Semantic Image Profiling for Historic Events: Linking Images to Phrases

Jia Chen[1], Qin Jin[2]* , Yifan Xiong[2]
[1]School of Computer Science, Carneige Mellon University
jiac@cs.cmu.edu
[2]School of information, Renmin University of China
{qjin, xiongyf}@ruc.edu.cn

ABSTRACT

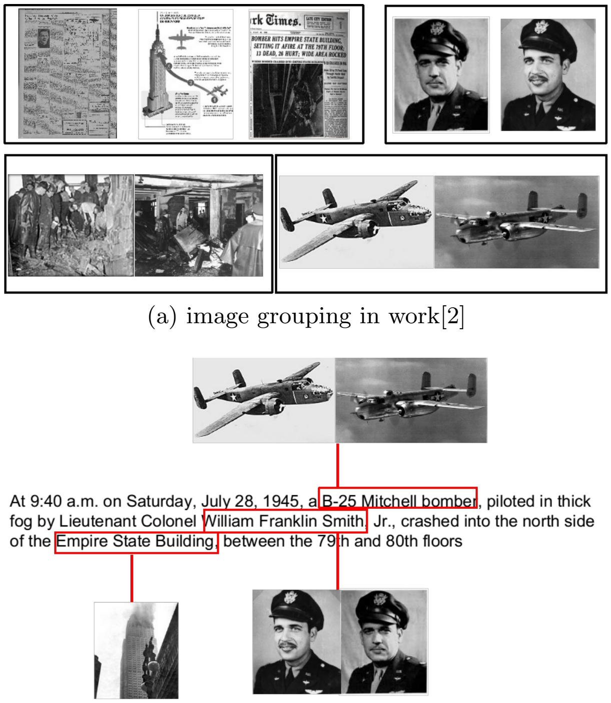
Automatically generating image profiles for historic events is desired for history knowledge preservation and curation. However, a simple profile with groups of related images lacks explicit semantic information, such as which images correspond to which aspects of the event. In this paper, we propose to add explicit semantic information to image profiling by linking images in the profile with related phrases in the event description. We measure the relevance of an image-phrase pair via a real-valued matching score. We exploit instance-wise ranking loss function to learn the matching score and we deal with two challenges: 1) how to automatically generate labeled positive data: we leverage out-of-domain labeled datasets to generate pseudo positive in-domain labels and propose a new algorithm (WIL4PPL) to robustly learn the model from the noisy pseudo positive labels; 2) how to automatically generate negative data: we propose a negative set generation algorithm to guide the model in learning which phrases and images to distinguish. We compare our model to three baselines and conduct detailed analysis and case studies to verify the quality of learnt semantic information. The extensive experiment results show the effectiveness of our proposed algorithms which significantly outperform the baselines.

Keywords

Historic Events; Image Profiling; Image-Phrase Linking

1. INTRODUCTION

Knowledge of history is precious and is recorded in both text and image media. There are many organized text resources for historic events online but relatively few organized image resources. To fill the gap, in [2] we proposed to automatically construct image profiles for historic events on the fly. The input of the task is one sentence event description containing 4Ws and the output is relevant image groups covering different aspects of the event.



(a) image grouping in work[2]

At 9:40 a.m. on Saturday, July 28, 1945, a B-25 Mitchell bomber, piloted in thick fog by Lieutenant Colonel William Franklin Smith, Jr., crashed into the north side of the Empire State Building, between the 79th and 80th floors

(b) image and phrase links

Figure 1: Illustration of linking images and phrases to enhance the semantic of image profiling for historic events

In work[2] we group images according to genre and visual content as shown in Figure 1(a). The genre groups include document/map, painting and photo. The visual content groups are consistent not only on visual appearance but also on semantic meaning. Both types of groups are useful in organizing the image profiles for the historic events. But such group based image profiles lack explicit semantic information. The genre grouping only focuses on genre, which does not carry explicit semantic information. The semantic information provided by content groups is implicit. That is, users have to figure out the meaning of each content group by themselves.

In this paper, we propose to improve image profiling by

*Qin Jin is the corresponding author

MM '16, October 15-19, 2016, Amsterdam, Netherlands
© 2016 ACM. ISBN 978-1-4503-3603-1/16/10…$15.00
DOI: http://dx.doi.org/10.1145/2964284.2964306

adding explicit semantic information automatically. To be specific, we add explicit semantic information by linking images to the corresponding noun phrases in the event description automatically. Figure 1 illustrates the links for related image-phrase pairs, contrasted with the grouping image profile proposed in [2]. Users can easily get the correspondence between images and text descriptions. For example, we see that the portrait photos are linked to phrase "Lieutenant Colonel Willima Ranklin Smith" and the plane images are linked to "B-25 Mitchell bomber". We also see that not all images needs to be linked to certain phrase in the event description, e.g. the news paper report photo. These images are relevant to the event but not relevant to any particular phrase. This suggests that image-phrase links and image groups complement each other for a better organized image profiling.

The automatic linking problem is a special version of general image annotation problem. It's specialized in the following two aspects. First, it targets on specific noun phrase (usually entities) rather than general words. Second, the event context is given, i.e. the event sentence and related images. Linking with specific noun phrase is annotation on a much finer scale, which is far more difficult than general words if the event context information is not given. But in our case, we have event context, which provides strong prior knowledge and narrows down the candidate scope.

Instead of directly formalize automatic linking as a classification problem, we model it as a real score and binarize the output by thresholding. To be specific, we compute a real-valued matching score between the image and the phrase. Under such formalization, we could exploit rank loss function to learn the matching score, which is shown to be more robust than $0-1$ loss function on relevance modeling tasks[17].

To learn the matching score through rank loss, we need both labeled positive data and automatically generated negative data. However, it is challenging to obtain both data in historic event domain.

Labeled positive data challenge: manual labeling the relevance of image-phrase pairs for historic events requires very strong background knowledge. The cost of collecting such labels is much higher than selecting a familiar object or scene concept for an image. We propose to leverage out-of-domain labeled dataset to automatically generate pseudo positive in-domain labels. However, the generated pseudo positive labels are not perfect. We use weighted instance learning to tackle the issue.

Negative pair generation challenge: there are two kinds of negative pairs for a positive image-phrase pair (i, w): the first kind of pairs share the same image i but with different phrase w and the second kind of pairs share the same phrase w but with different image i. These two kinds are symmetric and let's consider the first kind for the ease of explanation. Constructing a negative pair by simply picking a phrase that is different from the phrase in the groundtruth pair won't work in historic event domain since we may pick a phrase that is semantically similar to the phrase in the groundtruth. For example, the phrase "aircraft carrier USS Enterprise" is semantically similar to "destroyers USS Maddox". When we link images to phrases, we only consider candidate phrases within the same event. Thus we don't need to distinguish between these two phrases since they rarely appear in the text description of the same event. Such case exists widely in historic event domain. We design a negative pair generation algorithm that generates the proper negative pairs to train the model.

Up to now, we are still vague about the relevance of image-phrase pair. The noun phrase could describe a concrete object such as "B-25 bomber". In this case, it is relevant with an image of B-25. The noun phrase could also describe a very abstract concept such as "world record". Together with an image of an athlete breaking the world record, this image-phrase pair could also be considered as relevant. It is quite challenging to give a clear and general definition on the relevance of image-phrase pair. We therefore derive a case-based definition for image-phrase pair relevance. Under each case, we give a separate definition of relevance so that the groundtruth labeling used in evaluation is both consistent and reproducible.

We make the following contributions:
1. we add explicit semantic information to the image profiling by linking images with phrases.
2. we give a definition of relevance on image-phrase pairs which not only covers the diverse cases in historic event domain but also can be directly transformed to practical reproducible labeling guidance.
3. we propose to automatically generate pseudo positive in-domain data from out-of-domain datasets and leverage these pseudo data to learn the parameters.
4. we propose a negative pair generation algorithm to guide the model to learn what to distinguish in historic event domain.

The remainder of the paper is organized as follows. Section 2 presents related work. Section 3 introduces the definition of image-phrase relevance for historic events. Section 4 formalizes the image-phrase relevance problem as a problem of learning matching score with pseudo positive labels. Section 5 presents the solution to the problem. Section 6 evaluates the quality of image-phrase links and provides some detailed experiment analysis and section 7 draws some conclusions.

2. BACKGROUND AND RELATED WORK

Art, culture and history related multimedia data have received more attentions from the research community[13] [11][8][2][3] in recent years. Authors in work[13] collect an image dataset of artworks exhibited in the Rijksmuseum in Amsterdam and define challenges for visual classification and content-based retrieval of artistic content. Their focus is on collecting a public artwork dataset and designing challenging tasks that could become useful tools for museums. Authors in work[11] design a tool named eHeritage, including a creator module and a manipulator module, to preserve the precious traditional heritage Chinese shadow puppetry. Their focus is on building multimedia tool to help users generating multimedia data and animation in their user cases. Authors in work[8] propose a new language-dependent method for automatic discovery of adjective-noun structures and apply the pipeline on a social multimedia platform for creation of a large-scale multilingual visual sentiment concept ontology. Their focus includes not only collecting a public cross-lingual dataset for research community but also designing multimedia tools to help analyzing cultural difference of visual sentiments. In work[2], we propose to automatically construct an image profile given a one sentence description of the historic event so that a multimedia history event dataset could be automatically collected to preserve precious history knowledge.

All these works share two key elements: multimedia dataset and multimedia tool. These two key elements are interdependent. On the one hand, datasets in these domains are extremely valuable and precious to collect. Once provided a dataset of descent scale, researchers can train models that could finally be transformed to useful tools. On the other hand, once a multimedia extraction tool is built, it can be used to ease the manual labor effort in collecting multimedia data for art/culture/history preservation. This work shares these two key elements with all the above works. Our task is to link images with phrases to enhance the semantic information in image profiling of historic events, which could be finally transformed to a useful multimedia tool for history knowledge preservation. To build such a tool, we need labeled in-domain data to train the model. However, the cost of labeling image-phrase relevance in historic event domain is so high that training model after collecting enough labeled data is not a reasonable choice. We propose to automatically generate pseudo in-domain positive labels from out-of-domain datasets and design an algorithm WIL4PPL to learn the model from these noisy positive labels. Such interdependency between data and tools widely exists in multimedia problems of art, culture and history domain and our solution can be applicable to these problems.

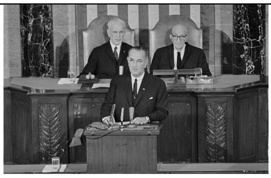
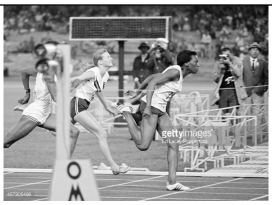
There have been many research works[1][4][19][6] on the image-word relevance problem in the form of image annotation. It is modeled as a matching problem. Authors in work[4] leverage machine translation technique to learn the matching between image and words. Authors in work[1] propose a graphical model to learn the matching between image and words. Authors in work[19] propose a multimodal embedding space to learn the matching between image and words. Authors in work[6] learn embedding space upon deep learning features of images and words. We also model the image-phrase linking problem as a matching problem. Different from these works, we don't have in-domain labeled training data and have to leverage pseudo positive labels based on out-of-domain dataset to train the model, which is very challenging. Our focus is on leveraging the noisy pseudo positive labels to train a matching model.

3. LINKING AN IMAGE TO A PHRASE

3.1 Preliminaries on image profiling task

In the historic event image profiling task[2], the input is a text description of an event e and a set of candidate images $\mathbf{I}_e$. The text description is usually composed of only one sentence. From the sentence, the phrase set $\mathbf{W}_e$ is extracted using standard NLP techniques[12]. Phrases vary from a celebrity name (e.g. "Queen Elizabeth II") to a mountain name (e.g. "Mount Everest"), from detailed description (e.g. "the ten second barrier") to general event name (e.g. "Montgomery Bus Boycott"), from nouns (e.g. "volcano") to verbs (e.g. "erupt"). The content of images varies from object to scene. This indicates that even within each modality (text or image), its type already varies. When considering relevance on the joint modality, for example linking images to texts, the situation becomes much more complex due to combination explosion. Furthermore, it is possible that some images may be irrelevant to any particular phrase in the event description while some images may be relevant to several phrases in the event description. Thus we need to give a clear definition of relevance between an image and a phrase before we try to formulate the problem.

Table 1: Two archetypes of relevance between an image and a phrase: the phrase is underlined in the event description

event description	image
United States President Lyndon B. Johnson proclaims his "Great Society" during his State of the Union address.	
Jim Hines of the USA becomes the first man ever to break the ten second barrier in the 100 metres Olympic final at Mexico City with a time of 9.95 sec. He would be the only man to do so until 1983.	

3.2 Definition of relevance between an image and a phrase

Consider the following two archetypes in Table 1. In the first archetype, the image shows a person standing in the front and the phrase is actually the name of the person in the image. This image-phrase pair is obviously relevant since the phrase describes the object in the image. In the second archetype, the image shows runners at the finish line and the phrase is abstract but contains important background information. This image-phrase pair is also relevant since the phrase could be used to caption this image in a news report. These are the two major archetypes of image-phrase relevance. On the text side, it covers concrete nouns and abstract nouns. On the image side, it has no restriction on image content.

We give the following definitions of image-phrase relevance archetypes:

content relevant an image-phrase pair is content relevant if the phrase is a noun phrase about concrete object or scene and the image contains the concrete object or scene.

background relevant an image-phrase pair is background relevant if the phrase is a meaningful abstract noun phrase in the event description that can be used to caption the image.

irrelevant the rest cases are considered as irrelevant.

4. PROBLEM FORMULATION

Given event e's image set $\mathbf{I}_e$ and phrase set $\mathbf{W}_e$, our problem is to classify an image-phrase pair (i, w) as a relevant pair or not, where $i \in \mathbf{I}_e$ and $w \in \mathbf{W}_e$. Instead of directly modeling relevance as binary 0-1 value, we propose to model the relevance of image-phrase pair (i, w) as a real valued matching score and thresholding it to get the binary classification output. To be specific, we measure the matching score between an image i and a phrase w by $s_{iw} = \mathbf{f}(i)^T\mathbf{M}\mathbf{f}(w)$, where $\mathbf{M}$ is a matching matrix, $\mathbf{f}(i)$ is the feature vector of image i and $\mathbf{f}(w)$ is the feature vector of word w. If the training data is not sufficient, we could mitigate the over-fitting problem by adding regularization on matching matrix $\mathbf{M}$'s rank r. To be specific, we add the regularization by decomposing $\mathbf{M}$ to $\mathbf{U}\mathbf{V}^T$, where the second dimension of $\mathbf{U}$ and $\mathbf{V}$ is r. Then, we learn $\mathbf{U}$ and $\mathbf{V}$ instead of $\mathbf{M}$.

4.1 Instance-wise ranking loss function

For a groundtruth image-phrase pair (i, w), the matching score s_{iw} should be higher than the matching score between the same image i and an unrelated phrase w' with a margin Δ. Similarly, it should also be higher than the matching score between the same phrase w_i and an unrelated image i' with a margin Δ. Such relation is summarized in the following equations:

$$s_{iw} - s_{iw'} \geq \Delta \qquad (1)$$

$$s_{iw} - s_{i'w} \geq \Delta \qquad (2)$$

Based on such relation, we introduce instance-wise ranking image loss $L_{iw}^{image}(i, w)$ and instance-wise ranking phrase loss $L_{iw}^{word}(i, w)$ for each groundtruth pair (i, w).

$$L_{iw}^{image}(\mathbf{M}) = \sum_{i' \in \bar{\mathbf{I}}_w} max(0, \Delta - s_{iw} + s_{i'w}) \qquad (3)$$

$$L_{iw}^{word}(\mathbf{M}) = \sum_{w' \in \overline{\mathbf{W}}_i} max(0, \Delta - s_{iw} + s_{iw'}) \qquad (4)$$

where $\bar{\mathbf{I}}_w$ denotes the negative image set of phrase w and $\overline{\mathbf{W}}_i$ denotes the negative phrase set of image i. Note that an image $i' \neq i$ doesn't mean that it is a negative image of phrase w. The same holds to image i when $w' \neq w$. Thus we have the following relations:

$$\bar{\mathbf{I}}_w \subseteq \mathbf{I} \setminus i \qquad (5)$$

$$\overline{\mathbf{W}}_i \subseteq \mathbf{W} \setminus w \qquad (6)$$

We leave the construction of $\bar{\mathbf{I}}_w$ and $\overline{\mathbf{W}}_i$ for a positive instance (i, w) to the solution section.

Table 2: Notations

$\mathbf{I}$	image set
$\mathbf{W}$	phrase set
$\mathbf{I}_e$	image set of event e
$\mathbf{W}_e$	phrase set of event e
$\bar{\mathbf{I}}_w$	negative image set of word w
$\overline{\mathbf{W}}_i$	negative word set of image i
$\mathbf{C}$	out-of-domain concept set
$\mathbf{T}$	all-domain text set
i	image index
w	phrase index
e	event index
c	concept index
t	text index
$\mathbf{f}(c), \mathbf{f}(w), \mathbf{f}(t)$	feature vector of concept c, phrase w and text t
$\mathbf{f}(i)$	feature vector of image i
$\mathbf{M}$	matching matrix between image feature and phrase feature, which has a low dimension decomposition: $\mathbf{M} = \mathbf{U}\mathbf{V}^T$

4.2 Weighted instance loss function for pseudo positive labels

As our task is to extract semantic links in historic event domain from scratch, we don't have any available in-domain labels yet. Manually labeling the historic event domain images requires volunteers with sufficient background knowledge. Such precious manual labeling is not scalable to collect enough data labels. Thus, we leverage the automatically generated pseudo positive labels to learn the matching score. We leave the details of the pseudo label generation to the solution section. A major issue of the automatically generated pseudo positive labels is that their confidence varies a lot. If we just preserve the top high confidence positive labels, the scale of positive data is not large enough to cover different image-phrase pairs. If we include all the pseudo positive labels and treat them equally, the training data will be very noisy. To tackle this issue, we propose to use weighted instance loss function on the pseudo positive labels:

$$L^{image}(\mathbf{M}; \alpha) = \sum_{iw} \alpha_{iw} L_{iw}^{image}(\mathbf{M}) \qquad (7)$$

$$L^{word}(\mathbf{M}; \alpha) = \sum_{iw} \alpha_{iw} L_{iw}^{word}(\mathbf{M}) \qquad (8)$$

where α_{iw} is the confidence score associated with the psuedo positive instance (i, w). We get the final loss function by convex combination of $L^{image}(i, w)$ and $L^{word}(i, w)$:

$$L(\mathbf{M}; \alpha, \theta) = \theta L^{image}(\mathbf{M}; \alpha) + (1 - \theta) L^{word}(\mathbf{M}; \alpha) \qquad (9)$$

This loss function can be minimized through standard stochastic gradient descent (SGD). The notations are summarized in Table 2.

5. SOLUTION

Our solution consists of three components: pseudo positive label generation, negative set generation and weighted instance learning.

5.1 Pseudo positive label generation

To generate pseudo positive labels for the historic event domain, we leverage out-of-domain labeled datasets to train and predict without additional manual labeling effort involved. Though we don't have an identical labeled dataset that covers both image ($\mathbf{I}$) and text modalities ($\mathbf{C}$) in the historic event domain, we have various public datasets that cover different modalities of different domains. For example we have image-text modality datasets that cover image in similar domains $\mathbf{I}$ but text in other different domains $\mathbf{C}$ such as ImageNet [16] and SUN [20]. They could be used to estimate the probability $p(c|i)$. We also have pure text modality datasets covering almost all domains $\mathbf{T}$, such as google news [14] which could be leveraged to estimate the probability $p(w|c)$. By properly arranging the domain sequence of public labeled datasets, we get the following Markov chain

$\mathbf{I} \rightarrow \mathbf{C} \rightarrow \mathbf{W}$:

$$p(w|i) = \sum_c p(w|c)p(c|i) \tag{10}$$

Here $p(c|i)$ is concept c's conditional probability on the given image i and $p(w|c)$ is phrase w's conditional probability on the given concept c. This Markov chain first generates out-of-domain concepts based on the given images, and then generates the in-domain phrase based on the out-of-domain concepts. Based on eq (10), we generate pseudo phrase labels for images by:

$$\hat{w} = arg \max_w p(w|i) \tag{11}$$

First, we select the suitable datasets for each edge in the Markov chain. For the $\mathbf{I} \rightarrow \mathbf{C}$ edge, we need datasets that cover most image domains. The major content of historic images is object and scene. For object content, ImageNet[16] is one of the most comprehensive public datasets. And for scene content, SUN[20] is one of the most comprehensive public datasets. The text of both datasets comes from WordNet [15], whose majorities are concept words from dictionary. The text modality domain is very different from history event domain, whose majorities are proper nouns. By utilizing pre-trained classifiers on these two image to text datasets, we are able to predict out-of-domain object and scene concepts $\mathbf{C}$ from images. For the $\mathbf{C} \rightarrow \mathbf{W}$ edge, we leverage a pure text dataset, google news [14]. It covers most of the words in almost all text domains $\mathbf{T}$. For phrases in $\mathbf{T}/(\mathbf{C} \cup \mathbf{W})$, we cover them by word composition from $\mathbf{T}$.

To estimate $p(c|i)$, we use the softmax output of the trained image concept classifier:

$$p(c|i) = softmax(\alpha_c^T \mathbf{f}(i)) \tag{12}$$

where α is the weight of the last full connect layer of the concept classifier.

To estimate $p(w|c)$, we learn vector representation of phrases from google news. We don't directly estimate $p(w|c)$ from this corpus because many phrases in the historic events are compound phrases such as "the United States Navy aircraft carrier USS Enterprise", which is a composite of "the United States Navy", "aircraft carrier", "USS Enterprise". There's no direct correpondence in the google news and we need to compound them from google news. Vector representation offers us the flexibility to composite new phrases from learned words. The vector representation is learned by maximizing the probability of the next word t given the previous words t':

$$p(t|t') = softmax(\mathbf{f}(t)^T \mathbf{f}(t')) \tag{13}$$

Once the vector representation of $\mathbf{T}$ is learned, we get the vector representation of a phrase in $\mathbf{W}$ by composition:

$$\mathbf{f}(w) = \sum_{t \in w} \mathbf{f}(t) \tag{14}$$

Given eq (14) and eq (13), we calculate $p(w|c)$ by:

$$p(w|c) = softmax(\mathbf{f}(w)^T \mathbf{f}(c)) \tag{15}$$

$$= \frac{exp(\mathbf{f}(w)^T \mathbf{f}(c))}{\sum_{w'} exp(\mathbf{f}(w')^T \mathbf{f}(c))} \tag{16}$$

Instead of enumerating the condition probabilities of the entire phrase set w, which could be extremely large, we restrict to calculate the condition probability only on the phrase set within event e of the image i for both eq (10) and (15). It is possible that phrases from other events are suitable for the image, but neglecting such phrases has neglectable impact on the recall of pseudo positive labels. In summary, a pseudo positive label is generated by:

$$\tilde{p}(w|c) = \frac{exp(\mathbf{f}(w)^T \mathbf{f}(c))}{\sum_{w' \in \mathbf{W}_e} exp(\mathbf{f}(w')^T \mathbf{f}(c))} \tag{17}$$

$$\tilde{p}(w|i) = \sum_c \tilde{p}(w|c)p(c|i) \tag{18}$$

$$\hat{w} = arg \max_{w \in \mathbf{W}_e} \tilde{p}(w|i) \tag{19}$$

By setting $\alpha_{wi} = \tilde{p}(w|i)$, we get the confidence score associated with the pseudo positive label.

5.2 Negative set generation

We construct negative phrase set $\overline{\mathbf{W}}_i$ for image i of positive image-phrase pair $(i,\, w)$ by excluding phrases w' that are similar to w. It is possible that some of these w' are hard negatives. But we don't need to introduce such hard negatives into the negative set since in the image profiling task, the possible phrase candidates are limited within an event. For example, we don't need to distinguish between "aircraft carrier USS Enterprise" and "destroyers USS Maddox" as they rarely appear in the same event. If we include such phrase w' in the $\overline{\mathbf{W}}_i$, the learning process has to optimize the parameter by distinguishing s_{iw} away from $s_{iw'}$ according to ranking word loss L_{iw}^{word}, which leads to contradictory condition. The same holds for negative image set $\overline{\mathbf{I}}_w$ for phrase w. We construct it by excluding images that are similar to i.

Now the negative set generation problem is transformed to the similar image/phrase searching problem. That is, we search similar images for each image and similar phrases for each phrase in the positive (image, phrase) set. We can approximate all these search operations by one clustering operation if we consider points in the same cluster are similar. When new positive pseudo labels are added, we could update the negative set by just assigning the image and phrase in the new pseudo label to the current image cluster and phrase cluster. We select affinity propagation clustering[5] to control the consistency of points within each cluster through the preference value.

5.3 Weighted instance learning

With the pseudo positive labels and negative set generation, we learn the parameters by mini-batch stochastic gradient descent. In each mini-batch, a record is a quadruple (i, w, i', w'), where i, w are the image and word in the positive image-phrase pair (i, w), $i' \in \overline{\mathbf{I}}_w$ and $w' \in \overline{\mathbf{W}}_i$. That is, each quadruple corresponds to one positive pair sample (i, w) and two negative pair samples, (i', w) and (i, w'). In each epoch, we traverse all positive pair samples in a shuffled order and the negative pairs are uniformly sampled with replacement from the generated negative set. That is, the model is fed with different negative pairs in each epoch but the same positive pairs in each epoch. Thus, the epoch number is proportional to the negative to postive ratio. The entire algorithm is summarized in Algorithm 1. Note that $\mathbf{U}$ and $\mathbf{V}$ are updated by weighted combination of gradients from each record. The weights correspond to the confidences of positive pseudo labels.

Algorithm 1: weighted instance learning for pseudo positive labels

```
input  : negative to positive ratio $r_{np}$, rank $r$ of the
         matching matrix, learning rate $\mu$
output: $\mathbf{U}$, $\mathbf{V}$
initilize $\mathbf{U}, \mathbf{V}$ from random distribution or from a
pretrained matching model;
for $e \leftarrow$ to $\frac{r_{np}}{2}$ do
    sample negative pair $(i', w), (i, w')$ for each positive
    $(i, w)$;
    foreach mini-batch $\mathbf{B}$, where $b \in \mathbf{B}$ are indices of
    postive pair labels do
        $\mathbf{U} \leftarrow \mathbf{U} + \mu \sum_{b \in \mathbf{B}} \alpha_b \frac{\partial L}{\partial \mathbf{U}}$;
        $\mathbf{V} \leftarrow \mathbf{U} + \mu \sum_{b \in \mathbf{B}} \alpha_b \frac{\partial L}{\partial \mathbf{V}}$
```

6. EXPERIMENT

This section is organized as follows. First we describe the dataset, groundtruth collection, evaluation metric and baselines. Then we evaluate the quality of our models on both content relevance and background relevance, analyze the quality of pseudo positive labels and the impact of negative set construction. We move on to do case studies on the generated image-phrase links and present a distribution of the links on phrase generality and visual semantics. Finally we discuss the difficulty of manual labeling in historic event domain.

6.1 Dataset

The image profiling historic event (IPHE) dataset in work[2] spans from 1816 to 2004 and contains 615 historic events with $17,504$ images and $1,890$ phrases. The scale of the dataset is relatively small which limits the number and diversity of potential image-phrase links. Thus, the dataset needs to be expanded for investigating the image-phrase linking task and for making evaluation more stable. The IPHE dataset is collected by parsing historic events from Wikipedia articles. We follow the same approach. The difference lies in the article selection strategy. IPHE selects articles that are related to locations from a predefined landmark list. We select articles by enumerating the months and years in the same history timespan of IPHE. Our strategy exploits the resource of a historic event list page for each month and year from Wikipedia. Many items in the event list have links to event article pages as well. Our strategy can expand the scale of the dataset by an order of magnitude.

The expanded dataset, called expanded IPHE, contains $3,226$ historic events with $185,173$ images and $16,596$ phrases. A comparison between IPHE and expanded IPHE is summarized in Table 3. We also show the distribution of the events by year in Figure 2. We see that there are three peaks and one plateau in the timespan, corresponding to four wars in the history: American Civil War (1861-1865), World War I (1914-1918), World War II (1939-1945) and Cold War (1947-1991). This shows that our expanded IPHE dataset covers major historic events.

6.2 Groundtruth collection, evaluation metric and baselines

Groundtruth collection: To label the groundtruth for relevant image-phrase pairs, we transform the definition in section 3 to a guideline for manual labeling as shown in Table 4. It involves two checks before assigning a label. The first check is on the type of noun-phrase: whether it's a concrete object/scene or an abstract phrase? This check guides the worker to one of lines 2 or 7. The worker needs to have a second check on image content together with the phrase. The criteria of the second check are different in these two cases.

Table 3: Comparison of the dataset scale

	$\sharp$ events	$\sharp$ phrases	$\sharp$ images
IPHE dataset	615	$1,890$	$17,504$
expanded IPHE dataset	$3,226$	$16,596$	$185,173$

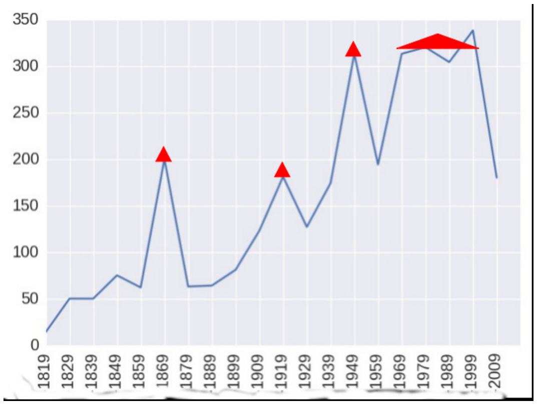


Figure 2: Event distribution on year: the three peaks and one plateau in the dataset correspond to American Civil War (1861-1865), World War I (1914-1918), World War II (1939-1945) and Cold War (1947-1991)

With this guidance, we invite 10 volunteers to label the relevance on randomly sampled $1,000$ image-phrase pairs. On average it takes a volunteer worker around 30 seconds to label one image-phrase pair. As a compromise to the high cost of manual labeling, we assign one volunteer worker rather than three volunteer workers to each pair. This set of $1,000$ image-phrase pairs constitutes our test set. The pseudo positive labels used to train our model are generated

Table 4: Guidance for manual labeling based on the relevance definition

```
1:  check noun phrase type
2:  if the phrase is a concrete noun phrase of object
    or scene
3:      if the image contains the object or scene
4:          label content relevant
5:      else
6:          label irrelevant
7:  if the phrase is a meaningful abstract noun
    phrase
8:      if the phrase can caption the image
9:          label background relevant
10:     else
11:         label irrelevant
```

Table 5: Distribution of relevance archetypes in the historic event image profiling dataset

content relevant	345
background relevant	250
irrelevant	405

outside the test set. The distribution of the two relevance archetypes in historic event image profiling is shown in Table 5. We see that more than 1/3 pairs belong to content relevance archetype and near 1/3 pairs belong to background relevance archetype.

Evaluation metric: For each image-phrase pair, our algorithm generates a matching score and then thresholds it to get a binary relevance classification. The threshold can be tuned. We use two evaluation metrics: average precision (AP) and area under ROC curve(AUC). For AP, we evaluate the corresponding precision at each threshold value t and summarize them by averaging. It focuses on precision. For AUC, we plot the ROC curve by true positive rate tpr and false positive rate fpr at each threshold and calculate the area under the ROC curve. It reflects the probability that a relevant positive image-phrase pair will be ranked higher than an irrelevant image-phrase pair.

$$p(t) = \frac{|\{wi|s_{wi} > t\} \cap grountruth|}{|\{wi|s_{wi} > t\}|} \tag{20}$$

$$AP = \bar{p(t)} \tag{21}$$

$$AUC = \int tpr(t) fpr(t)' \mathrm{d}t \tag{22}$$

Baselines: We introduce three baselines. 1) *pseudo positive label generation algorithm (PPLG)*: it is described in section 5.1. In this baseline, we train models by out-of-domain data. To be specific, VGG16[18] is exploited to predict $p(c|i)$ and public pretrained model from [14] is used to predict $(w|c)$. 2) DeViSE[6]: it learns an embedding model which scales to new concepts and vocabularies. There is no public trained DeViSE model available. We reimplement it on joint dataset of Imagenet[16] and SUN[20][1]. 3) finetuned DeViSE: we directly finetune the above DeViSE model using top 1000 confident pseudo positive labels and set the negative word set to $\mathbf{W} \setminus w$.

6.3 Evaluation on quality of image-phrase links

We compare three models based on our algorithm of weighted instance learning for pseudo positive labels (WIL4PPL) by setting the hyper-parameter θ (see equation (9)) to $0, 0.5, 1$ respectively. When $\theta = 0$, we only use the ranking word loss L^{word} to train our model. When $\theta = 1$, we only use the ranking image loss L^{image} to train our model. Better value of θ could be found by tuning on a validation set. But considering that manual labeling cost in historic event domain is extremely high, we choose to use the whole labeled set as test set to get a stable evaluation result. To train WIL4PPL, we initialize the parameter $\mathbf{U}$ and $\mathbf{V}$ from the pre-trained DeViSE model (our second baseline). The learning rate μ is set to 0.001 and the epoch number is set to 10, which is equivalent to 20 : 1 negative to positive ratio. We set the regularization rank to 300 and the margin Δ to 0.1.

6.3.1 *Comparison on overall relevance*

We compare our method to the three baselines on overall relevance, mixing content relevance and background relevance together. As shown in Table 6, we see that our best model (WIL4PPL($\theta = 1$)) achieves the best performance on both AP (0.661) and AUC (0.666). On AUC metric, our best model achieves significant improvement, 7.4% boosting over PPLG and 3.4% boosting over DeViSE. All the three WIL4PPL models outperform all the baselines.

The inferior performance of PPLG baseline shows that the generated pseudo positive labels are noisy. Comparing WIL4PPL to PPLG shows that our model robustly learns from the noisy pseudo positive labels. On the contrast, the finetuned DeViSE baseline performs much worse than original DeViSE on both metrics. This shows that directly exploiting pseudo labeled positive data and vanilla negative set construction may hamper the original model. WIL4PPL($\theta = 0$) performs much better than finetuned DeViSE, 2.5% better on AP and 1.8% better on AUC. Contrasting WIL4PPL($\theta = 0$) with the finetuned DeViSE, we find that they have similar loss functions, ranking word loss, except that WIL4PPL($\theta = 0$)'s loss function is weighted on the noisy pseudo positive instances. The negative set construction is also different. It shows that careful negative set construction with weighted instance loss function helps to improve the model using the pseudo positive labels.

6.3.2 *Comparison on content relevance and background relevance separately*

To study what is learned by WIL4PPL, we compare our method to baselines on each relevance archetype separately in Table 7 and Table 8. Comparing the performance on content relevance (Table 7) to that on background relevance (Table 8), we see two consistent trends:

1. all methods perform much better on content relevance than background relevance. Our explanation is that background relevance is more abstract and difficult to be captured by models. We have found that volunteer workers are more likely to disagree on background relevance than on content relevance, sometimes due to the lack of background history knowledge.
2. the improvement on AUC metric of all the WIL4PPL models over pretrained baselines (PPLG and DeViSE) are far more significant on background relevance than on content relevance. This suggests that pretrained models handle the content relevance very well and our models manage to learn the background relevance given the pseudo positive labels.

The performance change of finetuned DeViSE across content relevance and background relevance is not only interesting but also inspiring. Compared to DeViSE, the performance of finetuned DeViSE drops significantly on content relevance but increases quite a lot on background relevance. At first glance, we attribute it to overfitting as we only use top confident pseudo labels to finetune DeViSE. But this is definitely not the only reason. We further look at the pseudo positive labels and find that most of them are related to concrete relevance. This is justified by comparing PPLG's performance on content relevance to background relevance in Table 7 and 8. Recall that finetuned DeViSE uses a vanilla negative set construction. In such construction, the learning process needs to distinguish scores between two semantically similar words such as "aircraft carrier USS

[1] We use $1,000$ objects concepts in ILSVRC2012 from imagenet and 205 scene concepts from SUN

Table 6: Comparison with baselines on overall relevance: numbers in the bracket show the improvement over the best baseline (PPLG, DeViSE, finetuned DeViSE)

	AP	AUC
PPLG	0.642	0.592
DeViSE	0.646	0.632
finetuned DeViSE	0.623	0.616
WIL4PPL($\theta = 0$)	0.648(+0.2%)	0.634(+0.2%)
WIL4PPL($\theta = 0.5$)	0.649(+0.3%)	0.639(+0.7%)
WIL4PPL($\theta = 1$)	**0.661**(+**1.5**%)	**0.666**(+**3.4**%)

Table 7: Comparison with baselines on content relevance: numbers in the bracket show the improvement over the best baseline (PPLG, DeViSE, finetuned DeViSE)

	AP	AUC
PPLG	0.576	0.628
DeViSE	0.569	0.654
finetuned DeViSE	0.534	0.623
WIL4PPL($\theta = 0$)	0.572(−0.4%)	0.658(+0.4%)
WIL4PPL($\theta = 0.5$)	0.576(0.0%)	0.668(+1.4%)
WIL4PPL($\theta = 1$)	**0.586**(+**0.1**%)	**0.682**(+**2.8**%)

Table 8: Comparison with baselines on background relevance: numbers in the bracket show the improvement over the based pretrained baseline (PPLG and DeViSE)

	AP	AUC
PPLG	0.273	0.494
DeViSE	0.308	0.587
finetuned DeViSE	0.323	0.621
WIL4PPL($\theta = 0$)	0.308(+0.0%)	0.621(+3.4%)
WIL4PPL($\theta = 0.5$)	0.304(−0.4%)	0.638(+5.1%)
WIL4PPL($\theta = 1$)	**0.313**(+0.5%)	**0.643**(+5.6%)

Enterprise" and "destroyers USS Maddox". And such case is very common in the historic event domain dataset. Therefore, it indicates that negative set construction is very important for the relevance classification task. We can further improve our model's performance via improving negative set construction algorithm.

6.3.3 *Comparison of three WIL4PPL models*

In Table 7 and 8, three WIL4PPL models show consistent performance trend on all metrics: WIL4PPL($\theta = 1$) > WIL4PPL($\theta = 0.5$) > WIL4PPL($\theta = 0$). In these three WIL4PPL models, the weights of pseudo positive labels are the same in their loss functions. So the pseudo positive labels have the same impact on all the models. At first glance, we would attribute the difference to the weightings of ranking image loss and ranking word loss. But combination of ranking image loss and ranking word loss has also been successfully exploited in several works[10][9]. It is unusual that model WIL4PPL($\theta = 0.5$) works consistently worse than WI4PPL($\theta = 1$). We attribute this to the imperfect word negative set construction. Recall that on the word side, we cluster words based on features learned from text corpus only. Words that correspond to the same visual content may not be in the same cluster as they are clustered based on their textual meaning. This leads to that the learning process has to distinguish the score between words that share similar visual content. There is no such problem in the image negative set construction. Again, this indicates that negative set construction is very important and we have the potential to improve our model performance further by improving negative set construction in the future work.

6.4 Analysis on the Generated Image-Phrase Links

We set the threshold at $fpr = 0.1, tpr = 0.2$ to generate image-phrase links. Altogether $59,653$ image-phrase links are generated automatically. We first study some cases of the links. Then we automatically categorize phrases to four different categories and list the distribution of the generated links under these four categories.

Case study: Table 9 shows eight image-phrase links generated by our model. The first row shows four content relevant links and the second row shows four background relevant links. For content relevant links, our model could match images to very specific phrase such as celebrity name, ship serial name, place name within an event. For background relevant links, our model outputs meaningful links that are educational to users without enough background knowledge.

Category study: we categorize the phrases from two dimensions. Each dimension contains two categories and we get a 2×2 table for phrase categorization. The first dimension is the generality of a phrase. It includes two categories: common noun phrase and proper noun phrase. This could be automatically categorized based on whether there is a capital letter in the phrase. The second dimension is the visual semantic of a phrase. It includes two categories: object and scene. We consider the concepts from ImageNet dataset as object concepts and concepts from SUN dataset as scene concepts. We calculate cosine similarity between feature vectors of noun phrases and these concepts. If the similarity between the phrase and any object concept is greater than 0.4, we categorize it as object semantic. The same holds for categorization of scene semantic. Note that it is possible that a phrase can be categorized to both object semantic and scene semantic. As shown in Table 10, the proper noun phrase occupies roughly 1/3 of the generated links, which indicates that these links provide rich event-specific semantic information. The object noun phrase occupies roughly 3/4 of the generated links, which indicates that these links contain rich specific image information.

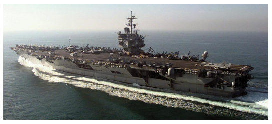
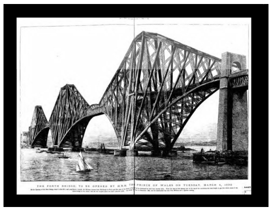
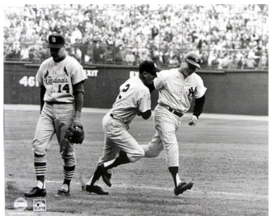
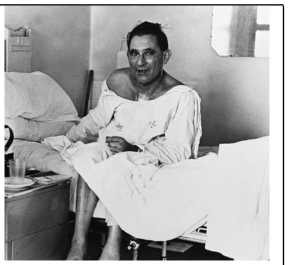
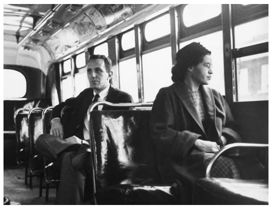
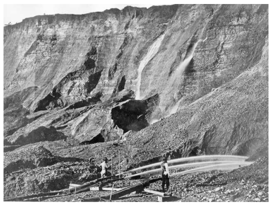
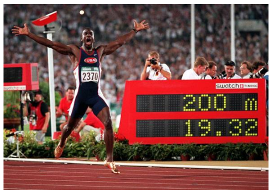
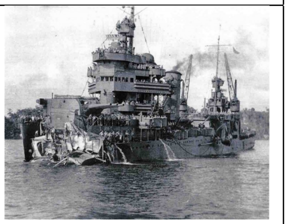
Table 10: Pair Distribution in Our Dataset

	common noun	proper noun
object	$32,650$	$12,846$
scene	$9,641$	$4,516$

6.5 Discussion on difficulty of manual labeling

Recall that we invite 10 volunteer workers to label the image-phrase pair and each image-phrase pair is labeled by one volunteer worker. We randomly select some labeled pairs in the test set and randomly distribute them to several volunteer workers who didn't label them in the first round.

Table 9: Case Study on the generated image-phrase links: the first rows shows four content relevant links and the second rows shows four background relevant links.

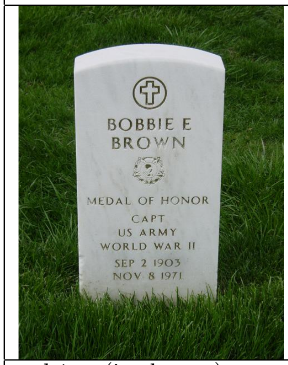
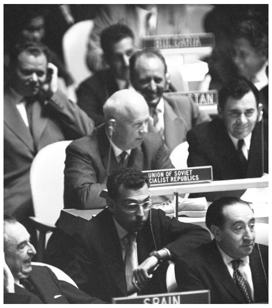
 The United States Navy aircraft carrier USS Enterprise	 longest bridge	 New York Highlander Jack Chesbro	 Groote Schuur Hospital
 the Montgomery Bus Boycott	 the San Francisco Gold Rush	 the world record	 World War II Guadalcanal Campaign

On such pairs, we collect more than one labels from different volunteer workers. We study such pairs by face-to-face interviewing the volunteer workers for the reason of their choice. Table 11 show two typical cases.

In the first case, vol.2 and vol.5 make the "background relevant" choice while vol.1 makes the "irrelevant" choice. After interviewing, we find out that vol.2 and vol.5 looked at the image carefully and read the text on the memorial while vol.1 just glanced at the image and failed to connect the memorial to World War II. In the second case, vol.2, vol.5, vol.7, vol.9 choose "irrelevant" while volunteer 6 chooses "content relevant". We ask vol.6 why she chooses content relevant and she tells that there is a white shoe in the center of the image though it is so easy to miss because it is mixed up with the background.

What are the volunteer workers doing in the average 30 seconds labeling progress? Sometimes they need to search on the web to get the background knowledge of the phrase especially for the proper noun phrases. Sometimes, they have to look at the images very carefully to recognize the linkage between image detail and the phrase. It is much more painful than just to label a general concept. Thus, we choose to learn with pseudo positive labels rather than to wait until enough clean labels are collected.

Table 11: Illustratinon of difficulty of Manual labeling: the first row shows the phrase, the second row shows the image and the third row shows the manual label result.

World War II	his shoe
 BOBBIE E BROWN MEDAL OF HONOR CAPT US ARMY WORLD WAR II SEP 2 1903 NOV 8 1971	
vol.1 (irrelevant), vol.2 (background relevant),vol.5 (background relevant)	vol.2 (irrelevant), vol.5 (irrelevant), vol.6 (content relevant), vol.7 (irrelevant), vol.9 (irrelevant)

7. CONCLUSION

In this paper, we propose to improve image profiling for historic events by adding explicit semantic information, linking images to the corresponding noun phrases in the event description, so that users can easily understand the correspondence between images and text descriptions. To achieve this goal, we need to judge whether an image and a phrase is related. Such judgement is made by calculating the relevance of an image-phrase pair via a real-valued matching score and then comparing it to a threshold. The relevance of an image-phrase pair can be quite diverse. We therefore derive a case-based definition of image-phrase pair relevance. Under each case, we give a separate definition of relevance that could be used to make reproducible judgement of relevance. We exploit instance-wise ranking loss function to learn the matching score between an image-phrase pair, which is shown to be more robust then 0-1 loss function on relevance modeling tasks. We deal with two challenges in this learning task: 1) how to automatically generate labeled positive data: we leverage out-of-domain labeled dataset to generate pseudo positive labels and propose a new algorithm (WIL4PPL) that can robustly learn the model from the noisy pseudo positive labels; 2) how to automatically generate negative data: we propose a negative set generation algorithm to automatically generate more proper negative data for model training. We compare our proposed algorithm to three baselines. We evaluate the quality of learnt semantic information - the linkage between related images and phrases - via detailed analysis and case studies. Comprehensive experiments show that our proposed algorithm outperforms the baselines significantly. In the future work, we will further improve the quality of semantic information in image profiling, for example, utilizing self-paced learning [7] for adjusting instance weights in the learning process and improving negative set generation for model training.

8. ACKNOWLEDGMENTS

We thank NVIDIA Corporation for their donation of GPUs. This research was supported by the Fundamental Research Funds for the Central Universities and the Research Funds of Renmin University of China (No. 14XNLQ01) and the Beijing Natural Science Foundation (No. 4142029). This research was also supported in part by the National Science Foundation (NSF) under grant number IIS-1251187. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

9. REFERENCES

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

[2] J. Chen, Q. Jin, Y. Yu, and A. G. Hauptmann. Image profiling for history events on the fly. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*, pages 291–300, 2015.

[3] J. Chen, Q. Jin, W. Zhang, S. Bao, Z. Su, and Y. Yu. Tell me what happened here in history. In *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, pages 467–468, 2013.

[4] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision - ECCV 2002, 7th European Conference on Computer Vision, Copenhagen, Denmark, May 28-31, 2002, Proceedings, Part IV*, pages 97–112, 2002.

[5] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[6] A. Frome, G. S. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2121–2129, 2013.

[7] L. Jiang, D. Meng, S. Yu, Z. Lan, S. Shan, and A. G. Hauptmann. Self-paced learning with diversity. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2078–2086, 2014.

[8] B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S. Chang. Visual affect around the world: A large-scale multilingual visual sentiment ontology. In *Proceedings of the 23rd Annual ACM Conference on Multimedia Conference, MM '15, Brisbane, Australia, October 26 - 30, 2015*, pages 159–168, 2015.

[9] A. Karpathy, A. Joulin, and F. Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 1889–1897, 2014.

[10] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *CoRR*, abs/1411.2539, 2014.

[11] M. Lin, Z. Hu, S. Liu, M. Wang, R. Hong, and S. Yan. eheritage of shadow puppetry: creation and manipulation. In *ACM Multimedia Conference, MM '13, Barcelona, Spain, October 21-25, 2013*, pages 183–192, 2013.

[12] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, USA, 1999.

[13] T. Mensink and J. van Gemert. The rijksmuseum challenge: Museum-centered visual recognition. 2014.

[14] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[15] G. A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, Nov. 1995.

[16] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[17] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 815–823, 2015.

[18] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[19] 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.

[20] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13-18 June 2010*, pages 3485–3492, 2010.