [42]</td>
<td>0.255</td>
<td>0.202</td>
<td>0.229</td>
</tr>
<tr>
<td>GMLDA [24]</td>
<td>0.272</td>
<td>0.232</td>
<td>0.252</td>
</tr>
<tr>
<td>C² SUR-10</td>
<td>0.273 ± 0.002</td>
<td>0.262 ± 0.003</td>
<td>0.268 ± 0.003</td>
</tr>
<tr>
<td>C² SUR-100</td>
<td>0.236 ± 0.004</td>
<td>0.250 ± 0.008</td>
<td>0.243 ± 0.006</td>
</tr>
<tr>
<td>C² SUR-200</td>
<td>0.278 ± 0.002</td>
<td>0.253 ± 0.002</td>
<td>0.266 ± 0.002</td>
</tr>
<tr>
<td>German</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C² SUR-10</td>
<td>0.284 ± 0.002</td>
<td>0.263 ± 0.003</td>
<td>0.276 ± 0.003</td>
</tr>
<tr>
<td>C² SUR-100</td>
<td>0.236 ± 0.004</td>
<td>0.250 ± 0.008</td>
<td>0.243 ± 0.006</td>
</tr>
<tr>
<td>C² SUR-200</td>
<td>0.278 ± 0.002</td>
<td>0.253 ± 0.002</td>
<td>0.266 ± 0.002</td>
</tr>
<tr>
<td>Spanish</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C² SUR-10</td>
<td>0.250 ± 0.001</td>
<td>0.268 ± 0.002</td>
<td>0.259 ± 0.002</td>
</tr>
<tr>
<td>C² SUR-100</td>
<td>0.258 ± 0.008</td>
<td>0.243 ± 0.004</td>
<td>0.251 ± 0.006</td>
</tr>
<tr>
<td>C² SUR-200</td>
<td>0.267 ± 0.003</td>
<td>0.244 ± 0.002</td>
<td>0.256 ± 0.003</td>
</tr>
</tbody>
</table>

Table 1. Text and Image Query Comparison (Wiki)

6 Conclusion

In this proposal, I presented my research on jointly learning heterogeneous features generated from two different modalities mainly polylingual text and image or videos. I aim to do this by segregating the approach to two different tasks. In the first task, textual information and media (image or video) is mapped to the same category with cross-modal retrieval. While in the second task, more sophisticated approaches are used to generate one modality from another. Inherently, these tasks provide better support to search, recommendations, analytics and advertising based multimedia applications.

7 Acknowledgements

I would like to thank Achim Rettinger (rettinger@kit.edu) and Rudi Studer (rudi.studer@kit.edu) for their guidance. This research has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 611346.

References


23. Vuli, I., and Moens, M. F. Probabilistic models of cross-lingual semantic similarity in context based on latent cross-lingual concepts induced from comparable data. In EMNLP. (2014)