Understanding Cultural Differences in News Videos via Image-Text Embedding

Yicun Liu, Supervisor: John R. Kender

Columbia University

Abstract. In our time, international events are widely covered around the globe, and most of them are available as news videos from different cultural perspectives. However, searching for another culture's response upon a specific event can be potentially challenging, as there usually exist language boundaries and user-preference disparities between one culture and another. In this report, we aim to capture the cultural difference by detecting the culture-specific tags: given a short video clip about a global event, we aim to retrieve the most relevant textual tags under another culture from the corresponding speech sequence. To achieve this goal, we first extract the visual and multilingual features by state-of-the-art neural models and then map them onto a joint latent embedding space through canonical correlation analysis (CCA) and its variations. To bridge the gap between different languages, we only extract the text features using native text models trained in that language. As temporal information is critical for video understanding, we propose several possible ways to embed the time stamp into the joint space. We show that for our dataset about 'AlphaGo vs. Human' news event, culture-specific tags can be significantly divergent.

Keywords: Multimodal Embedding, Video-Text Retrieval, Canonical Correlation Analysis, CCA

1 Introduction

Web-archived news videos are growing substantially on media sharing platforms, and a considerable amount of them are focusing on the coverage of international events. As the largest content sharing platform, YouTube nowadays has over 1

million hours of uploaded news video per week, sharing media coverage of domestic and international events in more than 100 languages [1, 2]. Among them, eye-catching international events like health epidemics, elections, terrorists, financial incidents, natural disaster, and sports games usually draw the most attention across the globe. Even though the event covered might overlap, news videos can entail unique information of the audience group, revealing the cultural preferences, living behaviors and user patterns in different regions of the world. Nevertheless, due to the flood of the massive media contents available online, the video selection step and interpretation step can be heavy work and exploiting the cultural difference is nearly impossible to conduct manually.

Researchers to date have long realized this dilemma, and there already exists a considerable amount of work focusing on digesting the media content and providing users with a more concise summary of the contents. Early research takes frame segments and extracts visual features from frames to conduct scene based classification [3]. Aroused by the more abstract classification task, there is some later research work focusing on automatically synthesizing text summary for short videos [4]. As speech is also the main resource for information in news videos, solely relying on visual features for news video understanding could be less effective. Multimodal models are then introduced in [5], utilizing both visual and semantic features sources for generating video description. As the comparison, cultural difference received less attention in previous research, although news videos indeed provide a fair testbed for understanding the attitudes and biases of different cultures towards the same international event. In the field of information retrieval, researchers have long been interested in summarizing similarities and differences among related documents [6]. In particular, Nakasaki [7] analyzed the text portion of multimedia pages, and conduct a cross-cultural comparison on the expressed facts and opinions. Lin [8] also observed significant cultural difference 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.

Prior to our work, Tsai made the first attempt in the detection of cultural specific tags from news videos, which is similar to regular image-text retrieval tasks [9]. Each time we use a few keyframes from an international event to conduct a

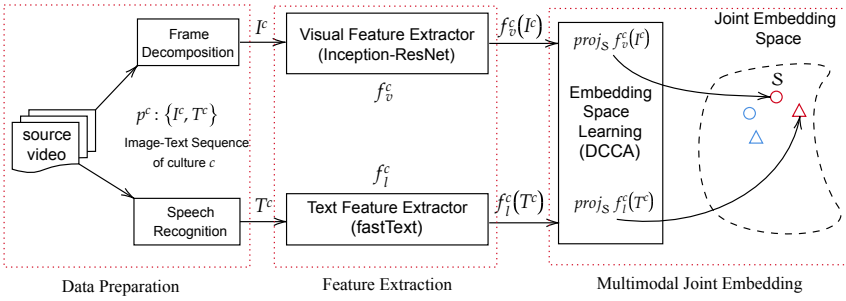


Fig. 1. Overview of the proposed pipeline. The total pipeline includes three major steps: 1. Data preparation, which includes crawling and converting video to image-text sequence pair p^c. 2. Feature extraction, which includes using pretrained expert network to extract the visual and language features. 3. Multimodal joint embedding, which aims at learning a shared space $\mathbb{S}$ where the projection of different features can be maximumly correlated.

query, and we need to find the most appropriate textual tags in the background culture. International news event appearing as videos in different cultures, only near-duplicate key-frames, and short video descriptions are considered in the embedding space. For non-English descriptions, they should be translated into English to extract text features. While this design is quite straightforward, there also exist certain limitations. First, video descriptions can be very short and less informative in the cultural analysis, like the fact that most descriptions on YouTube are under 100 words. Moreover, descriptions mainly serve as a plain summary of the news event and are usually mixed with noisy texts like the YouTube channel's advertisements. The word choices in video descriptions are more informative and objective, compared with the word choices in the editorial and criticism in the news video. Second, such an 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, one shared imagery only corresponds 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 less accurate results in the overall retrieval task. Third, using auto-translation to convert descriptions from non-English to English could introduce deficiency for comprehension.

To conquer the limitations as mentioned above, we propose to conduct cultural analysis purely on the imagery and speech available in news videos. For speech, we used speech recognition package to convert speech into text transcript and encoded it by native language embedding models. For imagery, according to the time stamps in the text transcript, we choose the key-frames which are centered by the nearby texts. 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. 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. 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 unified size.

With features vectors for imagery and text at hand, the next critical problem is how to construct an effective metric to calculate the semantic similarity between multimodal data, in our case is the text features and visual features. Before us, there are merely research in multimodal learning from an inter-cultural 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. Previously, there exists work approaching this problem by learning a linear transformation [10] or using ranking loss [11, 12]. 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 canon-

ical analysis based embeddings, which search for linear (CCA) [13] or non-linear (KCCA) [14] projections to maximize the correlation between features from different modalities. Followed by recent work in [9], 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.

The main contributions of our work can be summarized as follows:

- We reproduce the results of prior research in [9], which aims at detecting cultural-specific tags in several international events (AirAsia, Ebola, etc.). We rewrite and reorganize a majority portion of the code to best fit the current DL frameworks and website APIs. Additionally, we prepare a new dataset focusing on the feedback of 'AlphaGo vs. Human' in both the US and China.
- 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 mistranslation error in the previous setting.
- We investigate the performance of language and visual models and propose to replace them with more powerful frameworks (Inception-ResNet and fastText). Furthermore, we propose several ways for future improvement, considering alternatives of CCAs for constructing the latent embedding space and more efficient encoding of the temporal information.

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 task is 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 [15]. Later on, Karpthy first introduced CNN-based video classification on a large-scale dataset containing over 1 million videos over 487 classes [3]. The method took advantage of local spatiotemporal information in their network design, which achieved significant boost comparing with feature methods. There is further improvement in [16] 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 [17] based on sparse and incomplete tags, which is common in data from stream provider. There are also efforts in generating video summary in [18, 19], 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 [21], scene understanding [22] 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 to determine the cultural preference. In some scenarios, it is of great interest to know the relationship between images and its 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 [24]: (1) A shared embedding space for 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 MSCOCO dataset, but it is designed to operate in single image and text description instead of consecutive frames and speech sequence 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 [13]. As an improvement, a kernelizable, non-linear version of CCA (KCCA) is then introduced in [14], 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 [25], 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 that problem, a deep learning paralleled version of CCA (DCCA) is introduced in [26], 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 dataset. Additionally, the proposed DCCA framework in [26] 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) [27], 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 [28], which is a deep version of GCCA [29] 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 [10] and DeVISE [12] 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 [30, 31]. Such design enforces to 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 [32]. Similar to many classification tasks, there is also a soft assignment of the matching decision [32], 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 [33]. To train a similarity measurement network (as a branch of the two-branch network), Wang [24] 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 and Cultural Differences

Despite the exponential growth rate of the online social media contents, the query method of the large volume of multimedia data is still unequally developed. Research shows that people tend to annotate the 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 video covering an international event, US audiences tend to find such video on a media channel (e.g., CNN) on YouTube. However, in China, people tend to go to 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 habit in another country or culture, thus leads to low efficacy to conduct the multilingual query.

Previously, there also existed some research work in the multilingual query. Popescu [34] first propose 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 naturally tag their images when posting online [35]. The tags that are used initially provides 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 [36] 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 [37] 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 [37] 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 [38] revealed more cultural differences in the image tagging task: US people tend to assign the first tag to the primary object to the image, while Chinese people are more likely to assign the first tag describing the overall relationship between objects and the atmosphere of the image.

2.4 Multimodal News Event Analysis

For news event tracking and analysis, there exist a bunch of work relying on multimodal features. Early work in [39] 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 [40] 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 elements like people, places involved, and model contextual relationships between elements in the hierarchy. Jou [41] took a step further, extracting who, what, when, and where from news data. Wang [42] proposed to integrate multimodal features using the conditional random field (CRF) to segment the news stories.

Tsai [43] 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 [44] proposed a spatial-temporal multimodal TwitterLDA model which uses five Twitter cues including text, image, location, timestamp, hashtag and modeled topics as location-specific distributions. Chen [45] created a Visual-Emotional LDA (VELDA) model which associates images and texts both visually and emotionally for image retrieval. Up until now, few work [9] considers cultural differences in multimodal news analysis, and very few of them explicitly model the temporal information in their framework.

3 Data Preparation

In this section, we will briefly discuss how to prepare the dataset for the cultural analysis task. Before us, there is few publicly available dataset dedicated for cross-cultural news video analysis, and the only AirAsia/Ebola dataset in [9] has been gradually outdated. Thus, we decide to collect a new dataset based on the 'AlphaGo vs. Human' news event for our further analysis. We choose the AlphaGo event based on the consideration that it has been drawing a lot of public attention worldwide, and none of the previous datasets are tech relevant. 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.

3.1 News Videos Collection

User Patterns In the original work carried out in [9], 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 account registered on YouTube, keeping a daily routine of uploading news videos covering domestic and international events. For US audience in the cultural study, YouTube could be the largest and most convenient source for us to collect news videos.

Fig. 2. 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 includes the translated English version (from Chinese) from auto translator for comparison. It can be noticable that the translation has some lsight issue, such as tranlate Chines word 'diliutian(day 6)' to '6th', omitting the 'tian' character.

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 website, 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 contents. 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 [9] 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 [46]. 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 [46] 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 in 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' 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 that relevance of the downloaded videos.

3.2 Data Cleaning and Preprocessing

Speech Recognition With source files of the downloaded videos, the next step is to convert videos files into time series of image and text. We convert all the downloaded video file 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 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 Google's speech recognition model to convert the wav files into text files. The speech recognition service supports more than 20 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.

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 does 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 exists several popular methods for the task of Chinese word segmentation. Stanford Word Segmentor [47] supports multilingual word segmentation, including segmentation options for Chinese. There also exists Chinese word segmenter [48] making use of lexicon features. With external lexicon features, the segmenter segments more consistently and also achieves 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 [49], 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 like English, words after tokenization can be directly used.

Temporally Linked Image-Text Pair We use 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 5-minute duration, there will be more than 300 image-text pairs available.

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.

3.3 Overview of Cultural Datasets

In the previous work conducted by Tasi [9], three datasets are proposed, based on the international news event Ebola Virus, AirAsia Flight 8501 and Zika Virus. We additionally add AlphaGo vs. Human news event to study cultural difference towards advanced technology and its relationship with human beings. 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 US and 44000 keyframes of South Africa are available.

AlphaGo vs. Human There are in total 400 videos and metadata, in approximately 1:1 (China: US) ratio, in a date range from 03/09/16 to 03/23/17. In total 1000 keyframes of the US and 600 keyframes of China are available.

4 Methods

4.1 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.

4.2 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 [9, 24, 43, 50] 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 pretrained on 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 are 2048 and 512.

In order to obtain better quality visual features specific for our task, we made two modifications: 1. We use more advanced classification architecture Inception-ResNet proposed in [52] 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 $\times$ 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 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 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 main language a, and then we can use the text embedding model for language a to handle all the scenarios. In previous work of Tsai [9], 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, error in translation can never be avoidable. There are always difficult words for a language translation model, where it cannot find direct word-to-word translation but using word-to-phrase instead. In order 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 [53, 54] 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 [55]. The fastText model works surprisingly well, which is close to the other models. The language features generated by fastText is the feature vector of 256 entries. For a text sequence containing n words, we flatten n feature vectors into the averaged feature vector.

4.3 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) [13]. 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:

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

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 [56].

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) [26] 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 multivariate optimization problem has no closed form solution, but the optimal solution can be approximated by 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 to first 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:

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

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.

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 modals, generalized canonical correlation analysis (DCCA) is proposed in [29], with its deep learning version DGCCA later proposed in [28]. The objective of GCCA is to find a shared representation 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 datapoints (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 eigen-decomposition 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.

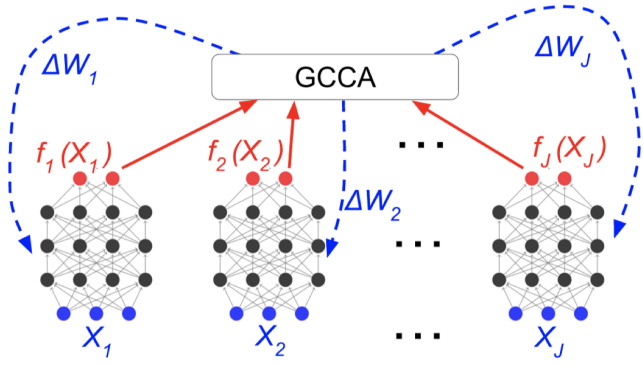


Fig. 3. Schematic Illustration of DGCCA, which is originally proposed in [28]. The f_j denotes the j-th (branch) of the network to project j-th input data entity X_j into the GCCA joint embedding space. Our objective function aims at maximize the total sum of the inter-input correlation, which calculates the correlation between each of pair of the two inputs.

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).

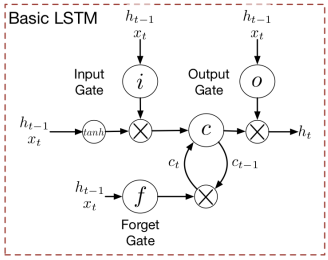
Three-Way Embeddings Assume the near-duplicate image pairs can be viewed as equal, the previous near-duplicate image-text pair $\{(I^m, T^m), (I^n, T^n)\}$ can be simplified to $\{I^{mn}, T^m, T^n\}$, where $I^{mn} = I^m = I^n$. For the objective function of GCCA, now the data matrix of X involves three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

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

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.

4.4 Temporal Relationship Reasoning

Why Encode Temporal Dependency In the previous work [9] 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



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

Fig. 4. Illustration of LSTM unit from [18]. 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.

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 [18] and visual question answering (VQA) [57]. 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 usu-

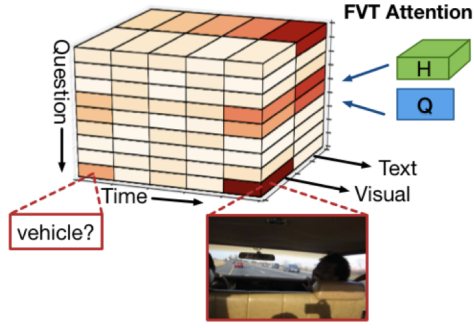


Fig. 5. Illustration of the data matrix H and query matrix Q to encode sequence of multimodal information according to the temporal order. This time-based sequence encoding method is originally used for VQA task in [57]. Slightly differnt from the (question) text-image query in QVA, the query matrix Q in our task is directly the image features of the near-duplicate frame/sequence.

ally 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 [57] 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 Experiments

5.1 Implementation Details

In the previous work [9], we have three datasets: Ebola Virus, Zika Virus, Air Asia Flight 8501. We collect additional dataset based on the feedback of the US and Chinese people on 'AlphaGo vs. Human' news event. For speech recognition, we use the pretrianed 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 extract image-text sequence pair using a 5 second sliding window, the sampling rate for keyframes is set to be 2 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 outputed by the last two fully connected layer in Inception-ResNet model and load existing weight pretrained 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 [58]

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.

	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

Table 1. Performance of intra-culture image-text queries. The Ebola, AirAsia, Zika dataset is bsaed on prior work in [9], and AlphaGo vs. Human is our new dataset.

5.2 Experimental Results

For each news event, we generate two culture-specific image-text embedding space for two different cultures. For AirAsia Flight 8501 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, 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. Further ablation study needs to be carried to determine which proposed component leads to significant accuracy boost.

6 Conclusion and Future Work

In this report, 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 [9], 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 image-text sequence. We use implicit encoding method to ensure the image and language features are chronologically related in our analysis. Furthermore, we rewrite most of the code in [9] 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. Conduct an 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 other methods covered in the related work (ranking-based, metric-learning based, classification based, etc.) to learn a better joint embedding space learning. 4. Enforce strict constraints on intra-sequence chronological order or use the explicit encoding of the temporal information (as proposed in section 4.4). 5. Extend our dataset to more news events and culture groups. Because there are 150+ multilingual native language embedding models available, we may able to gather more cross-cultural feedbacks based on other languages.

References

1. Burgess, J., Green, J.: YouTube: Online video and participatory culture. John Wiley & Sons (2018)
2. Xie, L., Natsev, A., Kender, J.R., Hill, M., Smith, J.R.: Visual memes in social media: tracking real-world news in youtube videos. In: Proceedings of the 19th ACM international conference on Multimedia. (2011)
3. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. (2014) 1725–1732
4. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K.: Sequence to sequence-video to text. In: Proceedings of the IEEE international conference on computer vision. (2015)
5. Ramanishka, V., Das, A., Park, D.H., Venugopalan, S., Hendricks, L.A., Rohrbach, M., Saenko, K.: Multimodal video description. In: Proceedings of the 2016 ACM on Multimedia Conference. (2016)

6. Mani, I., Bloedorn, E.: Summarizing similarities and differences among related documents. Information Retrieval (1999)
7. Nakasaki, H., Kawaba, M., Utsuro, T., Fukuhara, T.: Mining cross-lingual/cross-cultural differences in concerns and opinions in blogs. In: International Conference on Computer Processing of Oriental Languages. (2009)
8. Lin, B.Y., Xu, F.F., Zhu, K., Hwang, S.w.: Mining cross-cultural differences and similarities in social media. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. (2018)
9. Tsai, C.Y., Kender, J.R.: 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)
10. Weston, J., Bengio, S., Usunier, N.: Wsabie: Scaling up to large vocabulary image annotation. In: IJCAI. (2011)
11. Weston, J., Bengio, S., Usunier, N.: Large scale image annotation: learning to rank with joint word-image embeddings. Machine learning **81**(1) (2010) 21–35
12. Frome, A., Corrado, G.S., Shlens, J., Bengio, S., Dean, J., Mikolov, T., et al.: Devise: A deep visual-semantic embedding model. In: Advances in neural information processing systems. (2013) 2121–2129
13. Hardoon, D.R., Szedmak, S., Shawe-Taylor, J.: Canonical correlation analysis: An overview with application to learning methods. Neural computation **16**(12) (2004) 2639–2664
14. Lai, P.L., Fyfe, C.: Kernel and nonlinear canonical correlation analysis. International Journal of Neural Systems **10**(05) (2000) 365–377
15. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. (2008)
16. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015)
17. Wang, J., Zhu, X., Gong, S.: Video semantic clustering with sparse and incomplete tags. In: AAAI. (2016)
18. Zhang, K., Chao, W.L., Sha, F., Grauman, K.: Video summarization with long short-term memory. In: European Conference on Computer Vision. (2016)
19. Mahasseni, B., Lam, M., Todorovic, S.: Unsupervised video summarization with adversarial lstm networks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)

20. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation **9**(8) (1997) 1735–1780
21. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015)
22. Song, S., Lichtenberg, S.P., Xiao, J.: Sun rgb-d: A rgb-d scene understanding benchmark suite. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2015)
23. Jiang, Y.G., Ye, G., Chang, S.F., Ellis, D., Loui, A.C.: 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. Wang, L., Li, Y., Huang, J., Lazebnik, S.: Learning two-branch neural networks for image-text matching tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence (2018)
25. Klein, B., Lev, G., Sadeh, G., Wolf, L.: Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. arXiv preprint arXiv:1411.7399 (2014)
26. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning. (2013) 1247–1255
27. Wang, W., Yan, X., Lee, H., Livescu, K.: Deep variational canonical correlation analysis. arXiv preprint arXiv:1610.03454 (2016)
28. Benton, A., Khayrallah, H., Gujral, B., Reisinger, D.A., Zhang, S., Arora, R.: Deep generalized canonical correlation analysis. arXiv preprint arXiv:1702.02519 (2017)
29. Coelho, C.A.: Generalized Canonical Analysis. PhD thesis (1992)
30. Karpathy, A., Joulin, A., Fei-Fei, L.F.: Deep fragment embeddings for bidirectional image sentence mapping. In: Advances in neural information processing systems. (2014) 1889–1897
31. Socher, R., Karpathy, A., Le, Q.V., Manning, C.D., Ng, A.Y.: Grounded compositional semantics for finding and describing images with sentences. Transactions of the Association of Computational Linguistics (2014)
32. Jabri, A., Joulin, A., van der Maaten, L.: Revisiting visual question answering baselines. In: European conference on computer vision. (2016)
33. Rohrbach, A., Rohrbach, M., Hu, R., Darrell, T., Schiele, B.: Grounding of textual phrases in images by reconstruction. In: European Conference on Computer Vision. (2016)

34. Popescu, A., Kanellos, I.: 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)
35. Bergsma, S., Van Durme, B.: Learning bilingual lexicons using the visual similarity of labeled web images. In: IJCAI Proceedings-International Joint Conference on Artificial Intelligence. (2011)
36. Bao, P., Hecht, B., Carton, S., Quaderi, M., Horn, M., Gergle, D.: Omnipedia: bridging the wikipedia language gap. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2012) 1075–1084
37. Clough, P., Müller, H., Deselaers, T., Grubinger, M., Lehmann, T.M., Jensen, J., Hersh, W.: The clef 2005 cross–language image retrieval track. In: Workshop of the Cross-Language Evaluation Forum for European Languages, Springer (2005)
38. Dong, W., Fu, W.T.: Cultural difference in image tagging. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, ACM (2010) 981–984
39. Wu, X., Ngo, C.W., Hauptmann, A.G.: Multimodal news story clustering with pairwise visual near-duplicate constraint. IEEE Transactions on Multimedia (2008)
40. Li, W., Joo, J., Qi, H., Zhu, S.C.: Joint image-text news topic detection and tracking by multimodal topic and-or graph. IEEE Trans. Multimedia (2017)
41. Jou, B., Li, H., Ellis, J.G., Morozoff-Abegauz, D., Chang, S.F.: 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)
42. Wang, X., Xie, L., Lu, M., Ma, B., Chng, E.S., Li, H.: Broadcast news story segmentation using conditional random fields and multimodal features. IEICE TRANSACTIONS on Information and Systems (2012)
43. Tsai, C.Y., Xu, R., Colgan, R.E., Kender, J.R.: 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)
44. Cai, H., Yang, Y., Li, X., Huang, Z.: What are popular: exploring twitter features for event detection, tracking and visualization. In: Proceedings of the 23rd ACM international conference on Multimedia, ACM (2015)
45. Chen, T., SalahEldeen, H.M., He, X., Kan, M.Y., Lu, D.: Velda: Relating an image tweet's text and images. In: AAAI. (2015)
46. Yao, M.: You-get: Dumb downloader that scrapes the web (2018)
47. Huihsin, T., Chang, P., Andrew, G., Jurafsky, D., Manning, C.: A conditional random field word segmenter. In: Fourth SIGHAN Workshop. (2005)

48. Chang, P.C., Galley, M., Manning, C.D.: Optimizing chinese word segmentation for machine translation performance. In: Proceedings of the third workshop on statistical machine translation, Association for Computational Linguistics (2008) 224–232
49. Cai, D., Zhao, H., Zhang, Z., Xin, Y., Wu, Y., Huang, F.: 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), Vancouver, Canada, Association for Computational Linguistics (July 2017) 608–615
50. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016)
51. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
52. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. (2017)
53. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018). (2018)
54. Mikolov, T., Grave, E., Bojanowski, P., Puhrsch, C., Joulin, A.: Advances in pre-training distributed word representations. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018). (2018)
55. Li, S., Zhao, Z., Hu, R., Li, W., Liu, T., Du, X.: 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), Association for Computational Linguistics (2018) 138–143
56. Johnson, R.A., Wichern, D.W.: Multivariate analysis. Encyclopedia of Statistical Sciences **8** (2004)
57. Liang, J., Jiang, L., Cao, L., Li, L.J., Hauptmann, A.: Focal visual-text attention for visual question answering. In: Conference on Computer Vision and Pattern Recognition (CVPR). (2018)
58. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)