Tagging and Browsing Videos According to the Preferences of Differing Affinity Groups

Yicun Liu

Department of Computer Science

Columbia University

Abstract

As the digital archive of our world grows exponentially, international events are widely covered, and most of them are available as news videos from diverse cultural perspectives. Although understanding each culture is a meaningful task, searching for another culture's response upon news events can be quite challenging, as there exist language boundaries and user disparities between one culture and another.

In this work, we aim to discover the cultural differences by identifying the culture-specific tags towards news videos: given a short video clip about an international news event, we retrieve the most relevant feedback from one specific cultural affinity group. To achieve this goal, we start by finding the determining characteristics of the news events that got widely covered in different cultures. We crawled thousands of news videos from major video sharing sites across the globe, and gather a new dataset covering several iconic news events with culturally differing feedback. Furthermore, we propose a pre-filtering step based on the video transcript to reduce the portion of non-relevant contents (such as advertisements) in the crawled data.

Report of COMS 6901 Projects in Computer Science, work supervised by Prof. John R. Kender.

To bridge the modality gap between language and vision, we extract visual and multilingual features from state-of-the-art embedding models and mapped them onto a joint latent embedding space through canonical correlation analysis (CCA), which produces a potential correlation matrix between image candidates and text candidates. To further understand the cultural differences, we conduct non-negative factorization on the correlation matrix to extract visual and language concepts for demonstration and evaluation. We test our pipeline on several popular international news events, and experiments confirm the effectiveness of our framework.

1 Introduction

Web-archived news videos are growing substantially on online multimedia sharing platforms, and a considerable amount of them are focusing on the coverage of international news events. As the largest content sharing platform, YouTube nowadays has over one million hours of uploaded news video per week, sharing the media coverage of domestic and international events in more than one hundred languages [6, 66]. Among them, eye-catching international events such as health epidemics, presidential elections, terrorist and anti-terrorist activities, financial incidents, natural disasters, and sports games draw tremendous attention from the audience across the globe. Since the events covered in different countries have significant overlap, it can be a promising source to explore feedback from different audience groups, revealing their unique cultural preferences, living habits, and other usage patterns. Nevertheless, due to the flood of the massive online media contents available, the video selection step and interpretation step can be heavy work. Moreover, the task of exploiting the cultural difference usually requires bilingualism and cultural contexts (e.g., memes, histories, etc.), which brings unbearable costs to conduct manually, and might introduce new bias from the labeler.

Researchers to date have long realized this dilemma, as there already exists a considerable amount of work focusing on digesting the multimedia contents and providing

users with concise summaries. Early research takes frame segments and extracts visual features from frames to conduct scene classification [27]. Some later research focuses on automatically synthesizing text summary for short videos [57]. Since speech also conveys essential information in news videos, solely relying on visual features in news video understanding is suboptimal due to lack of language concepts. To address this problem, multimodal models are introduced in [46], which utilizes both visual and text features to generate short video descriptions. In parallel to the generation frameworks, there are also previous works on retrieval tasks such as tagging. Automatic tagging procedures are proposed in [50], which exploits content redundancy to generate the most suitable textual tags for potential video recommendation. Tag recommendation is later enhanced by various ranking algorithms in [39], which produces relevant tags in a user-preferred ranking order rather than random order. However, most of the generation or retrieval models are designed in a low-level semantic level, where only little attention has been drawn on the differences in tags across user affinity groups.

In fact, international news videos can be considered as a valuable resource for mining such culturally-based differences. Since many cultural affinity groups may watch the same international news event, it is possible to compare the differences in the video languages (such as real-time editorial comments) towards similar news imagery, and discover the cultural tendencies. Although few research to date has made similar attempts before, comparing language segments in news reports is not a fundamentally new task. In the field of information retrieval, researchers are also interested in summarizing similarities and differences among related documents [42] towards the same topic. In particular, Nakasaki [44] analyzed the text portion of multimedia pages, and conduct a cross-cultural comparison on commonly used language expressions on facts and opinions. Lin [38] also observed significant cultural differences towards one common concept or term on social media. However, there exists only limited work using temporally relevant visual and language features to exploit cultural differences, and no comprehensive study of this topic has appeared. All these works confirm that the cultural difference indeed exists and could be potentially interesting to study.

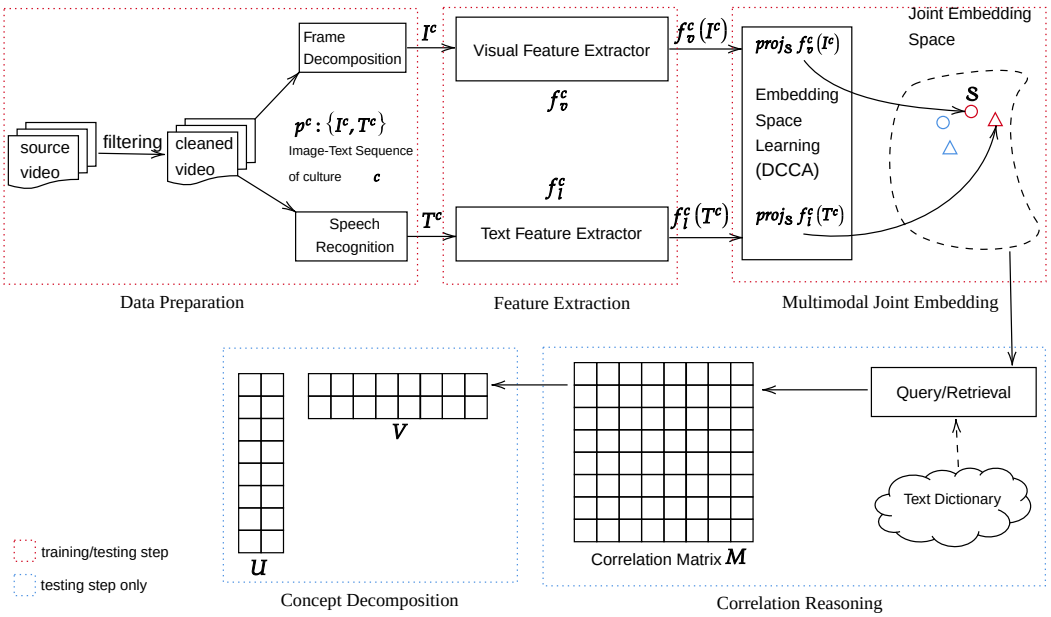


Figure. 1. Overview of the pipeline of the proposed framework. The pipeline includes five steps: **1. Data Preparation**, which includes crawling, filtering, and converting video to image-text sequence pair p^c. **2. Feature Extraction**, which includes using pre-trained expert networks to extract visual and language features. **3. Multimodal Joint Embedding**, which aims at learning a shared space $\mathbb{S}$ where the projection of different features can be maximumly correlated. **4. Correlation Reasoning**, which queries each of the tags in the text dictionary to generates a $N \times K$ correlation matrix $\mathcal{M}$, where N is the number of images, K is the top-K limit of the most related text tags for each image. **5. Concept Decomposition**: we perform non-negative matrix decomposition on the correlation matrix to generate the visual concept vector u and language concept vector v. Note that only steps 1-3 are required during the training step of the framework, and steps 4-5 are only for testing and validation.

Prior to our work, Tsai [55] made the first attempt to detect cultural specific tags from news videos, which uses a similar framework to regular image-text retrieval tasks. Under Tsai's setting, each time a few keyframes are randomly selected from a news video, and the correlation between the random frames and tags from a dictionary is computed. To find the most appropriate textual tags in the background culture, ranking is performed based on the correlation between the keyframes and candidate tags. While this design is quite straightforward, there also exist certain limitations:

1. In Tsai's work [55], only descriptions are used instead of transcripts. Descriptions are usually concise and objective. In fact, most descriptions for videos on YouTube are less than 100 words and usually contains only heading of the news. Even worse, some news videos serve as summary of daily news, which usually contains topics from several news. Similar to the cocktail party effect[1], it is hard to separate the text belong to a specific news with texts from other news, without the guidance of the corresponding news imagery. Additionally, the word choices in video descriptions are more objective, compared with the subjective words in the editorial and criticism made by the reporters and the news hosts.

2. In Tsai's work [55], the embedding process takes no account of the chronological order in the video, which could serve as a critical hint to link the text (speech) and imagery together. Intuitively, the shared imagery only correspond to the speech near its occurrence. Without taking temporal information into consideration, the intrinsic mapping between image and text could be entirely out of order, leading to many noisy image-to-text pairs for the following embedding process.

3. In Tsai's work [55], language embedding is only conducted with models trained on English corpus. Machine translation is used to convert descriptions from non-English to English could introduce deficiency for comprehension. As shown in Figure 3, confidence of machine translation could be significantly lower when translating from non-spacing languages to spacing languages. As name entities usually appears in the news and it is extremely hard to translate, translating those name entities into English could be a sub-optimal choice.

To conquer the limitations as mentioned above, we propose to conduct cultural analysis purely on the imagery and speech available in news videos. To take the chronological

[1] The cocktail party effect is the phenomenon of the brain's ability to focus one's auditory attention (an effect of selective attention in the brain) on a particular stimulus while filtering out a range of other stimuli, as when a partygoer can focus on a single conversation in a noisy room [2].

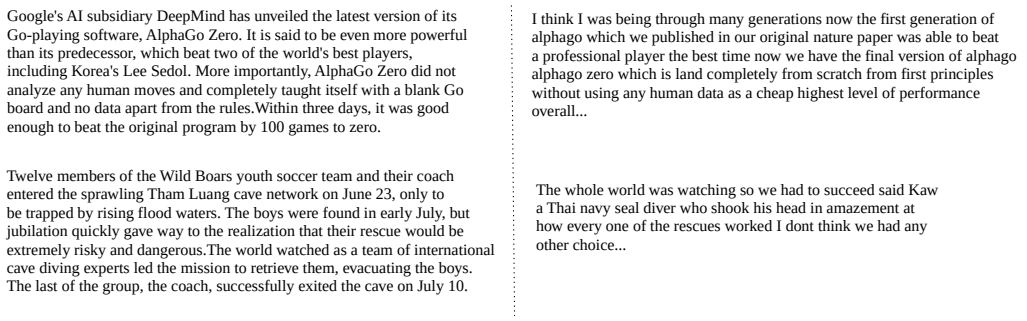


Figure. 2. Clips of descriptions and transcripts of the crawled news videos. On the left are clips of video descriptions on YouTube, and on the right are clips of video transcripts generated by speech recognition package. Note that, in most cases, the transcribed texts do not have punctuation. We show two pairs of clips in this example, where the first one is about 'AlphaGo vs. Human' news event, and the second one is about 'Thailand Cave Rescue' news event. It is worth noticing that the content of the description usually gives a brief idea of what happened, which is more subjective. On the other hand, the content of the transcripts can be originally interviews, debates and other discussions, which involves a more objective view of the international event.

order into consideration, we propose to utilize the real-time language information in the news video, rather than short descriptions. This will includes editorial comments, critics, dialogue made by reporters, guests and news hosts, which entails more subjective perspective towards the news video. We use speech recognition packages to convert speech into text transcripts and encoded them by native language embedding models. To reflect the chronological order in the imagery, we choose the key-frames which are centered by the nearby texts according to the timestamps in the transcript. There are several benefits in doing that:

1. It is easy to find that speech sequences in the news video is significantly lengthier than video descriptions, and they usually include subjective editorial and criticism, which is more revealing for our task of cross-cultural analysis. For a news video of approximately 3 minutes in length, there are typically 300-500 words, compared with less than 100 words in the video description. An example of these differences is shown in Figure 2.

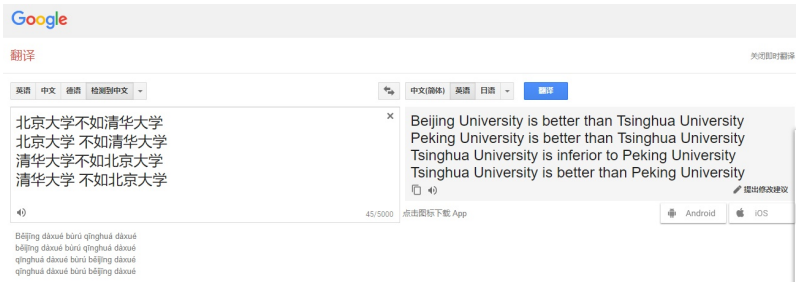


Figure. 3. One widely-discussed false translation example made by previous version of Google's Online Translation: the machine translation between non-spacing languages (such as Chinese) and spacing languages (such as English) is prone to such adversarial examples. By adding space to the Chinese sentences which doesn't change the actual semantic meaning, machine translation can be fooled and generate English sentence of the reverse meaning.

2. This approach ensures that speech sequences and image sequences are chronologically aligned, and the temporal association between the text and image can be effectively encoded in the multimodal embedding process.

3. Despite the advances in machine translation, the vast divergence between one language and another can still lead to inevitable cases of less accurate translation, even for human interpretation. In our method, through native word embedding models, we can bypass the language gaps and encode multilingual texts into language feature vectors with language models trained on native corpus.

With features vectors for imagery and text at hand, the next problem is how to bridge the modality gap between feature vectors of vision and language perspective, and how to construct an effective metric to calculate the semantic similarity between imagery and text. Before us, there is merely research in multimodal learning from the intercultural perspective. One popular solution is to learn a mapping to project the language features and visual features into a joint embedding space, where vectors from various modalities can be compared side by side after the projection. Previously, there exists work approaching this problem by learning a linear transformation [64] or using ranking loss [63, 16] to sort the text candidates with similarity score. For each image in the

training step, they use a single-directional ranking loss that applies a margin-based penalty to any incorrect annotations which are ranked higher than the correct one. Later work proposed bi-directional ranking loss to add additional guarantee for the rank of the sentences instead of individual words, where correct sentences should be ranked higher than the incorrect ones. Although the ranking loss approach is proved to perform well, it requires additional negative sampling to matching the positive sampling, which is not suitable for our cross-cultural analysis task. Another alternative is canonical analysis based embedding, which search for linear (CCA) [19] or non-linear (KCCA) [32] projections to maximize the correlation between features from different modalities. Followed by prior work in [55], we propose to use CCA to construct the joint latent space for our cultural analysis task. We assume that we have two pairs of image-text feature vectors from two different cultures, and we aim to find a joint latent space where the projections of the two image-text pairs lead to the maximized correlation.

Assume a joint embedding space can be effectively learnt by CCA and its variants, we can conduct comprehensive query to obtain a correlation matrix between texts and images. Although correlation itself can be measured quantitatively, the correlation value usually cannot provide enough interpretability for further human-in-the-loop analysis. Different from previous work in [55], where only recall of matched tags is recorded for evaluation, we take the framework further and try to explain the dominating facts inside correlation matrix. Inspired by general decomposition-based recommendation system setting [47, 31, 17], we proposed to first rescale the correlation value to zero to one, and then decompose the rescaled correlation matrix as the form of matrix multiplication of two thin matrices, where the two thin matrices should correspond to the dominating factors in image and language dimension, respectively. Among many matrix decomposition techniques, we choose the non-negative matrix decomposition [34] algorithm to extract concept vectors in both vision and language perspective. By analyzing the concept vectors, we hope to visualize the 'focusing point' of one culture towards a specific news video, and compare the difference of 'focusing points' in difference culture, to reveal more cultural differences.

This work can be treated as a systematic extension of prior work by Tsai [55], but differs from prior work in the following ways:

- We study the user preferences of mainstream video sharing websites in the US and China, which helps us to design an automatic pipeline to target and crawl news videos from web archives for different countries. We analyze which kind of news topics are most popular in the two cultural affinity groups, and find certain characteristics of three types of topics that can be potential news topic for our automatic video crawling pipeline.

- We construct both syntactic-based and semantic-based pre-filtering approaches to alleviate the non-relevant and advertisement contents in the news video. The proposed syntactic-based filtering approach is based on term frequency—inverse document frequency (TF-IDF) score of the statistical blacklist and whitelist. The semantic-based filtering approach is based on building Supporting Vector Machine (SVM) binary classifier on top of feature vectors extracted from language embedding neural networks.

- We exploit several potential drawbacks of the previous method, illustrating why the original speech is preferred in the cross-cultural comparison task. As an improvement, we propose to utilize chronological ordering as the linkage of speech and image and use native language embedding models to bridge the multilingual gap and bypass the error in machine translation.

- We investigate more recent language and visual embedding models and the effect of introducing more context in the language embedding. Furthermore, we propose alternatives of CCAs for constructing the latent embedding space, and more efficient framework to encode the temporal information.

- We propose to conduct non-negative matrix factorization (NMF) on the correlation matrix generated from CCA. The decomposed thin matrices serve as dominating concepts from image and language perspective respectively, which provides more interpretability for the evaluation and visualization process.

2 Related Work

2.1 Video Analysis and Image-Text Matching

The blooming of online video contents has long drawn the interest of researchers. In the multimedia field, analyzing and digesting the video contents is one of the core research topics. Early research on video analysis mostly focuses on effectively categorize videos or generating a concise summary for users. Traditional video classification tasks are constructed by three steps: (1) Extract local visual features that describe a region or patch (2) Combine extracted local features into a fixed-size video descriptor (3) Apply classifier to train on the resulting 'bug of patches.' Early work usually used hand-crafted feature extractor and dictionary-based K-Means to quantize the accumulated features over the duration of video [33]. Later on, Karpthy first introduced CNN-based video classification on a large-scale dataset containing over 1 million videos over 487 classes [27]. The method took advantage of local spatiotemporal information in their network design, which achieved a significant boost compared with feature methods. There is further improvement in [68] to handle full-length video. As the unsupervised alternative of video classification, video clustering has also received significant attention, Wang proposed semantic-based clustering of tagged videos [58] based on sparse and incomplete tags, which is common in data from stream provider. There are also efforts in generating video summary in [69, 41], which aimed at automatically selecting key-frames and utilized LSTM [20] to model the variable-range temporal dependency among video frames. There are also other research concentrating on analyzing more fine-grained visual attributes in videos: activity analysis [7], scene understanding [53] and consumer analysis [23].

Nevertheless, analyzing videos solely based on visual features can be less effective. As an example, for news videos covering the same international news, news imagery can be shared by multiple media channels, where the speech and comments play a more crucial role in determining the cultural preference. In some scenarios, it is of great interest to know the relationship between images and their descriptions. The

image-text matching task is later proposed with an emphasis on retrieval of the most relevant text given an image, or finding the most appropriate image given the text descriptions. Completing the matching task usually involves two components [59]: (1) A shared embedding space to represent language and visual features (2) A metric to fuse the features in the embedding space and calculate the similarity score. The proposed methods achieve impressive accuracy in the MS COCO dataset, but it is designed to operate in a single image and text description instead of consecutive frames and speech sequences in videos, which generally have a much longer duration.

2.2 Multimodal Representation Learning

CCA-based Method A popular baseline in multimodal representation learning is the Canonical Correlation Analysis (CCA), which constructs a linear transformation to maximize the correlation between two projected vectors from the two views [19]. As an improvement, a kernelizable, non-linear version of CCA (KCCA) is then introduced in [32], which aims at finding the maximized correlated non-linear projection in the reproducing kernel Hilbert space. Not limited to its simplicity, CCA works surprisingly well for most language and image embedding task [29], capable of competing with many state-of-the-art methods if the multimodal input features are properly generated.

The main shortcoming of CCA and KCCA is its high memory cost and testing speed. Upon testing, it is required to load the entire dataset to compute the covariance matrix, which consumes a considerable amount of time for inference. To alleviate this problem, a deep learning paralleled version of CCA (DCCA) is introduced in [1], which does not require an inner product and does not require to reference the training data in the testing step. As a parametric method, its training speech scales with the dataset and inference speech is always constant. This property provides DCCA a dominating advantage when handling large-scale datasets. Additionally, the proposed DCCA framework in [1] includes a non-saturating sigmoid function based on the cube root.

Apart from DCCA, there also exist other deep learning variations. Wang proposed deep variational canonical correlation analysis (DVCCA) [61], which adds a lower bound of

the data likelihood by parameterizing the posterior probability of the latent variables. Deep generalized canonical correlation analysis (DGCCA) is proposed by Benton [4], which is a deep version of GCCA [13] and targets at learning non-linear transformation of an arbitrary number of views. Unlike standard CCA, the number of views in the generalized version is not fixed, thus suitable for multimodal learning of many views.

Ranking-based Method The ranking loss is a widely used technique for optimizing many multimodal embedding models. Early research work like WSABIE [64] and DeVISE [16] applied single-directional ranking loss in the training step of the linear transformation for visual and language features. Both framework also introduces a margin-based penalty to any incorrect annotations when get ranked higher than correct ones when describing an image. There are also work associating the bi-directional ranking loss, which adds the missing link in the opposite direction [26, 52]. Such design enforces a stronger guarantee to the ranking order: for any annotation, the corresponding image should get ranked higher than those unrelated images.

Classification-based Method Learning the similarity between multimodal features can also be formulated as a typical classification problem. For example, given a visual feature x and a language feature y, the core idea is to answer whether or not x and y matches each other [22]. Similar to many classification tasks, there is also a soft assignment of the matching decision [22], which includes a softmax function to predict whether the input image and question match with each other. A two-branch network is later introduced using classification loss to match visual and lingual features for zero-shot learning [48]. To train a similarity measurement network (as a branch of the two-branch network), Wang [59] proposed to use the non-exclusive logistic regression loss to replace in the ranking loss, treating each phrase-regio pair as an independent binary classification problem.

2.3 Multilingual Query

Despite the exponential growth rate of the online social media content, the query method of the large volume of multimedia data is still unequally developed. Research shows that people tend to annotate multimedia sources by their native languages, and it is difficult for ordinary people to conduct a cross-lingual query in the multimedia effectively. Not only does the differences between different languages forms such boundary, but also the divergent user habits and preferences for increases the cultural gap. For example, when searching for news videos covering an international event, US audiences tend to find such videos on a media channel (e.g., CNN) on YouTube. However, in China, people tend to go to a localized video sharing website (e.g., Bilibili or iQiYi) in search of such news videos. In real-world cases, most people are unfamiliar about such user habits in another country or culture, thus leads to low efficacy to conduct a multilingual query.

Previously, there also existed some research work in the multilingual query. Popescu [45] first proposes a multilingual dataset on FLICKR image (MLFLICKR Dataset), which builds a cross-lingual query platform on FLICKR. First, they translate the query into different languages and further verified the return results by determining whether or not they are visually similar. Bergsma later observed that users tend to tag their images when posting online [5] naturally. The tags which are used initially provide the perfect link between the language part and the visual part. This discovery enables them to generate tag translations by finding tag-image pairs that share a high level of similarity in the corresponding visual part. Bao [3] proposed Omnipedia, which is a unified framework to retrieve the Wikipedia insights from another language. The contents retrieved includes text, images, hyperlinks, and videos in Wikipedia in 25 languages. Clough [12] indicated that for cross-language image retrieval task, the actual language used in textual tags should not affect the overall accuracy of the retrieval. It was also suggested in [12] that users from different cultures tended to spread their attention differently when viewing the image associated with texts of their native language. Similar research conducted in [15] revealed more cultural differences in the image tagging task:

US people tend to assign the first tag to the primary object to the image. In contrast, Chinese people are more likely to ascribe the first tag describing the overall relationship between objects and the atmosphere of the image.

2.4 News Event Analysis

For news event tracking and analysis, there exist a bunch of work relying on multimodal features. Early work in [65] proposed a constraint-driven co-clustering algorithm (CCC), which utilized the near-duplicate key-frame constraints on top of the text, to mine topic-related stories and the outliers. Li [36] proposed a multimodal topic and-or graph (MT-AOG) to jointly represent textual and visual elements of news stories and their latent topic structures. Their framework is designed to describe the hierarchical composition of news topics by semantic aspects like people, places involved, and model contextual relationships between elements in the hierarchy. Jou [25] took a step further, extracting who, what, when, and where from news data. Wang [62] proposed to integrate multimodal features using the conditional random field (CRF) to segment the news stories. Tsai [56] exploited background statics from YouTube for news event understanding, like the metadata, video category, location, comments, and user preference to establish a PARAFAC co-clustering model and mine the latent factors.

In social event analysis, Cai [9] proposed a spatial-temporal multimodal TwitterLDA model that uses five Twitter cues, including text, image, location, timestamp, hashtag, and modeled topics as location-specific distributions. Chen [11] created a Visual-Emotional LDA (VELDA) model, which associates images and texts both visually and emotionally for image retrieval. Up until now, few work [55] considers cultural differences in multimodal news analysis, and very few of them explicitly model the temporal information in their framework.

2.5 Recommendation System

Recommendation (recommender) system has drawn significant amount of attention from Internet companies such as Youtube, Netflix, Spotify and IMDB [40]. Particularly,

many media companies offer practical recommender systems to their subscribers. For many contents spreaded on Youtube and other social media like Twitter, Facebook, they usually exist as multimodal data, where speech, imagery and other metadata are used as input to the recommendation systems. The way that recommendation system exploits multimodal data is usually through effective embedding models from vision, language, and other perspectives. With the multimodal features at hand, designers of recommendation systems also need to think about how to bridge the multimodal gap to make the final predictions. For mainstream recommendation systems, there are usually three types of the method: content-based, and collaborative filtering.

Content-Based Filtering Content-based methods make recommendations based on the description of the items. Nowadays, it is combined with other methods and use more information about items and users. Content-based recommendation systems consist of three major parts from a high level architectural point of view [49]; first it does the preprocessing on items with a content analyzer, then a profile learner learns about users, and finally, the filtering component finds a set of appropriate recommendations.

Collaborative Filtering Collaborative Filtering method is to predict user-specific rating of items based on a sparse ranking matrix or other user patterns [30]. The sparse ranking matrix records user's historical ratings of many items. In order to extract useful information, collaborative filtering method need to relate the two fundamentally different things: users and items. There are mainly two types of methods for collaborative filtering: latent factor models and neighborhood models. Neighborhood models is designed to analyze the intrinsic relationship inside the user group and item group, so that similarities between different items and users can be computed. Based on the similarities, it can recommend user's with item choices of the most similar other user, or tag the item with the identity of the most similar items. Latent factor models usually involves matrix decomposition such as SVD. It tries to decompose the low-rank user-item rating matrix into two thin matrices, representing the important properties for both users and items. The latent space tries to explain ratings by characterizing both products and users on factors automatically inferred from user feedback. Our

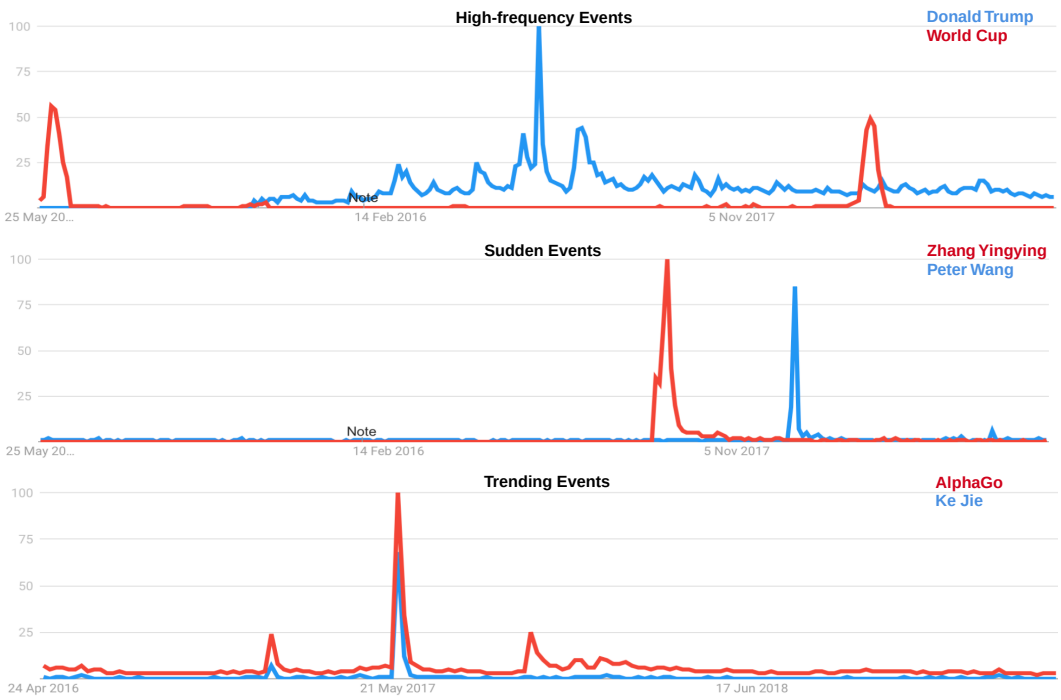


Figure. 4. Examples of the three types of international news events: For each type, we use the name of the person or the topic as the search keyword in google trend, and we plot the magnitude of the attention each event received in the vertical axis. It can be observed that for the **High-frequency Events** type, such as 'Donald Trump' and 'World Cup', the search popularity remains very high for a long time duration (usually more than three months). For the **Sudden Events** type, there is a sudden peak of popularity. In contrast, the rest of the time, the popularity is close to zero, which is ideal for our framework to conduct culture analysis based on the near-duplicate frames. The **Trending Events** type performs similar to the **Sudden Events** type, except that it might have several peaks when public exposure draws increasing attention from the news media.

proposed non-negative matrix factorization [34] is similar to the traditional matrix decomposition techniques used in the latent factor models. The difference is that in our case, the 'rating' (correlation) matrix is generated from canonical correlation analysis, and we use the decomposition techniques to better explain the concepts in both image and language, from a cultural perspective.

3 Which News Event?

Before start gathering the news videos, the first problem is: which news event should be chosen for our cross-cultural analysis task? Although international news gets widely covered across the globe, not all news event is fit for our job. Since we need imagery a the key to retrieve the most relevant cultural-specific textual tag, we expect the new coverage in both cultures shares a large portion of the similar images, such as iconic photos of a specific person.

We use Google Trend data to narrow down our search. We use the top 100 searches in different regions of the world to consider which kind of event is proper for our cultural analysis task. We identify three types of international news events:

1. **High-frequency Events:** The person or topic of the event is frequently mentioned, in which case the person is highly possible to be a public figure. Since the person or the topic frequently appears under the spotlight, there could be thousands of images identified as correlated, where a near-duplicate frame is challenging to find.

2. **Sudden Events:** This type of event usually occurs suddenly and unexpectedly, in which the person or topic in the event is not frequent in the previous news report. An example of these types of events could be a natural disaster or a human-made tragedy. Since the person of the topic is merely reported before the event, there doesn't exist a vast choice of the image in the news video, where a certain portion of the images is similar or shared.

3. **Trending Events:** This type of event is usually about some previously unpopular domains for the main public, such as scientific and medical research, which means that the images before the event merely appear as headlines of the news media. The event is usually an exhibition, a game, or an IPO, which gains significant public attention. Since the person of the topic is merely reported before the event, there doesn't exist a vast choice of the image in the news video, where a certain portion of the images are similar or shared.

Some examples of the three types of events are shown in Figure 4. In the three types of events, the latter two types could be ideal for our analysis, considering the near-duplicate images are easy to find. For the **Trending Events** type, we choose the 'AlphaGo vs. Human' and 'Chinese Moon Rover' events as the target for our analysis. For the **Sudden Events** type, we choose the 'Thailand Cave Rescue', 'Florida High School Shooting', and 'UIUC Zhang Yingying Missing' as the target for our analysis. Note that some of the technical parts of the crawling and refinement will be omitted, and this section only serves as an overview of our data preparation process.

4 Data Preparation

In this section, we will discuss how to prepare the datasets for the cultural analysis task. Before us, there are few publicly available datasets dedicated to cross-cultural news video analysis, and the only relevant dataset gathered by Tsai in [55] has been gradually outdated. Thus, we decide to collect a new dataset based on recent international news events for our further analysis.

4.1 News Videos Collection

User Patterns In the original work carried out in [55], Tsai proposed to collect news videos in the US and China, using YouTube and Baidu search respectively. Currently, in the US, YouTube has over 170 million users, and over 300 million videos are being watched every day. There are plenty of well-known mass media accounts registered on YouTube, keeping a daily routine of uploading news videos covering national and international events. For the US audience in the cultural study, YouTube could be the largest and most convenient source for us to collect news videos.

In China, the top popular video sharing websites include Tencent, Youku, iQiYi, and Bilibili. The operation pattern of Tencent, Youku, iQiYi is closer to traditional content sharing websites, whereas Bilibili focuses more on young-age users, with particular emphasis on social interaction within its user group. There are in total of over 600 million users of the four major video sharing websites, where Tencent has the largest

user community, and Bilibili offers the most original video content. Similar to YouTube, there also exists some media accounts on these mainstream video sharing platforms, which includes mass media accounts covering all domestic and international news event and smaller media accounts with the special focus on a specific group of users sharing the same interest.

Availability and Crawling Difficulty YouTube provides user-friendly APIs for people to download and analyze the videos, metadata, comments, and statistics. As for the video-sharing websites in China, only Bilibili provides developer APIs similar to YouTube, while the rest of the sites can only be crawled by manually crafted downloaders. We find that the design of the other websites updates frequently so that the legacy web crawler used by Tsai in [55] has been outdated and does not work for concurrent website layouts. Fortunately, there exists a bunch of helpful off-the-shelf third-party packages that can be used for spidering videos and metadata from these sites [67]. Also, utilizing a search engine to find Chinese news videos by keywords can also be possible, but the source of the videos can be quite broad, and a unified API or toolkit supporting all the websites is still absent. In our case, we decide first to write query script on the four major websites to obtain the URL of each video, and then we pass the collected URLs to a third-party toolkit [67] to finish the downloading process. As for the video description, we also write a customized script for manual download from the four websites.

Search Query In our initial trial, we directly use 'AlphaGo News' and its corresponding Chinese version as the keyword for search in both English and Chinese websites. The returned results include both news coverage and the live videos of the games where AlphaGo competed with Go players named Lee Sedol and Ke Jie. We then limit the length of the video to be less than 5 minutes, expecting the modified search query can filter out the live game videos. This simple modification works well for the top returned results, where the search engines on YouTube or Chinese websites listed them as the 'most relevant' results. However, we still observed some unmatched video as we enlarge the query range from 500 to 5000 and allow the websites listed those 'less relevant'

Figure. 5. Examples of the real-time transcripts of news videos generated by Google speech recognition service. On the left is from news videos with native English speech, using English mode. On the right is from news videos with native Chinese speech, using Simplified Chinese mode. For each of the paragraph in the transcript, a confidence score will be generated, indicating how confident is the speech recognition service towards the paragraph. For each word in the paragraph, start timestamp, end time stamp are appended as metadata for the word. For English transcript, Google speech recognition can handle the speech well and successfully recognize name entities like `Rothesay`. For Chinese transcript, it performs much worse than English version, leaving several mistakes for characters with similar pronunciation, such as 餐甜 (Can Tian, not a commonly used word) versus 三点 (San Dian, meaning three o'clock), and fail to recognize the name of the Chinese player 柯洁 (Ke Jie, wrongly recognized as 可杰 Ke Jie, although with similar pronunciation but it is not a commonly used word).

videos. To refine the search query, we crawled the video descriptions for the top 100 videos and collected the most frequent words into a whitelist, and we also established a blacklist containing the unique words in the unmatched videos. Each time we conduct a query on a specific website, we use both the whitelist and blacklist to check the description of the returned video and decide whether or not to filter this video. With this strategy, it can automatically filter out most of the unmatched videos. After that, we also conduct a manual check of the video topics to ensure the relevance of the downloaded videos.

4.2 Data Preprocessing

Speech Recognition With source files of the downloaded videos, the next step is to convert video files into time series of images and text. We convert all the downloaded video files into mp4 standard and extract its audio file of wav format. The next step is to obtain the transcripts of the audio files and split them into separate words according to the standard NLP pipeline. Although we find that some videos indeed comes with the original subtitles, the majority of the downloaded videos do not come with audio text. To solve that problem, we propose to use off-the-shelf online speech recognition service to convert the wav files into text files. Mainstream speech recognition service supports multiple languages, delivering very high-quality recognition and robust to different levels of noise. We find that for the news audio, a majority of the audio is recorded in the newsroom or public interview, which leads to impressive accuracy of recognition in both Chinese and English. For text files obtained, we remove all the boilerplate or other linguistically insignificant content, including the name entities.

We compare the speech recognition performance of different online speech recognition service. For news video with English speech, Google's speech recognition performs well can can well recognize the main structure of each sentence. As shown in Figure 5, Google's speech recognition performs relatively well on English speech, leaving spacing between words, which ease the further punctuation. Most of the name entities can be recognized as it supports adding a 'rarely used words' dictionary as prior knowledge to initiate the transcription process. Nevertheless, accuracy in Chinese transcription is significantly less than English transcriptions. For default setting, there is no spacing between words, which is a natural thing in Chinese but potentially makes it hard to read and predict further punctuation. Moreover, even with user-predefined 'rare words' dictionary including the name entities, the speech recognition still struggles to recognize names, only leaving weird words of similar pronunciations.

To address the problem in speech recognition to provide news video dataset with higher quality, we investigate native speech recognition providers such as iFlyTek[2] and Baidu[3]. We find that the two speech recognition service has special focus on Chinese speech and supports multiple Chinese dialects. The output Chinese transcripts have auto punctuation and even without pre-defined 'rare word' dictionary, they are still recognize most of the Chinese name entities, including the name of the go player and the name of the town that the game between AlphaGo and human was held.

Word Tokenization The job of tokenization, or word segmentation is rather straightforward for English, where it is mainly based on the spacing and punctuation. However, for languages that do not require spacing between words, word segmentation can be a challenge. In Chinese, in a long sentence, there might be several possible ways for segmentation, and the final choice largely depends on the semantic meaning of the sentence. There also exist several popular methods for the task of Chinese word segmentation. Stanford Word Segmenter [21] supports multilingual word segmentation, including segmentation options for Chinese. There also exists a Chinese word segmenter [10] making use of lexicon features. With external lexicon features, the segmenter segments more consistently and also achieves a higher F measure when we train and test on the bakeoff data. We also test the performance of the news segmentation on the neural-based method [8], and we find that it outperforms the former two methods, especially for those unusual word pairings and trendy words. Additionally, we remove some of the name entities for both of the Chinese and English words, because the rare name entities are usually poorly encoded in most standard word2vec models.

Stemming and Lemmatization We use the NLTK package to conduct the stemming and lemmatization for English. Because Chinese words do not have such variations as English, words after tokenization can be directly used.

[2] https://www.xfyun.cn/

[3] https://ai.baidu.com/tech/speech/asr

Figure. 6. Near-duplicate keyframes across cultures with different texts. The upper pair is the near-duplicate keyframes about the AirAsia flight news. The lower pair is the near-duplicate keyframes about the AlphaGo vs.Human. We also include the translated English version (from Chinese) from an auto translator for comparison. It can be noticeable that the translation has some slight issue, such as translate Chines word '第六天 (Di Liu Tian, day 6)' to '6th', omitting the '天'(Tian) character.

Chronological Image-Text Pair We use the OpenCV package to decompose the video into a sequence of consecutive frames. According to the timestamp in the audio transcript, we use a sliding window of five seconds to create an image-text pair. For each time interval of five seconds, we uniformly sampled n frames to be the imagery, in association with the texts in the transcripts to construct an image-text pair. Such an approach ensures that the imagery and text are chronologically linked together. If there is no text appearing on the transcript for that 5 seconds interval, we drop the frames and jump to the next word in the transcript, and use its timestamp as the beginning anchor of the next interval. The total number of available image-text pairs largely determines by the word distribution in the transcript. On average, for a news video of a 5-minute duration, there will be more than 300 image-text pairs available.

word	frequency	word	freqency	word	frequency	word	frequency
subscribe	70	twitter	67	instagram	63	twitter	24
vice	60	facebook	54	http	51	follow	15
here	46	visit	36	more	31	bussiness	10

Table. 1. Top blacklist words and their corresponding frequencies on the 'AlphgoGo vs. Human' datasets. We manually identify some templates advertisement text sequences from the video description and calculate their corresponding frequencies based on the collected advertisement sequences.

4.3 Finding Near-Duplicate Images

Near-Duplicate Keyframe Pair To prepare for the training and testing set for our task, we need to find the near-duplicate keyframe in news video from another culture. We establish the set of the near-duplicate keyframes by first selecting an image-text pair in culture A, and then we conduct a greedy search for all the image-text pair in culture B to find the most similar frame. For example, in the AlphaGo vs. Human event, many news videos in the US and China share a similar keyframe, showing the interview of the Go player after the game versus AlphaGo, which can be counted as near-duplicate frames. In our design, we calculate the cosine similarity between visual feature vectors as the distance measurement. We keep only the cross-cultures pairs whose distance is below a threshold τ. This method works quite well over 90% accuracy. After automatic detection, we conduct a manual double-check on the keyframes pair to ensure there are no incorrect pairs in our dataset.

4.4 Video Content Pre-filtering

Noisy Information As we want to scale up the number of videos gathered for each news event, the quality of the news videos remains a problem. For a search using keywords about a news event on video sharing websites such as YouTube and Bilibili, the query results are usually sorted according to their relevance/viewing times. In most case, the first several pages of results will be of higher quality (videos published by large media such as CNN/South China Morning Post), while the remaining results usually contain more noise (videos published by individual uploaders). Noisy news videos often

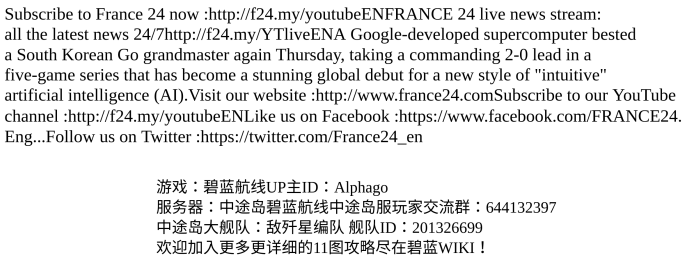


Figure. 7. Example illustrating the challenges of noisy data: Both the English and Chinese text sequences are crawled for the 'AlphaGo vs. Human' news event. On the top is the description of an English video crawled from YouTube, where the description is not informative as others but contains several links of advertisements. On the bottom is the description of a news video crawled from Bilibili, which is irrelevant with our topic but it falsely appears in the results of the search query because the uploader is named as 'alphago'.

contain information irrelevant to the query, which could be advertisement or coverage of other topics. Since there are still a significant amount of valuable feedbacks in the lower-ranked videos of our query, we come out with several ways to filter out the noisy videos in the data cleaning process. Typical examples of the challenges from the noisy data are illustrated in the Figure 9.

Targeting Blacklist Words Advertisements can be identified by some frequently used advertising words (in most cases, to advertise its own video channel or sell products). Considering we have both descriptions and transcripts at hand, where the text sequence containing advertisement in the description is relatively easy to identify, we decide to construct a 'ads word bank' by manually picking out the suspicious text sequences of advertisements. Then we calculate the frequency of words appearing in the 'ads word bank' and consider the Top-K of them as the K blacklist words. Empirically, since the descriptions are short, manually selecting such sequences can be done in a short time, and our further test shows that the 'ads word bank' from 300 videos can already generate good blacklist candidates.

Targeting Whitelist Words Despite the ads word can indeed help eliminate the low-quality news videos of advertising tendency, there still exist some noisy videos which are

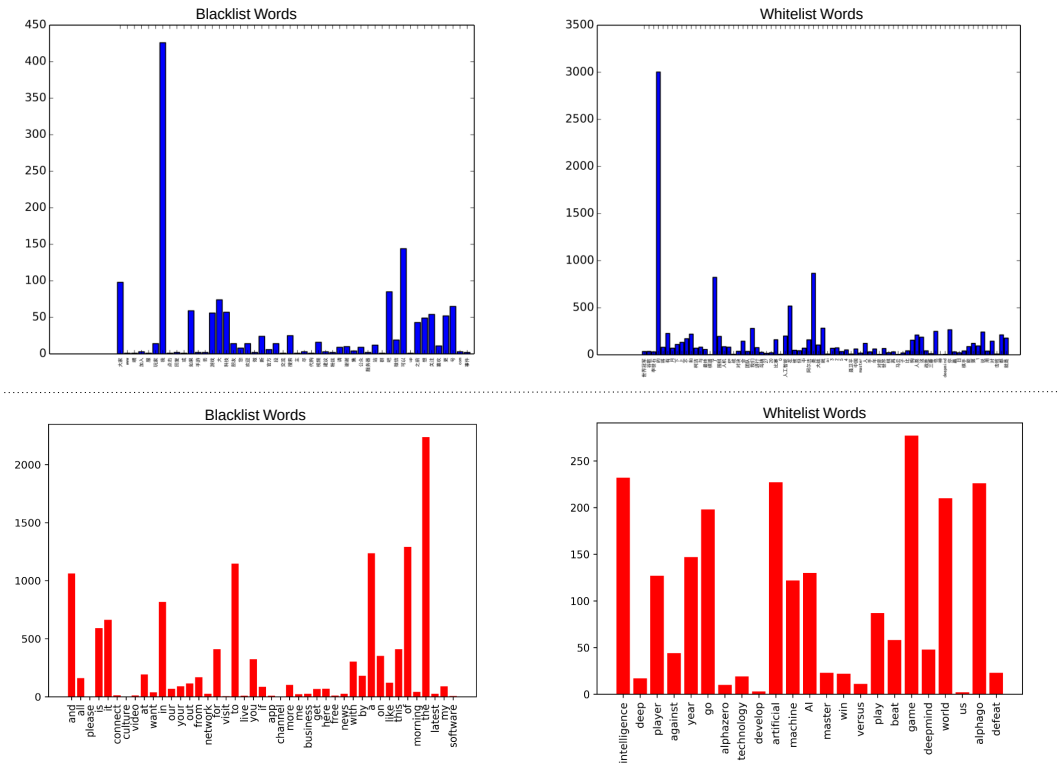


Figure. 8. Occurrences of the blacklist words and whitelist words in Chinese (blue) and English (red) on 'AlphaGo vs. Human Event'. The blacklist words and whitelist words are generated from descriptions of the gathered videos. With the two lists at hand, we calculate the frequency of each word based on the transcripts in the entire dataset. It can be observed from the ticks in the x-axis that the blacklist indeed contains words closely correlated with advertisements, such as *join*, *bussiness*, *free*, *follow*, *please*, *welcome* (translated from Chinese). The whitelist words are closely correlated with the topic, where *alphago*, *play*, *win*, *AI*, *game*, *word champion* (translated from Chinese) are with high frequency.

irrelevant to our topic (One example is shown as the Chinese case in Figure 9). Thus, we further need a whitelist to ensure that the video content and our topic is closely related. Similar to blacklist, we start from the description to manually construct an 'essential word bank', using the remaining sequences in the description after the ads sequences have been picked out. Then we calculate the frequency of words appearing in the 'keyword bank' and consider the Top-K of them as the K whitelist words.

Removing Frequent Words After gathering the whitelist and blacklist for our topic, we first need to eliminate the frequent words in both lists (such as prepositions: *of, with, about...*). The frequent words are usually universally frequently used in the language, regardless of the context and topic. Here we conduct a simple word count process based on the entire descriptions and choose the Top-K frequent words, and we remove these words from both the blacklist and whitelist.

Algorithm 1: Black/White-List Joint Filtering Algorithm

Input video description set D, advertisement word bank $M = \varnothing$, key word bank $N = \varnothing$, blacklist $\mathcal{B} = \varnothing$, whitelist $\mathcal{W} = \varnothing$, blacklist threshold α, whitelist threshold β;

Output decision (as string) for each video;

for $d \in D$ **do**

  select advertisement sequence d_A from d;

  $M = M \cap d_A$;

  $N = N \cap (d - d_A)$;

end

$\mathcal{B}$ = Top-K(M), $\mathcal{W}$ = Top-K(N);

set $a = 0, b = 0, c = 0$;

for $d \in D$ **do**

  for $w \in d$ **do**

    if $w \in \mathcal{B}$ **then**

      $a = a + 1$;

    else if $w \in \mathcal{W}$ **then**

      $b = b + 1$;

    $c = c + 1$;

  end

  if $a/c > \alpha$ **then**

    return 'advertising';

  if $b/c < \beta$ **then**

    return 'irrelevant';

  return 'normal';

end

BW-List Joint Filtering Filtering is conducted based on the blacklist (tendency towards ads) and whitelist (measure of relevance). For each video description sequence, we check its IoU (intersection of union) with the blacklist and whitelist, respectively. Two thresholds for the IoU are set as hyper-parameters. If the IoU of the description

with the blacklist is above the threshold, we mark the video as 'advertising'. Likewise, if the IoU of the description with the whitelist is below the threshold, we mark the video as 'irrelevant'. For the remaining video neither marked as 'advertising' nor 'irrelevant', we mark it as 'normal'. We only use the videos marked as 'normal' for further analysis. Since the thresholds are hyper-parameters, we will discuss the tweaking of the two thresholds in the experiment section.

4.5 Overview of Cultural Datasets

In the previous work conducted by Tasi [55], three datasets are proposed based on the international news event Ebola Virus, AirAsia Flight 8501, and Zika Virus. We additionally several international news events to study cultural differences towards various topics. All of the four news events are long-termed (2 months to 1 year).

AirAsia Flight 8501 There are in total 1000 videos and metadata, in approximately 1:1 (China: US) ratio, in a date range from 12/28/14 to 01/15/15. In total, 4300 keyframes of the US and 2000 keyframes of China are available.

Ebola Virus There are in total 3100 videos and metadata, in approximately 1:3 (Europe: US) ratio, in a date range from 8/21/14 to 11/30/14. In total, 27000 keyframes of the US and 9000 keyframes of Europe are available.

Zika Virus There are in total 1700 videos and metadata, in approximately 7:10 (South Africa: US) ratio, in a date range from 12/01/15 to 02/15/16. In total, 61000 keyframes of the US and 44000 keyframes of South Africa are available.

AlphaGo vs. Human There are in total 1000 videos and metadata, in approximately 6:4 (China: US) ratio, in a date range from 03/09/16 to 03/23/17. In total, 2500 keyframes of the US and 2000 keyframes of China are available.

Thailand Cave Rescue There are in total 400 videos and metadata, in approximately 1:1 (China: US) ratio, in a date range from 06/23/18 to 09/01/18. In total, 1000 keyframes of the US and 600 keyframes of China are available.

Chinese Moon Rover Yutu There are in total 400 videos and metadata, in approximately 1:1 (China: US) ratio, in a date range from 01/11/10 to 03/01/19. In total, 600 keyframes of the US and 700 keyframes of China are available.

Florida High School Shooting There are in total 600 videos and metadata, in approximately 3:7 (China: US) ratio, in a date range from 02/14/18 to 05/01/18. In total, 1500 keyframes of the US and 700 keyframes of China are available.

UIUC Zhang Yingying Kidnapping There are in total 500 videos and metadata, in approximately 6:4 (China: US) ratio, in a date range from 09/07/17 to 06/01/18. In total, 1000 keyframes of the US and 1200 keyframes of China are available.

5 Methods

5.1 Problem Formulation

In this work, we focus on a special type of multimodal retrieval task: image-text retrieval in the multi-cultural setting. As introduced in the previous section, we use the timestamp in audio and video to create multiple text-image pairs, where the texts and images in each pair are temporally linked together. Each text-image pair contains a text sequence and an image sequence, and both sequences may contain an arbitrary number of texts or images.

For a news video v^c from a specific cultural c, we first create multiple image-text pairs. In each valid s second time interval, we define a image-text pair $p^c : (I^c, T^c)$. The image sequence I^c stands for n uniformly sampled frames in that time interval. The text sequence T^c stands for words within that time interval. We use I^c_k to denote the k-th frame of the image sequence, and we use T^c_k to denote the k-th word of the text sequence. Note that the length of the image sequence I^c is set to be n, the actual number of n can be manually determined (in our case, we uniformly sample 10 frames within that time interval), and the length of the text sequence T^c depends on the speech

rate and actual situation in the original news video. In the real-world case, the length of T^c is usually around 10 to 20 words.

We formulate the our cross cultural image-text retrieval task as follows: given a image-text pair $p^m : (I^m, T^m)$ from culture m, our goal is to detect the most suitable text tag t^n from another culture n to describe the image sequence I^m.

5.2 Intra-Modal Feature Extraction

The intuitive idea is to convert the images sequence I^c into a visual feature vector $\mathcal{F}_v(I^c)$, where $\mathcal{F}_v$ denotes the visual feature extractor. Similarly, we need to convert text sequence T^c into a language feature vector $\mathcal{F}_l(T^c)$, where $\mathcal{F}_l$ denotes the textual feature extractor. With two intermediate feature vectors in their modalities, we aim to bridge the gap between modalities and learn the transformation functions to map the visual features and text features into a joint embedding space $\mathbb{S}$. In the embedding space $\mathbb{S}$, cross-modal features can be compared by their corresponding projections onto that space $\mathcal{F}_v(I^c) \rightarrow \mathbb{S}$, and $\mathcal{F}_l(T^c) \rightarrow \mathbb{S}$. For simplicity, we abuse the notation a little bit and use the right arrow $\rightarrow$ to denote projection function $proj_{\mathbb{S}}$.

Visual Features The most common choice for visual embedding is to use the feature maps from the last few layers of many advanced image classification models. A bunch of previous work [55, 59, 56, 60] used the second last layer of VGG-19 or VGG-16 model [51], which is a fully connected layer. The standard procedure is to use weights pre-trained on a large-scale image classification dataset. For each image, it is proposed first resize the image to 256 $\times$ 256 and randomly cropped in ten different ways into 224 $\times$ 224: ten four corners, the center, the mirrored version by flipping the x-axis. The classification network encodes the inputted image, and the output dimension of the two FC layers on the image side is 2048 and 512.

In order to obtain better quality visual features specific to our task, we made two modifications: 1. We use more advanced classification architecture Inception-ResNet proposed in [54] to replace the older VGG model. This architecture entails residue block

to link the high-level and low-level features and is proved to have a higher score in the visual feature extraction task. 2. We propose to enlarge the spatial size of the image to 512×512, which is the image size in ImageNet. We think in the image-text retrieval task across culture, larger image size ensures the near-duplicate keyframes to be closer to entirely-duplicate, which can lead to more convincing performance for learning the joint embedding space. For an image sequence of n frames, we flatten n feature vectors into the averaged feature vector for simplicity. We find that for a short duration of the time interval, averaging the visual feature vectors generally will not jeopardize the performance of feature embedding. In the near-duplicate keyframe detection task, even if we use averaged feature vector, the accuracy of detection is nearly the same as using a single feature vector. ,

Textual Features Usually, there exists a language gap between two cultures (e.g., the US and China), and how to conduct multilingual text embedding should be carefully considered. Intuitively, we can first translate the texts from all other languages towards the primary language a, and then we can use the text embedding model for language a to handle all the scenarios. In the previous work of Tsai [55], all the non-English texts are first translated into English, and then she used an English word2vec model to obtain the textual features for all languages. Although using machine translation can be an effective method to bridge the multilingual gap, errors in translation can never be avoidable. There are always tricky words for a language translation model, where it cannot find direct word-to-word translation but using word-to-phrase instead. To conduct better image-text retrieval, we propose to use multilingual language embedding models instead of translating other languages into English.

We propose to use fastText [18, 43] to conduct native language embedding task. This framework supports multilingual text embedding for 157 languages, where the model for each language is trained on large-scale corpus like Wikipedia. To validate its performance, we select the Chinese language as the testbed to compare the performance with other state-of-the-art methods [35]. The fastText model works surprisingly well, which is close to the other models. The language features generated by fastText is the

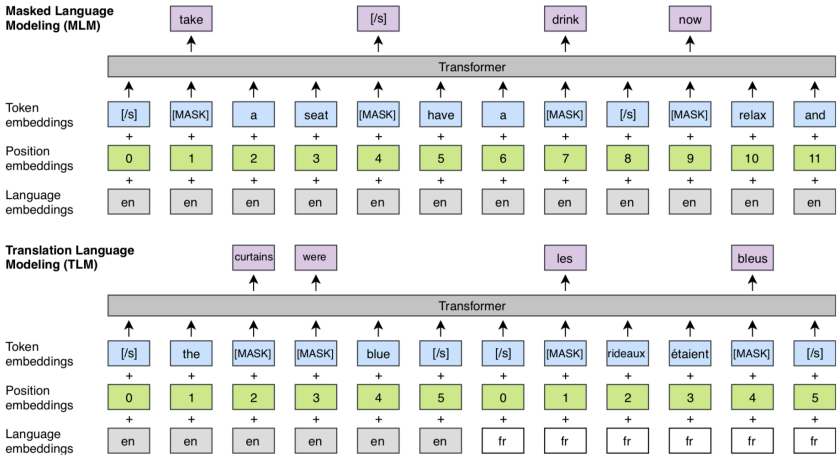


Figure. 9. Example of the Masked Language Modeling used in BERT's [14] training regime. An illustration is adapted from the original paper. Unlike the translation language modeling task, some words of a sentence are randomly masked out, leaving the network to predict the left and right missing words based on the remaining words. Note that the notion of 'left' and 'right' can be far away words, which gives the model the capability to capture the contextual meaning of the sentence.

feature vector of 256 entries. For a text sequence containing n words, we flatten n feature vectors into the averaged feature vector.

Contextualize Text Features The training of word2vec doesn't consider much of the contextual information, which means word2vec might be an effective embedding for words of short phrases but not suitable for long sentences. As its improvement, LSTM and Transformer-based model are later proposed, such as BERT [14], which is a state-of-the-art model for many language tasks. BERT' s critical technical innovation is applying the bidirectional training of Transformer, a popular attention model, to language modeling. This is in contrast to previous efforts that looked at a text sequence either from left to right or combined left-to-right and right-to-left training. BERT details a novel technique named Masked Language Model (MLM), which allows bidirectional training in models in which it was previously impossible. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the Transformer encoder reads the entire sequence of words at once. Therefore it is

considered bidirectional. This characteristic allows the model to learn the context of a word based on all of its surrounding words.

5.3 CCA-Based Multimodal Embeddings

Canonical Correlation Analysis To start with, we consider a two-culture setting: there are a cultures m and and a culture n. Within each culture we have many many image-text pairs, here we denote two sets of image-text pairs as P^m and P^n, where $p^m = (I^m, T^m),\ p^m \in P^m$ and $p^n = (I^n, T^n),\ p^n \in P^n$. Given the original image sequence I^m and text sequence T^m, we use the two feature extractor to convert them into visual-textual feature pairs $f^n : (\mathcal{F}_v(I^m), \mathcal{F}_l(T^m))$ and $f^n : (\mathcal{F}_v(I^n), \mathcal{F}_l(T^n))$, where $\mathcal{F}_v$ denotes the visual feature extractor and $\mathcal{F}_l$ denotes the textual feature extractor.

We assume that the I^m and I^n are near-duplicate keyframe pairs sharing similar imagery, which is denoted as (I^n, I^m). Then we extend the near-duplicate keyframe pairs to its corresponding texts $\{(I^m, T^m), (I^n, T^n)\}$. Our goal is to find a joint embedding space $\mathbb{S}$, where the projection of the two pairs can be best associated.

If use correlation of the projected vectors to represent the association, the method is actually equal to the design of canonical correlation analysis (CCA) [19]. Given two sets of random vectors, CCA aims at find a linear combination of the two vectors $\{v_1, v_2\}$ to represent the joint embedding space $\mathbb{S}$, and the correlation of the projected vectors should be maximized:

$$\underset{\mathbb{S}}{\text{argmax}}\ corr(proj_{\mathbb{S}} v_1,\ proj_{\mathbb{S}} v_2),\ \mathbb{S} \in span(v_1, v_2)$$

Two-Way Embeddings Assume that we have two near duplicate images sand we find their corresponding texts $\{(I^m, T^m), (I^n, T^n)\}$. To apply CCA on our cross culture analysis task, $\{(I^m, T^m), (I^n, T^n)\}$ can be further represented as the three ordinary

pairs that exploit already known image-image and image-text matching:

$$\{I^m, T^m\},\ \{I^n, T^n\},\ \{I^m, I^n\}$$

The former two pairs indicate the linkage inside culture m and culture n, and the third pair indicate that I^m and I^n should be near-duplicated. We use the three pairs as input to CCA.

Let $X \in \mathbb{R}^{D_x}$ be the collection of the left elements of those three types pairs, and let $Y \in \mathbb{R}^{D_y}$ be the collection of the right elements of those three types of pairs. The objective of CCA is to find $u_x \in \mathbb{R}^{D_x}$ and $u_y \in \mathbb{R}^{D_y}$ such that the projection of X, Y onto u_x and u_y are maximally correlated:

$$\begin{aligned}(u_x^*, u_y^*) &= \underset{u_x, u_y}{\operatorname{argmax}}\ corr(u_X^T X, u_y^T Y) \\ &= \underset{u_x, u_y}{\operatorname{argmax}}\ \frac{u_x^T \sum_{xy} u_y}{\sqrt{u_x^T \sum_{xx} u_x u_y^T \sum_{yy} u_y}}\end{aligned} \tag{1}$$

where $\sum_{xy}$ is the covariance matrix between the two views, and $\sum_{xx}$ and $\sum_{yy}$ is the covariance matrix within each view. For CCA problem, the optimal k-dimensional projection mappings are provided as a closed form solution by the rank-k singular value decomposition (SVD) of the $D_x \times D_y$ matrix $\sum_{xx}^{-\frac{1}{2}} \sum_{xy} \sum_{yy}^{-\frac{1}{2}}$, as proved by Johnson in [24].

Deep CCA The major drawback of CCA is that it needs to store all of the training set while testing, which costs a significant amount of memory. Another problem is that CCA only applies linear combination, thus performs poorly on two vectors with a non-linear relationship. To alleviate such problems, deep canonical correlation analysis (DCCA) [1] is proposed to capture the hidden non-linear relationship of the data. DCCA also does not require the training set upon testing, which also improves the time consumption of single testing to constant. In the DCCA model, we aim at learning the two branches of deep neural networks f and g to extract features from view X and view Y, respectively. The two-branch neural network is constrained by its final

objective function, which maximizes the correlation between the outputs of the two-branch network. The objective function can be expressed as:

$$(\mathbf{W}_f^*, \mathbf{W}_g^*, u_f^*, u_g^*) = \underset{u_f^*, u_g^*}{\operatorname{argmax}} \ corr(u_f^T f(X),\ u_g^T g(Y))$$

where $\mathbf{W}_f^*$ and $\mathbf{W}_g^*$ denotes the optimal weights learned by network f and g. Such a multivariate optimization problem has no closed-form solution, but the optimal solution can be approximated by the gradient descent approach. The weights $\mathbf{W}_f^*$ and $\mathbf{W}_g^*$ can be trained following standard deep learning pipeline, using backward propagation of the loss term.

For our task, we do not need to train the two embedding branch network f and g from scratch, which directly transforms the original image/text data to the shared embedding space. The reason for not doing that is mainly because model visual and language features explicitly by existing expert networks (e.g., VGG and word2vec) can achieve better results for feature representation. Instead, we propose to utilize the expert embedding frameworks first to extract visual and language features, then we pass the intermediate features to f and g, and learn the joint embedding. The language and the visual expert network can fully utilize the large-scale dataset (e.g., ImageNet and multilingual Wikipedia text dataset) and benefited by the high-quality features using transfer learning. The modified objective function for our task can be expressed as:

$$(\mathbf{W}_f^*, \mathbf{W}_g^*, u_f^*, u_g^*) = \underset{u_f^*, u_g^*}{\operatorname{argmax}} \ corr(u_f^T f(\mathcal{F}_v(I^c)),\ u_g^T g(\mathcal{F}_l(T^c)))$$

The $\mathcal{F}_l$ denotes the expert language feature extractor, which is fastText in our case. The $\mathcal{F}_v$ denotes the expert visual feature extractor, which is Inception-ResNet, in our case. The $\mathcal{F}_v(I^c)$ and $\mathcal{F}_l(T^c)$ denotes the intermediate features outputted by the expert feature extractors. When training, we load the pertained weights of $\mathcal{F}_v$ and $\mathcal{F}_l$ and only update weights for f and g. Considering the criteria to select near-duplicate keyframes have already guaranteed the similarity of the visual features, we only use the two cross-modal pairs $\{I^m, T^m\},\ \{I^n, T^n\}$ to train the f and g.

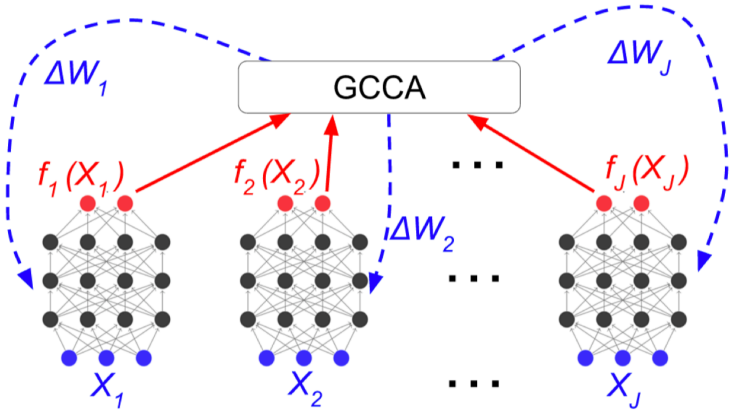


Figure. 10. Schematic Illustration of DGCCA, which is originally proposed in [4]. The f_j denotes the j-th (branch) of the network to project j-th input data entity X_j into the GCCA joint embedding space. Our objective function aims at maximize the total sum of the inter-input correlation, which calculates the correlation between each of pair of the two inputs.

Generalized Version of CCA Limited by the definition of the covariance, both CCA and DCCA can only handle inputs of two modalities. As an extension to handle an arbitrary number of modalities, generalized canonical correlation analysis (DCCA) is proposed in [13], with its deep learning version DGCCA later proposed in [4]. The objective of GCCA is to find a shared representation of G of J ($J \geq 2$) different views with maximum inter-correlation. The objective function can be written as:

$$U_j^* = \underset{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}}{\arg\min} \sum_{j=1}^{J} ||G - U_j^T X_j||_F, \text{ where } GG^T = I_r$$

where N is the number of data points (in our case is 3), d_j is the dimension of the j-th view, r is the dimension of the learned representation, and $X_j \in \mathbb{R}^{d_j \times N}$ is the data matrix for the j-th view. We can first find the eigendecomposition of a $N \times N$ matrix to solve GCCA, where the $N \times N$ matrix scales quadratically with the sampled size and leads to extreme memory consumption. Unlike CCA and DCCA, which only learn projections or transformations on each of the views, GCCA also learns a view-independent representation G that best reconstructs all of the view-specific representations.

Deep Generalized CCA For the gradient descent version of GCCA, the key idea is to construct a embedding network with J branches. Now we need to replace the data matrix X in the previous GCCA objective function with the feature matrix $f(X)$, where f denotes the network with J-branches. The objective function becomes:

$$U_j^* = \underset{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}}{\arg\min} \sum_{j=1}^{J} ||G - U_j^T f_j(X_j)||_F, \text{ where } GG^T = I_r$$

where $f_j(X_j)$ indicates applying j-th branch of the network on the j-th input.

For optimization, we define $C_{jj} = f(X_j)f(X_j)^T \in \mathbb{R}^{o_j \times o_j}$ to be the scaled empirical covariance matrix of the j-th network output. We define $P_j = f(X_j)^T \ C_{jj}^{-1} f(X_j) \in \mathbb{R}^{N \times N}$ to be the corresponding projection matrix of the data. Then the reconstruction error should be expressed as follows:

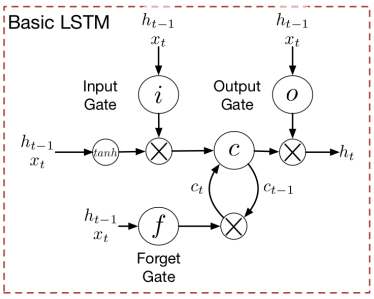
$$e = \sum_{j=1}^{J} ||G - U_j^T f_j(X_j)||_F^2 = \sum_{j=1}^{J} ||G - G f_j(X_j)^T C_{jj}^{-1} f_j(X_j)||_F^2 = rJ - Tr(GMG^T)$$

where minimizing the obejective function equals to maximizing $Tr(GMG^T)$, with the sum of eigenvectors $L = \sum_{i=1}^{r} \lambda_i(M)$. Taking derivation of L based on $f_j(X_j)$, we have:

$$\frac{\partial L}{\partial f_j(X_j)} = 2(U_j G - U_j U_j^T f_j(X_j))$$

Thus, the gradient is the difference between the r-dimensional auxiliary representation G embedded into the subspace spanned by the columns of U_j (the first term) and the projection of the actual data in $f_j(X_j)$ onto the subspace (the second term).

Three-Way Embeddings Assume the near-duplicate image pairs can be viewed as equal, the previous near-duplicate image-text pair $\{(I^m, T^m), (I^n, T^n)\}$ can be simplified to $\{I^{mn}, T^m, T^n\}$, where $I^{mn} = I^m = I^n$. For the objective function of GCCA, now the data matrix of X involves three inputs $\{I^{mn}, T^m, T^n\}$. The objective function



$$
\begin{aligned}
\boldsymbol{i}_t &= \text{sigmoid}(\boldsymbol{W}_i[\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{h}_{t-1}^{\mathrm{T}}]^{\mathrm{T}}) \\
\boldsymbol{f}_t &= \text{sigmoid}(\boldsymbol{W}_f[\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{h}_{t-1}^{\mathrm{T}}]^{\mathrm{T}}) \\
\boldsymbol{o}_t &= \text{sigmoid}(\boldsymbol{W}_o[\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{h}_{t-1}^{\mathrm{T}}]^{\mathrm{T}}) \qquad (1) \\
\boldsymbol{c}_t &= \boldsymbol{i}_t \odot \tanh(\boldsymbol{W}_c[\boldsymbol{x}_t^{\mathrm{T}}, \boldsymbol{h}_{t-1}^{\mathrm{T}}]^{\mathrm{T}}) \\
&\quad + \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t),
\end{aligned}
$$

Figure. 11. Illustration of LSTM unit from [69]. The memory cell is modulated jointly by the input, output, and forget gates to control the knowledge transferred at each time step. The $\odot$ denotes element-wise products.

of DGCCA becomes:

$$
\begin{aligned}
U_j^* = \underset{U_j \in \mathbb{R}^{d_j \times r}, G \in \mathbb{R}^{r \times N}}{\arg\min} \quad & ||G - U_j^T f^v(I^{mn})||_F \; + \; ||G - U_j^T f_m^l(T^m)||_F \\
& + ||G - U_j^T f_n^l(T^n)||_F, \text{ subject to } GG^T = I_r
\end{aligned} \qquad (2)
$$

where f_m^l stands for language embedding model for a language m and f^v stands for visual embedding model for imagery. Therefore, by training the common representation for the triplets $\{I^{mn}, T^m, T^n\}$, we can embed the three inputs into the joint embedding space. Using the same strategy as in DCCA, we first load the pretrained weights from expert networks (fastText(Eng), fastText(Chi), Inception-ResNet) to obtain the intermediate features, then optimize the weights in the three branches: f_v, f_l^n and f_l^m to obtain the final joint embedding.

5.4 Temporal Relationship Reasoning

Why Encode Temporal Dependency In the previous work [55] of a similar task, only keyframes and video descriptions are used. For imagery, each time only one frame is encoded, and there is no concern temporal relationship among the selected keyframes. For descriptions, there is no explicit timeline like the audio file, so she randomly selects n tags and averaging the language features vectors for embedding. For our baseline CCA-based model in section 4.3, one improvement is that we use speech sequence and image sequence for the encoding job. The temporal information is encoded implicitly because

the image-text pair in our dataset guarantees that in the original news video image, are text are chronologically close to each other. Nevertheless, such loose constraints can be potentially dangerous because there is no guarantee for the intra-chronological-order within a visual or language sequence. For example, for a sentence of three words 'Tom eats an apple', if we encode this sentence using word2vec and flatten the feature vectors, the vectors could represent 16 different results 'Tom eat(s) an apple', 'an apple eat (s) Tom', ... etc. On the other hand, pooling the visual/text feature vectors could be harmful for the accurate feature representation, features after pooling will be more ambiguous, and inferring original visual/text from features becomes more difficult after such temporal-pooling operations. For visual representation, this problem could be as equal severe as the frame numbers increases, uniformly sample the keyframes might omit important frames, and temporal order between consecutive frames are vital for the overall video understanding (hence, if the frames are completely out of order, even human beings are difficult to summarize the news in the videos).

LSTM-based Sequence Encoder A popular approach for encoding video/speech sequence is to use the Long-Short-Term-Machine (LSTM) [20], which has been used for video summarization [69] and visual question answering (VQA) [37]. LSTM is a bionically designed recurrent neural network that is adept at modeling long-range feature dependencies. The dependency can be continuous attributes such as time. At the core of the LSTM, there exist some memory cells c, which encode, at every time step, the knowledge of the inputs that have been observed up to that step. The cells are then modulated by non-linear gates, which are usually constructed by logistic (sigmoid) functions. The gates determine whether the LSTM should keep the information (if the gates return 1) or discard them (if the gates return 0). There are three gates: the input gate (i) controlling whether the LSTM considers its current input (x_t), the forget gate (f) allowing the LSTM to forget its previous memory (c_t), and the output gate (o) decides the amount of the memory to transfer to the hidden states (h_t). Together, they provide LSTM with the capability to learn complex long-term dependencies. In particular, the forget date serves as a time-varying data-dependent on/off switch to

selectively incorporating the past and present information. This design makes LSTM extremely suitable for encoding temporal sequences such as videos and speeches.

We propose to use separate LSTM networks to encode the visual and textual sequences, respectively, to capture the temporal dependency within each sequence. The inputs to the LSTM units are intermediate image/text features from the expert feature extraction network. Let d denote the size of the hidden state of the LSTM unit; The temporally encoded text and images are represented by $H \in \mathbb{R}^{2d \times T \times 2}$, where T denotes the maximum length of the sequence. The query is represented as a matrix Q of concatenated bi-directional LSTM outputs, i.e., $Q \in \mathbb{R}^{2d \times M}$, where M is the maximum length of the query.

Intra-Sequence Temporal Constraint Inspired by work in VQA [37] to maintain the temporal consistency, we introduce the temporal correlation matrix $C \in \mathbb{R}^{T \times T}$ as a constraint to ensure the data matrix H and query matrix Q is encoded chronologically. Let $h_i = H_{:i:} \in \mathbb{R}^{2d \times 2}$ to denote the visual/text representation for the i-th timestep in the multimodal data matrix H. The entry C_{ij} is calculated by:

$$C_{ij} = \tanh \sum_{k=1}^{K} \mathbf{w}_c^T (\mathbf{w}_h^T \cdot \text{sim}(h_{ik}, h_{jk}) + Q_{:M})$$

where K is the number of modalities, in our case, K = 2. The operator : is a slicing operator to extracts all elements from a dimension, where $h_{i1} = H_{:i1}$ $\text{sim}(h_{ik}, h_{jk})$ denotes the similarity between the two-modal features. The $\mathbf{w}_c \in \mathbb{R}^{2d \times 1}$ and $\mathbf{w}_h \in \mathbb{R}^{4d \times 2d}$ are parameters to learn.

5.5 Non-negative Matrix Factorization

In the test time, for each query of image I and text t, the canonical correlation analysis predicts a correlation score between -1 and 1. If in total, we consider N images for the query, and for each of the image, we predict the correlation between each of the text tag from a tag dictionary of size K, we can have a correlation matrix $\mathcal{M}$ of size $N \times K$.

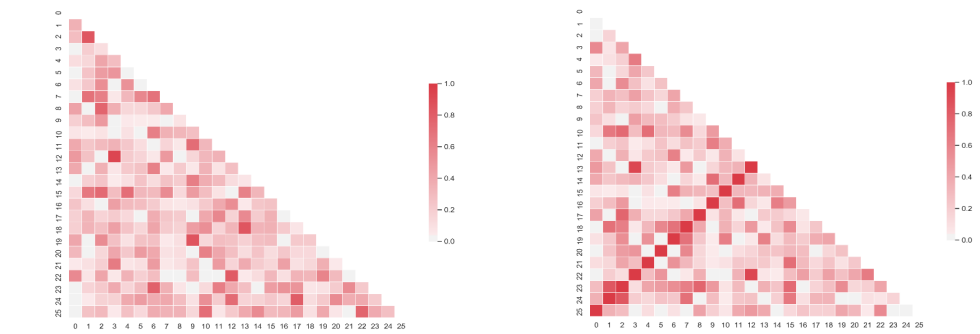


Figure. 12. Left: correlation matrix of 25 images and 25 candidate tags. Right: reorganized correlation matrix by Cuthill—McKee algorithm. After the ranking process, the correlation matrix appears as a pseudo-symmetric form.

However, the correlation value itself cannot provide more useful qualitative indications for human evaluation, since we only know if one image and one tag are correlated or not but cannot know the actual cause to explain such correlation. Up until now, the CCA completes the task of 'tagging' by finding the Top-K tags for each of the images, but the actual culture differences are hard to observe and explain.

Inspired by the latent models in collaborative filtering [30], we can treat the correlation matrix as a rating matrix such as IMDB user-item rating matrix [33], and conduct matrix decomposition based on that. We aim to decompose the correlation matrix into the multiplication of two small matrices, which representing the dominating concepts in visual and language perspective. While Singular Value Decomposition (SVD) is commonly used for standard latent model collaborative filtering, it generates three resulting matrices with potentially negative values, which is not suitable for our case. Instead, we choose to conduct non-negative matrix decomposition [34] on the correlation matrix $\mathcal{M}$. We assume M of size $N \times K$ to be approximated by the matrix product of two thin matrices $\mathcal{U}$ of size $N \times F$, and $\mathcal{V}$ of size $F \times K$, where:

$$\mathcal{M} \approx \mathcal{U}\mathcal{V}$$

and the number F is a user-defined hyper-parameter, usually explained at the number of latent factors or concepts. With a larger value of F, the final factors obtained after decomposition tend to be more fine-grained, and the approximation of the two small matrices towards the correlation matrix tends to be more precise.

To obtain an approximate factorization $\mathcal{M} \approx \mathcal{U}\mathcal{V}$, we first need to define cost functions which can quantify the quality of the approximation. Such a cost function can be constructed using some measure of distance between two non-negative matrices $\mathcal{U}$ and $\mathcal{V}$. One simple distance commonly used is the L2 distance:

$$||\mathcal{M} - \mathcal{U}\mathcal{V}||^2 = \sum_{ij}(\mathcal{M}_{ij} - \sum_{k}\mathcal{U}_{ik}\mathcal{V}_{kj})^2$$

which is lower bounded by zero. Another commonly used distance metric is:

$$D(\mathcal{M} \mid\mid \mathcal{U}\mathcal{V}) = \sum_{ij}(\mathcal{M}_{ij} log \frac{\mathcal{M}_{ij}}{\sum_k \mathcal{U}_{ik}\mathcal{V}_{kj}} - \mathcal{M}_{ij} + \sum_{k}\mathcal{U}_{ik}\mathcal{V}_{kj})$$

Since this metric is not symmetric on $\mathcal{M}$ and $\mathcal{U}\mathcal{V}$, strictly speaking, it cannot be called as 'distance', and here we refer it to 'divergence'. If

$$\sum_{ij}\mathcal{M} = \sum_{ij}\sum_{k}\mathcal{U}_{ik}\mathcal{V}_{kj} = 1$$

where $\mathcal{M}$ and $\mathcal{U}\mathcal{V}$ follows normal distribution, the divergence metric will be simplified to Kullback-Leibler divergence, which is also known as the relative entropy.

Since both of the metrics can be defined as matrix operation, we can use the chain rule to back-propagate the error to each element of $\mathcal{U}$ and $\mathcal{V}$, and update them using commonly-used optimizer like gradient descent (GD).

	Query	Recall@1	Recall@5	Recall@10
Ebola Virus	US images query EU tags	8.2	29.8	40.3
	EU images query US tags	9.5	29.5	44.1
AirAsia Flight	US images query CN tags	7.6	18.8	29.1
	CN images query US tags	9.3	23.3	36.4
Zika Virus	US image query SA tags	11.8	31.2	54.1
	SA images query US tags	9.6	32.9	52.7
AlphaGo vs. Human	US images query CN tags	10.1	25.4	41.3
	CN images query US tags	11.6	27.2	45.9
Thailand Cave Rescue	US images query CN tags	10.4	25.1	38.9
	CN images query US tags	10.9	27.0	41.8

Table. 2. Performance of intra-culture image-text queries. The Ebola, AirAsia, Zika dataset is bsaed on prior work in [55], Thailand Cave Rescue and AlphaGo vs. Human are our new datasets.

α	0.01	0.025	0.01	0.025	0.025	0.05	0.05
β	0.02	0.02	0.05	0.05	0.05	0.1	0.1
advertising	4/0	6/1	4/0	6/1	6/1	8/3	8/3
irrelevant	98/91	98/91	26/21	26/21	26/21	7/3	7/3

Table. 3. Experiments on the the two thresholds for the blacklist/whitelist filtering. We test it based on a test set of 150 videos for 'AlphaGo vs. Human' event. The 'advertising' and 'irrelavant' videos are manually picked out within the test dataset. For the performance metric, each x/y denotes the total positive/false positive. We find that $\alpha = 0.05$ and $\beta = 0.1$ could be relatively ideal for the BW-List Joint Filtering Task.

6 Experiments

6.1 Implementation Details

In the previous work [55], we have three datasets: Ebola Virus, Zika Virus, Air Asia Flight 8501. We collect additional datasets based on the feedback of the US and Chinese people on the 'AlphaGo vs. Human' news event. For speech recognition, we use the pre-trained model provided by Google Speech Recognition to convert English and Chinese speech sequences into text sequences. We use standard NLP pipelines provided by NLTK to remove the stop words, and then preprocess the texts extracted by WordNet's Lemmatizer. Then we obtain the image-text sequence pair using a 5-

second sliding window. The sampling rate for keyframes is set to be two frames per second. For visual feature embedding, we resize the images into 584×584 and conduct random cropping and flipping to generate resulted input frame of 512×512. We use feature output by the last two fully connected layers in the Inception-ResNet model and load existing weight pre-trained on ImageNet.

For near-duplicate keyframes, we set the threshold τ to be 20 to 45, and we observe that threshold at around 30 can already satisfy the requirement of very similar keyframes. For text embedding, we use the pre-trained word2vec model on multilingual the Wikipedia dataset to conduct the language embedding. To learn the joint embedding space, we modified the previous DCCA code based on MATLAB and reimplemented the pipeline on PyTorch. We use Adam [28] optimizer with parameter of $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The initial learning rate is set to be $1e^{-2}$, and the learning rate should automatically decrease by 10% after every 5 epoch.

6.2 Experimental Results

We first conduct a simple search for the hyper-parameters α and β used as the blacklist/whitelist filtering threshold. As shown in Table 3, the variations of α and β is set to be $0.01, 0.02, 0.025, 0.05$ and 0.10. For the testing set, we manually pick out the 'advertising' and 'irrelevant' videos within 150 videos, which is randomly sampled in the total dataset for the Chinese videos on 'AlphaGo vs. Human' news event. We use the total and true positive as the standard to evaluate the quality of the filtering. After several trails in tweaking, we find that $\alpha = 0.05$ and $\beta = 0.1$ could be relatively ideal for the filtering task.

For each news event, we generate two culture-specific image-text embedding space for two different cultures. For AirAsia Flight 8501, Thailand Cave Rescue, and AlphaGo vs. Human, one for the US and one for China. For Ebola Virus, one for the US and one for Europe. For Zika Virus, one for South Africa and one for the US. For each culture-specific joint embedding space, we randomly select 1000 images with the tags, or 10% or the dataset, whichever is smaller as the testing set. Each time of the query,

Figure. 13. Visual concepts extracted from non-negative matrix decomposition with the number of factors to be 10. The upper row represents videos for US audiences, which consists of critical images in the second row of the left decomposed visual matrix under US culture. The low row represents videos for China audiences, which consists of critical images in the second row of the left decomposed visual matrix under Chinese culture. It can be observed that the 'concept NO.2' in Chinese culture has a special focus on the go board and close view of the go chess, while 'concept NO.2' in US culture does not share the same characteristic.

we randomly select a keyframe from culture A to query the text tags in culture B. For performance evaluation, we use Recall@1, Recall@5, Recall@10 as the testing metric. For AlphaGo event, we achieved 45.9% in Recall@10, which is slightly higher than the previous results. Thailand Cave Rescue performs similar to the AlphaGo news event, where the Recall@10 is relatively lower. We further try to swap the word2vec text features with BERT features, which doesn't lead to a boost in the performance. Due to limited time, we haven't tested on all of our newly gathered datasets. Further ablation study needs to be carried to determine which proposed component leads to a significant accuracy boost.

We further experiment with matrix factorization of the correlation matrix. We treat 25 randomly selected images as one image group and construct a 25×25 correlation matrix, where the second dimension denotes the top 25 tags in all tag candidates. We conduct the experiment on the AlphaGo dataset and construct the correlation matrix

for both Chinese and US culture. The correlation matrix is first re-organized into pseudo-symmetric form by Cuthill—McKee algorithm. Then we experiment with the non-negative matrix factorization and set different values for the number of factors in the decomposition: 5, 10, 25, 20. Visualization of the decomposed concept matrices can be found in Figure 14. We find that setting the decomposed concept number to 5 or 10 usually gives the most explicit concept separation. As shown in Figure 13, we discover that with factor number to be 10, we can find similarities in the images of Chinese culture, where all of the images more or less contain imagery of the go board.

7 Conclusion and Future Work

In this work, we exploit the image-text retrieval task to discover the text tagging differences in cross-cultural and multilingual news videos. Based on the prior work conducted in [**?**], we propose several improvement ideas for visual/textual feature extraction and bridge the language gap using speech recognition and native text embedding models. We also particularly study on the temporal encoding method to extract the image-text sequence. We use an implicit encoding method to ensure the image and language features are chronologically related in our analysis. Furthermore, we rewrite most of the code in [55] to adopt the developing environments, transforming the entire pipeline to the native Python environment, which is much easier to use, update, and deploy.

For our feature work, several potential improvements can be explored: 1. Carry out the ablation study based on controlled variants to analyze the performance of our proposed components. 2. Apply the newest deep generalized CCA and conduct three-way embeddings. 3. Explore the LSTM temporal grounding for both visual and textual features, with emphasis on the Chronological ordering. 4. Find the problem of using contextualized text features in our model.

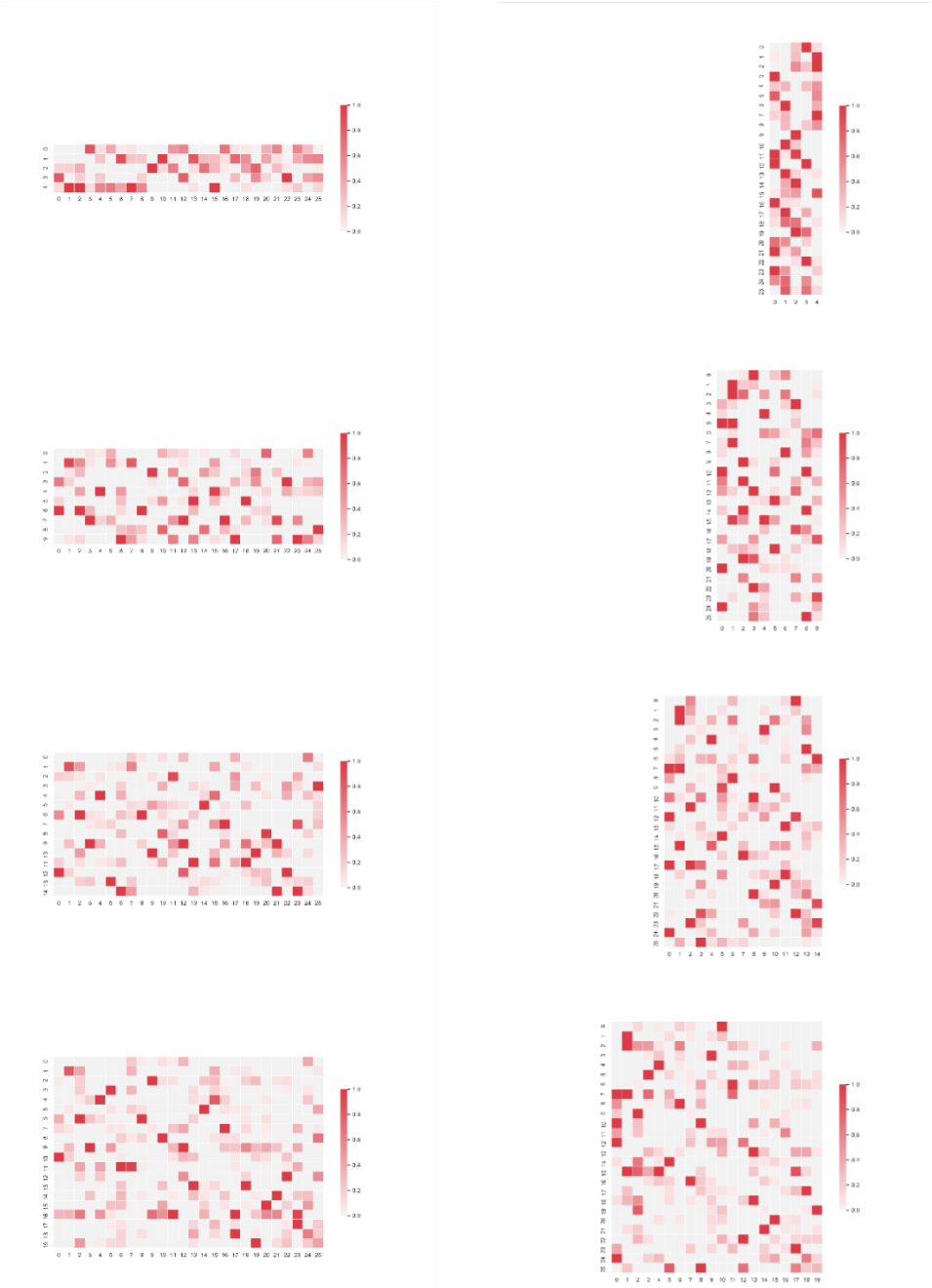


Figure. 14. Visualization of the decomposed two sub-matrices with different settings on the number of factors, on the 'AlphaGo vs. Human' dataset. Left column: decomposed latent matrix indicating scores of visual concepts. Right column: decomposed latent matrix indicating scores of language concepts. From top to bottom, we change the number of concepts to be 5, 10, 15, 20. For the most case of our experiments, we usually observed that non-negative matrix decomposition with a number of factors to be 10 archives the best separation between concepts.

References

[1] G. Andrew, R. Arora, J. Bilmes, and K. Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[2] B. Arons. A review of the cocktail party effect. *Journal of the American Voice I/O Society*, 12(7):35–50, 1992.

[3] P. Bao, B. Hecht, S. Carton, M. Quaderi, M. Horn, and D. Gergle. Omnipedia: bridging the wikipedia language gap. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1075–1084. ACM, 2012.

[4] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora. Deep generalized canonical correlation analysis. *arXiv preprint arXiv:1702.02519*, 2017.

[5] S. Bergsma and B. Van Durme. Learning bilingual lexicons using the visual similarity of labeled web images. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011.

[6] J. Burgess and J. Green. *YouTube: Online video and participatory culture*. John Wiley & Sons, 2018.

[7] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[8] D. Cai, H. Zhao, Z. Zhang, Y. Xin, Y. Wu, and F. Huang. Fast and accurate neural word segmentation for chinese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 608–615, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[9] H. Cai, Y. Yang, X. Li, and Z. Huang. What are popular: exploring twitter features for event detection, tracking and visualization. In *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015.

[10] P.-C. Chang, M. Galley, and C. D. Manning. Optimizing chinese word segmentation for machine translation performance. In *Proceedings of the third workshop on statistical machine translation*, pages 224–232. Association for Computational Linguistics, 2008.

[11] T. Chen, H. M. SalahEldeen, X. He, M.-Y. Kan, and D. Lu. Velda: Relating an image tweet's text and images. In *AAAI*, 2015.

[12] P. Clough, H. Müller, T. Deselaers, M. Grubinger, T. M. Lehmann, J. Jensen, and W. Hersh. The clef 2005 cross–language image retrieval track. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, 2005.

[13] C. A. Coelho. *Generalized Canonical Analysis*. PhD thesis, 1992.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] W. Dong and W.-T. Fu. Cultural difference in image tagging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 981–984. ACM, 2010.

[16] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, T. Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[17] C. A. Gomez-Uribe and N. Hunt. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):13, 2016.

[18] E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[19] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[20] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] T. Huihsin, P. Chang, G. Andrew, D. Jurafsky, and C. Manning. A conditional random field word segmenter. In *Fourth SIGHAN Workshop*, 2005.

[22] A. Jabri, A. Joulin, and L. van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, 2016.

[23] Y.-G. Jiang, G. Ye, S.-F. Chang, D. Ellis, and A. C. Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*. ACM, 2011.

[24] R. A. Johnson and D. W. Wichern. Multivariate analysis. *Encyclopedia of Statistical Sciences*, 8, 2004.

[25] B. Jou, H. Li, J. G. Ellis, D. Morozoff-Abegauz, and S.-F. Chang. Structured exploration of who, what, when, and where in heterogeneous multimedia news sources. In *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013.

[26] A. Karpathy, A. Joulin, and L. F. Fei-Fei. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[27] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[28] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[29] B. Klein, G. Lev, G. Sadeh, and L. Wolf. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399*, 2014.

[30] Y. Koren and R. Bell. Advances in collaborative filtering. In *Recommender systems handbook*, pages 77–118. Springer, 2015.

[31] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, (8):30–37, 2009.

[32] P. L. Lai and C. Fyfe. Kernel and nonlinear canonical correlation analysis. *International Journal of Neural Systems*, 10(05):365–377, 2000.

[33] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008.

[34] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[35] S. Li, Z. Zhao, R. Hu, W. Li, T. Liu, and X. Du. Analogical reasoning on chinese morphological and semantic relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 138–143. Association for Computational Linguistics, 2018.

[36] W. Li, J. Joo, H. Qi, and S.-C. Zhu. Joint image-text news topic detection and tracking by multimodal topic and-or graph. *IEEE Trans. Multimedia*, 2017.

[37] J. Liang, L. Jiang, L. Cao, L.-J. Li, and A. Hauptmann. Focal visual-text attention for visual question answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[38] B. Y. Lin, F. F. Xu, K. Zhu, and S.-w. Hwang. Mining cross-cultural differences and similarities in social media. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018.

[39] D. Liu, X.-S. Hua, L. Yang, M. Wang, and H.-J. Zhang. Tag ranking. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 351–360, New York, NY, USA, 2009. ACM.

[40] A. L. Maas, R. E. Daly, P. T. Pham, D. Huang, A. Y. Ng, and C. Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics.

[41] B. Mahasseni, M. Lam, and S. Todorovic. Unsupervised video summarization with adversarial lstm networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[42] I. Mani and E. Bloedorn. Summarizing similarities and differences among related documents. *Information Retrieval*, 1999.

[43] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[44] H. Nakasaki, M. Kawaba, T. Utsuro, and T. Fukuhara. Mining cross-lingual/cross-cultural differences in concerns and opinions in blogs. In *International Conference on Computer Processing of Oriental Languages*, 2009.

[45] A. Popescu and I. Kanellos. Multilingual and content based access to flickr images. In *Information and Communication Technologies: From Theory to Applications, 2008. ICTTA 2008. 3rd International Conference on*, 2008.

[46] V. Ramanishka, A. Das, D. H. Park, S. Venugopalan, L. A. Hendricks, M. Rohrbach, and K. Saenko. Multimodal video description. In *Proceedings of the 2016 ACM on Multimedia Conference*, 2016.

[47] P. Resnick and H. R. Varian. Recommender systems. *Communications of the ACM*, 40(3):56–59, 1997.

[48] A. Rohrbach, M. Rohrbach, R. Hu, T. Darrell, and B. Schiele. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, 2016.

[49] S. Seyednezhad, K. N. Cozart, J. A. Bowllan, and A. O. Smith. A review on recommendation systems: Context-aware to social-based. *arXiv preprint arXiv:1811.11866*, 2018.

[50] S. Siersdorfer, J. San Pedro, and M. Sanderson. Automatic video tagging using content redundancy. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 395–402, New York, NY, USA, 2009. ACM.

[51] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[52] R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association of Computational Linguistics*, 2014.

[53] S. Song, S. P. Lichtenberg, and J. Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[54] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[55] C.-Y. Tsai and J. R. Kender. Detecting culture-specific tags for news videos through multimodal embedding. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, Thematic Workshops '17, 2017.

[56] C.-Y. Tsai, R. Xu, R. E. Colgan, and J. R. Kender. News event understanding by mining latent factors from multimodal tensors. In *Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion*. ACM, 2016.

[57] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, 2015.

[58] J. Wang, X. Zhu, and S. Gong. Video semantic clustering with sparse and incomplete tags. In *AAAI*, 2016.

[59] L. Wang, Y. Li, J. Huang, and S. Lazebnik. Learning two-branch neural networks for image-text matching tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[60] L. Wang, Y. Li, and S. Lazebnik. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[61] W. Wang, X. Yan, H. Lee, and K. Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

[62] X. Wang, L. Xie, M. Lu, B. Ma, E. S. Chng, and H. Li. Broadcast news story segmentation using conditional random fields and multimodal features. *IEICE TRANSACTIONS on Information and Systems*, 2012.

[63] J. Weston, S. Bengio, and N. Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[64] J. Weston, S. Bengio, and N. Usunier. Wsabie: Scaling up to large vocabulary image annotation. In *IJCAI*, 2011.

[65] X. Wu, C.-W. Ngo, and A. G. Hauptmann. Multimodal news story clustering with pairwise visual near-duplicate constraint. *IEEE Transactions on Multimedia*, 2008.

[66] L. Xie, A. Natsev, J. R. Kender, M. Hill, and J. R. Smith. Visual memes in social media: tracking real-world news in youtube videos. In *Proceedings of the 19th ACM international conference on Multimedia*, 2011.

[67] M. Yao. You-get: Dumb downloader that scrapes the web, 2018.

[68] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015.

[69] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman. Video summarization with long short-term memory. In *European Conference on Computer Vision*, 2016.