A Low-Rank Approximation Approach to Learning Joint Embeddings of News Stories and Images for Timeline Summarization

William Yang Wang[1]*, **Yashar Mehdad**[3], **Dragomir R. Radev**[2], **Amanda Stent**[4]
[1]School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA
[2]Department of EECS, University of Michigan, Ann Arbor, MI 48109, USA
[3]Yahoo, Sunnyvale, CA 94089, USA and [4]New York, NY 10036, USA
yww@cs.cmu.edu, {ymehdad,stent}@yahoo-inc.com, radev@umich.edu

Abstract

A key challenge for timeline summarization is to generate a concise, yet complete storyline from large collections of news stories. Previous studies in extractive timeline generation are limited in two ways: first, most prior work focuses on fully-observable ranking models or clustering models with hand-designed features that may not generalize well. Second, most summarization corpora are text-only, which means that text is the sole source of information considered in timeline summarization, and thus, the rich visual content from news images is ignored. To solve these issues, we leverage the success of matrix factorization techniques from recommender systems, and cast the problem as a sentence recommendation task, using a representation learning approach. To augment text-only corpora, for each candidate sentence in a news article, we take advantage of top-ranked relevant images from the Web and model the image using a convolutional neural network architecture. Finally, we propose a scalable low-rank approximation approach for learning joint embeddings of news stories and images. In experiments, we compare our model to various competitive baselines, and demonstrate the state-of-the-art performance of the proposed text-based and multimodal approaches.

1 Introduction

Timeline summarization is the task of organizing crucial milestones of a news story in a temporal order, e.g. (Kedzie et al., 2014; Lin et al., 2012). A timeline example for the *2010 British Oil spill* generated by our system is shown in Figure 1. The task is challenging, because the input often includes a large number of news articles as the story is developing each day, but only a small portion of the key information is needed for timeline generation. In addition to the conciseness requirement, timeline summarization also has to be complete—all key information, in whatever form, must be presented in the final summary.

To distill key insights from news reports, prior work in summarization often relies on feature engineering, and uses clustering techniques (Radev et al., 2004b) to select important events to be included in the final summary. While this approach is unsupervised, the process of feature engineering is always expensive, and the number of clusters is not easy to estimate. To present a complete summary, researchers from the natural language processing (NLP) community often solely rely on the textual information, while studies in the computer vision (CV) community rely solely on the image and video information. However, even though news images are abundantly available together with news stories, approaches that jointly learn textual and visual representations for summarization are not common.

In this paper, we take a more radical approach to timeline summarization. We formulate the problem as a sentence recommendation task—instead of recommending items to users as in a recommender system, we recommend important sentences to a timeline. Our approach does not require feature engineering: by using a matrix factorization framework, we are essentially performing representation learn-

*This work was performed when William Wang and Dragomir Radev were visiting Yahoo NYC.

Proceedings of NAACL-HLT 2016, pages 58–68,
San Diego, California, June 12-17, 2016.

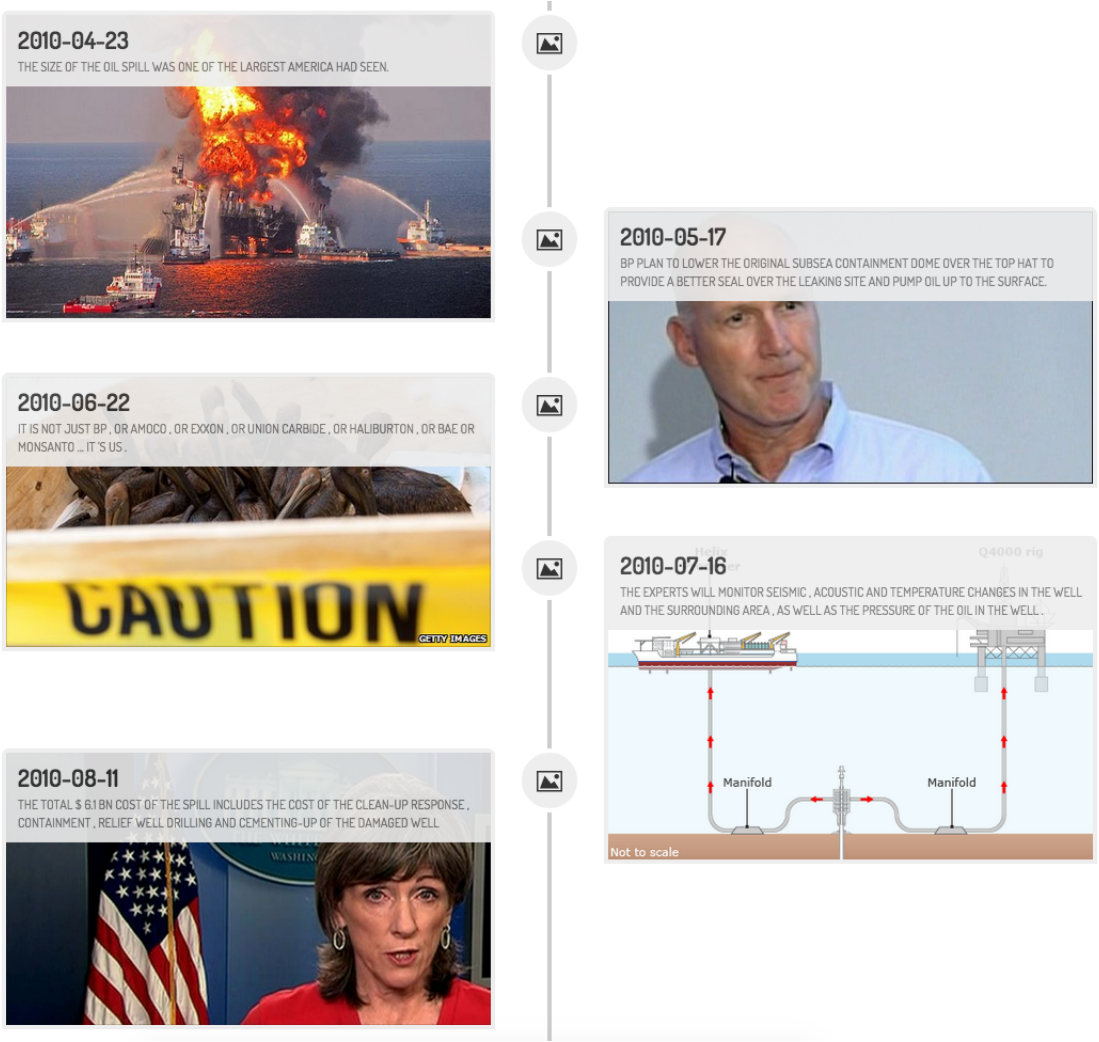

Figure 1: A timeline example for the BP oil spill generated by our proposed method. *Note that we use Yahoo! Image Search to obtain the top-ranked image for each candidate sentence.*

ing to model the continuous representation of sentences and words. Since most previous timeline summarization work (and therefore, corpora) only focuses on textual information, we also provide a novel web-based approach for harvesting news images: we query Yahoo! image search with sentences from news articles, and extract visual cues using a 15-layer convolutional neural network architecture. By unifying text and images in the low-rank approximation framework, our approach learns a joint embedding of news story texts and images in a principled manner. In empirical evaluations, we conduct experiments on two publicly available datasets, and demonstrate the efficiency and effectiveness of our approach. By comparing to various baselines, we show that our approach is highly scalable and achieves state-of-the-art performance. Our main contributions are three-fold:

- We propose a novel matrix factorization approach for extractive summarization, leveraging the success of collaborative filtering;
- We are among the first to consider representation learning of a joint embedding for text and images in timeline summarization;
- Our model significantly outperforms various competitive baselines on two publicly available datasets.

2 Related Work

Supervised learning is widely used in summarization. For example, the seminal study by Kupiec et al. (1995) used a Naive Bayes classifier for selecting sentences. Recently, Wang et al. (2015) proposed a regression method that uses a joint loss function, combining news articles and comments. Additionally, unsupervised techniques such as language modeling (Allan et al., 2001) have been used for temporal summarization. In recent years, ranking and graph-based methods (Radev et al., 2004b; Erkan and Radev, 2004; Mihalcea and Tarau, 2004; Fader et al., 2007; Hassan et al., 2008; Mei et al., 2010; Yan et al., 2011b; Yan et al., 2011a; Zhao et al., 2013; Ng et al., 2014; Zhou et al., 2014; Glavaš and Šnajder, 2014; Tran et al., 2015; Dehghani and Asadpour, 2015) have also proved popular for extractive timeline summarization, often in an unsupervised setting. Dynamic programming (Kiernan and Terzi, 2009) and greedy algorithms (Althoff et al., 2015) have also been considered for constructing summaries over time.

Our work aligns with recent studies on latent variable models for multi-document summarization and storyline clustering. Conroy et al. (2001) were among the first to consider latent variable models, even though it is difficult to incorporate features and high-dimensional latent states in a HMM-based model. Ahmed et al. (2011) proposed a hierarchical nonparametric model that integrates a Recurrent Chinese Restaurant Process with Latent Dirichlet Allocation to cluster words over time. The main issues with this approach are that it does not generate human-readable sentences, and that scaling nonparametric Bayesian models is often challenging. Similarly, Huang and Huang (2013) introduced a joint mixture-event-aspect model using a generative method. Navarro-Colorado and Saquete (2015) combined temporal information with topic modeling, and obtained the best performance in the cross-document event ordering task of SemEval 2015.

There has been prior work (Wang et al., 2008; Lee et al., 2009) using matrix factorization to perform sentence clustering. A key distinction between our work and this previous work is that our method requires no additional sentence selection steps after sentence clustering, so we avoid error cascades.

Zhu and Chen (2007) were among the first to consider multimodal timeline summarization, but they focus on visualization, and do not make use of images. Wang et al. (2012) investigated multimodal timeline summarization by considering cosine similarity among various feature vectors, and then using a graph based algorithm to select salient topics. In the computer vision community, Kim and Xing (2014) made use of community web photos, and generate storyline graphs for image recommendation. Interestingly, Kim et al. (2014) combined images and videos for storyline reconstruction. However, none of the above studies combine textual and visual information for timeline summarization.

3 Our Approach

We now describe the technical details of our low-rank approximation approach. First, we motivate our approach. Next, we explain how we formulate the timeline summarization task as a matrix factorization problem. Then, we introduce a scalable approach for learning low-dimensional embeddings of news stories and images.

3.1 Motivation

We formulate timeline summarization as a low-rank matrix completion task because of the following considerations:

- **Simplicity** In the past decade, a significant amount of work on summarization has focused on designing various lexical, syntactic and semantic features. In contrast to prior work, we make use of low-rank approximation techniques to learn representations directly from data. This way, our model does not require strong domain knowledge or lots of feature engineering, and it is easy for developers to deploy the system in real-world applications.

- **Scalability** A major reason that recommender systems and collaborative filtering techniques have been very successful in industrial applications is that matrix completion techniques are relatively sophisticated, and are known to scale up to large recommendation datasets with more than 100 million ratings (Bennett and Lanning,

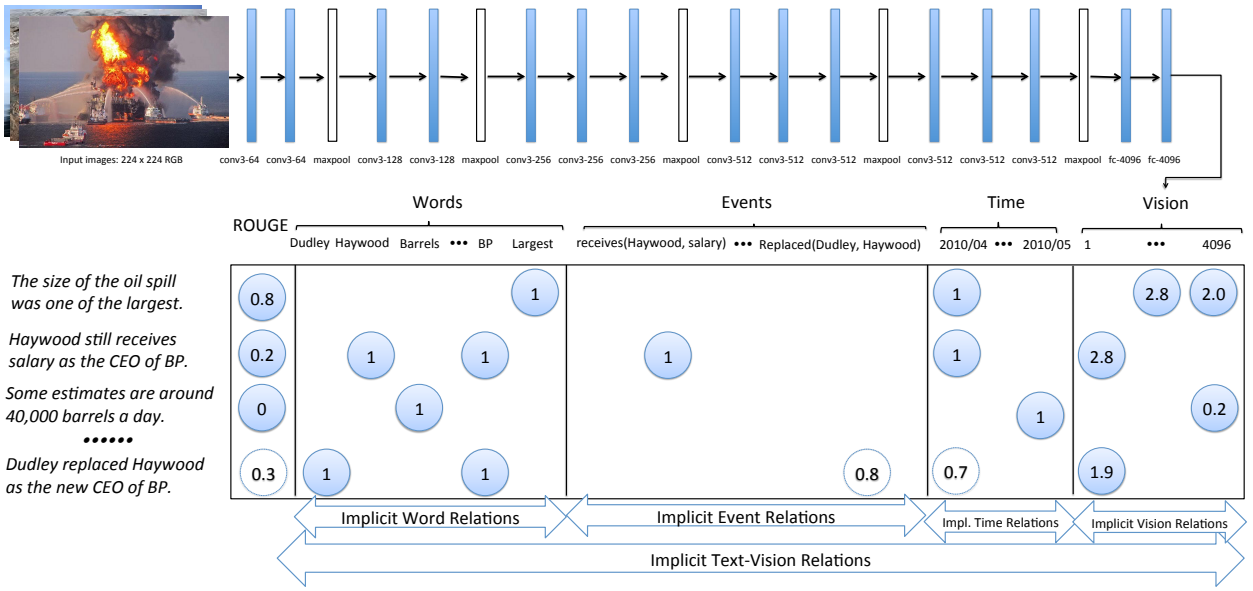

Figure 2: Our low-rank approximation framework for learning joint embedding of news stories and images for timeline summarization.

2007). Therefore, we believe that our approach is practical for processing large datasets in this summarization task.

- **Joint Multimodal Modeling** A key challenge of supervised learning approaches for summarization is to select informative sentences. In this work, we make use of multimodality to select important sentences.

3.2 Problem Formulation

Since the Netflix competition (Bell and Koren, 2007), collaborative filtering techniques with latent factor models have had huge success in recommender systems. These latent factors, often in the form of low-rank embeddings, capture not only explicit information but also implicit context from the input data. In this work, we propose a novel matrix factorization framework to "recommend" key sentences to a timeline. Figure 2 shows an overview of the framework.

More specifically, we formulate this task as a matrix completion problem. Given a news corpus, we assume that there are m total sentences, which are the rows in the matrix. The first column is the *metric* section, where we use ROUGE (Lin, 2004) as the metric to pre-compute a *sentence importance* score between a candidate sentence and a human-generated summary. During training, we use these scores to tune model parameters, and during testing, we predict the sentence importance scores given the features in other columns. That is, we learn the embedding of important sentences.

The second set of columns is the *text feature* section. In our experiments, this includes word observations, subject-verb-object (SVO) events, and the publication date of the document from which the candidate sentence is extracted. In our preprocessing step, we run the Stanford part-of-speech tagger (Toutanova et al., 2003) and MaltParser (Nivre et al., 2006) to generate SVO events based on dependency parses. Additional features can easily be incorporated into this framework; we leave the consideration of additional features for future work.

Finally, for each sentence, we use an image search engine to retrieve a top-ranked relevant image, and then we use a convolutional neural network (CNN) architecture to extract visual features in an unsupervised fashion. We use a CNN model from Simonyan and Zisserman (2015), which is trained on the ImageNet Challenge 2014 dataset (Russakovsky et al., 2014). In our work, we keep the 16 convolutional layers and max-pool operations. To extract

neural network features, we remove the final fully-connected-1000 layer and the softmax function, resulting in 4096 features for each image.

The total number of columns in the input matrix is n. Our matrix M now encodes preferences for a sentence, together with its lexical, event, and temporal attributes, and visual features for an image highly relevant to the sentence. Here we use i to index the i-th sentence and j to index the j-th column. We scale the columns by the standard deviation.

3.3 Low-Rank Approximation

Following prior work (Koren et al., 2009), we are interested in learning two low-rank matrices $P \in \mathbf{R}^{k \times m}$ and $Q \in \mathbf{R}^{k \times n}$. The intuitition is that P is the embedding of all candidate sentences, and Q is the embedding of textual and visual features, as well as the sentence importance score, event, and temporal features. Here k is the number of latent dimensions, and we would like to approximate $M_{(i,j)} \simeq \vec{p_i}^T \vec{q_j}$, where $\vec{p_i}$ is the latent embedding vector for the i-th sentence and $\vec{q_j}$ is the latent embedding vector for the j-th column. We seek to approximate the matrix M by these two low-rank matrices P and Q. We can then formulate the optimization problem for this task:

$$\min_{P,Q} \sum_{(i,j) \in M} (M_{(i,j)} - \vec{p_i}^T \vec{q_j})^2 + \lambda_P ||\vec{p_i}||^2 + \lambda_Q ||\vec{q_j}||^2$$

here, λ_P and λ_Q are regularization coefficients to prevent the model from overfitting. To solve this optimization problem efficiently, a popular approach is stochastic gradient descent (SGD) (Koren et al., 2009). In contrast to traditional methods that require time-consuming gradient computation, SGD takes only a small number of random samples to compute the gradient. SGD is also natural to online algorithms in real-time streaming applications, where instead of retraining the model with all the data, parameters might be updated incrementally when new data comes in. Once we have selected a random sample $M_{(i,j)}$, we can simplify the objective function:

$$(M_{(i,j)} - \vec{p_i}^T \vec{q_j})^2 + \lambda_P (\vec{p_i}^T \vec{p_i}) + \lambda_Q (\vec{q_j}^T \vec{q_j})$$

Now, we can calculate the sub-gradients of the two latent vectors $\vec{p_i}$ and $\vec{q_j}$ to derive the following variable update rules:

$$\vec{p_i} \leftarrow \vec{p_i} + \delta(\ell_{(i,j)} q_j - \lambda_P \vec{p_i}) \quad (1)$$

$$\vec{q_j} \leftarrow \vec{q_j} + \delta(\ell_{(i,j)} p_i - \lambda_Q \vec{q_j}) \quad (2)$$

Here, δ is the learning rate, whereas $\ell_{(i,j)}$ is the loss function that estimates how well the model approximates the ground truth:

$$\ell(i,j) = M_{(i,j)} - \vec{p_i}^T \vec{q_j}$$

The low-rank approximation here is accomplished by reconstructing the M matrix with the two low-rank matrices P and Q, and we use the row and column regularizers to prevent the model from overfitting to the training data.

SGD-based optimization for matrix factorization can also be easily parallelized. For example, HOGWILD! (Recht et al., 2011) is a lock-free parallelization approach for SGD. In contrast to synchronous approaches where idle threads have to wait for busy threads to sync up parameters, HOGWILD! is an asynchronous method: it assumes that because text features are sparse, there is no need to perform synchronization of the threads. In reality, although this approach might not work for speech or image related tasks, it performs well in various text based tasks. In this work, we follow a recently proposed approach called fast parallel stochastic gradient descent (FPSG) (Chin et al., 2015), which is partly inspired by HOGWILD!.

3.4 Joint Modeling of Mixed Effects

Matrix factorization is a relatively complex method for modeling latent factors. So, an important question to ask is: in the context of timeline summarization, what is this matrix factorization framework modeling?

From equation (1), we can see that the latent sentence vector $\vec{p_i}$ will be updated whenever we encounter a $M_{(i,\cdot)}$ sample (e.g., all the word, event, time, and visual features for this particular sentence) in a full pass over the training data. An interesting aspect about matrix factorization is that, in addition to using the previous row embedding $\vec{p_i}$ to update the variables in equation (1), the column embedding $\vec{q_j}$ will also be used. Similarly, when updating the latent column embedding $\vec{q_j}$ in equation (2), the pass will visit all samples that have non-zero items in that

column, while making use of the $\vec{p_i}$ vector. Essentially, in timeline summarization, this approach is modeling the mixed effects of sentence importance, lexical features, events, temporal information, and visual factors. For example, if we are predicting the ROUGE score of a new sentence at testing, the model will take the explicit sentence-level features into account, together with the learned latent embedding of ROUGE, which is recursively influenced by other metrics and features during training.

Our approach shares similarities with some recent advances in word embedding techniques. For example, word2vec uses the continuous bag-of-words (CBOW) and SkipGram algorithms (Mikolov et al., 2013) to learn continuous representations of words from large collections of text and relational data. A recent study (Levy and Goldberg, 2014) shows that the technique behind word2vec is very similar to implicit matrix factorization. In our work, we consider multiple sources of information to learn the joint embedding in a unified matrix factorization framework. In addition to word information, we also consider event and temporal cues.

3.5 The Matrix Factorization Based Timeline Summarization

We outline our matrix factorization based timeline summarization method in Algorithm 1. Since this is a supervised learning approach, we assume the corpus includes a collection of news documents S, as well as human-written summaries H for each day of the story. We also assume the publication date of each news document is known (or computable).

During training, we traverse each sentence in this corpus, and compute a *sentence importance score* (I_i) by comparing the sentence to the human generated summary for that day using ROUGE (Lin, 2004). If a human summary is not given for that day, I_i will be zero. We also extract subject-verb-object event representations, using the Stanford part-of-speech tagger (Toutanova et al., 2003) and MaltParser (Nivre et al., 2006). We use the publication date of the news document as the publication date of the sentence. Visual features are extracted using a very deep CNN (Simonyan and Zisserman, 2015). Finally, we merge these vectors into a joint vector to represent a row in our matrix factorization framework. Then, we perform stochastic gradient descent

Algorithm 1 A Matrix Factorization Based Timeline Summarization Algorithm

Input: news documents S, human summaries H for each day t.
procedure TRAINING(S^{tr}, H)
  for each training sentence S_i^{tr} in S^{tr} **do**
    $I_i \leftarrow$ ComputeImportanceScores(S_i^{tr},H_t)
    $\vec{E_i} \leftarrow$ ExtractSVOEvents(S_i^{tr})
    $\vec{D_i} \leftarrow$ ExtractPublicationDate(S_i^{tr})
    $\vec{V_i} \leftarrow$ ExtractVisualFeatures(V_i^{tr})
    $\vec{M_i} \leftarrow$ MergeVectors($I_i, \vec{E_i}, \vec{D_i}, \vec{V_i}$)
  end for
  for each epoch e **do**
    for each cell i, j in M **do**
      $\vec{p_i}^{(e)} \leftarrow \vec{p_i}^{(e)} + \delta(\ell_{(i,j)} q_j^{(e)} - \lambda_P \vec{p_i}^{(e)})$
      $\vec{q_j}^{(e)} \leftarrow \vec{q_j}^{(e)} + \delta(\ell_{(i,j)} p_i^{(e)} - \lambda_Q \vec{q_j}^{(e)})$
    end for
  end for
end procedure
procedure TESTING(S^{te})
  for each test sentence S_i^{te} in S^{te} **do**
    $\vec{E_i} \leftarrow$ ExtractSVOEvents(S_i^{te})
    $\vec{D_i} \leftarrow$ ExtractPublicationDate(S_i^{te})
    $\vec{V_i} \leftarrow$ ExtractVisualFeatures(S_i^{te})
    $\vec{M_i} \leftarrow$ MergeVectors($\vec{E_i}, \vec{D_i}, \vec{V_i}$)
    $I_i \leftarrow$ PredictROUGE($\vec{M_i}, P, Q$)
  end for
  for each day t in S^{te} **do**
    $H^{te} \leftarrow$ SelectTopSentences(S_t^{te},$\vec{I_t}$)
  end for
end procedure

training to learn the hidden low-rank embeddings of sentences and features P and Q using the update rules outlined earlier.

During testing, we still extract events and publication dates, and the PredictROUGE function estimates the sentence importance score I_i, using the trained latent low-rank matrices P and Q. To be more specific, we extract the text, vision, event, and publication date features for a candidate sentence i. Then, given these features, we update the embeddings for this sentence, and make the prediction by taking the dot product of this i-th column of P (i.e., $\vec{p_i}$) and the ROUGE column of Q (i.e., $\vec{q_1}$). This predicted scalar value I_i indicates the likelihood of the sentence being included in the final timeline summary. Finally, we go through the predicted results of each sentence in the timeline in temporal order, and

include the top-ranked sentences with the highest sentence importance scores. It is natural to scale this method from daily summaries to weekly or monthly summaries.

4 Experiments

In this section, we investigate the empirical performance of the proposed method, comparing to various baselines. We first discuss our experimental settings, including our primary dataset and baselines. Then, we discuss our evaluation results. We demonstrate the robustness of our approach by varying the latent dimensions of the low-rank matrices. Next, we show additional experiments on a headline-based timeline summarization dataset. Finally, we provide a qualitative analysis of the output of our system.

4.1 Comparative Evaluation on the 17 Timelines Dataset

We use the 17 timelines dataset which has been used in several prior studies (Tran et al., 2013b; Tran et al., 2013a). It includes 17 timelines from 9 topics[1] from major news agencies such as CNN, BBC, and NBC News. Only English documents are included. The dataset contains 4,650 news documents. We use Yahoo! Image Search to retrieve the top-ranked image for each sentence.[2] We follow exactly the same topic-based cross-validation setup that was used in prior work (Tran et al., 2013b): we train on eight topics, test on the remaining topic, and repeat the process eight times. The number of training iterations was set to 20; the k was set to 200 for the text only model, and 300 for the joint text/image model; and the vocabulary is 10K words for all systems. The common summarization metrics ROUGE-1, ROUGE-2, and ROUGE-S are used to evaluate the quality of the machine-generated timelines. We consider the following baselines:

[1]The nine topics are the BP oil spill, Egyptian protests, Financial crisis, H1N1, Haiti earthquake, Iraq War, Libya War, Michael Jackson death, and Syrian crisis.

[2]We are not aware of any publicly available dataset for timeline summarization that includes both text and images. Most of these datasets are text-only, not including the original article file or links to accompanying images. We adopted this Web-based corpus enhancement technique as a proximity for news images. Our low-rank approximation technique can be applied to the original news images in the same way.

(a) ROUGE:0 (b) ROUGE:.009.

Figure 3: Examples of retrieved Web images. The left image was retrieved by using a non-informative sentence: "*The latest five minute news bulletin from BBC World Service*". The right image was retrieved using a crucial sentence with a non-zero ROUGE score vs. a human summary, "*Case study : Gulf of Mexico oil spill and BP On 20 April 2010 a deepwater oil well exploded in the Gulf of Mexico*".

- **Random**: summary sentences are randomly selected from the corpus.
- **MEAD**: a feature-rich, classic multi-document summarization system (Radev et al., 2004a) that uses centroid-based summarization techniques.
- **Chieu et al.** (Chieu and Lee, 2004): a multi-document summarization system that uses TFIDF scores to indicate the "popularity" of a sentence compared to other sentences.
- **ETS** (Yan et al., 2011b): a state-of-the-art unsupervised timeline summarization system.
- **Tran et al.** (Tran et al., 2013b): another state-of-the-art timeline summarization system based on learning to rank techniques, and for which results on the 17 Timelines dataset have been previously reported.
- **Regression**: a part of a state-of-the-art extractive summarization method (Wang et al., 2015) that formulates the sentence extraction task as a supervised regression problem. We use a state-of-the-art regression implementation in Vowpal Wabbit[3].

We report results for our system and the baselines on the 17 timelines dataset in Table 1. We see that the random baseline clearly performs worse than the other methods. Even though Chieu et al. (2004)

[3]https://github.com/JohnLangford/vowpal_wabbit

Methods	ROUGE-1	ROUGE-2	ROUGE-S
Random	0.128	0.021	0.026
Chieu et al.	0.202	0.037	0.041
MEAD	0.208	0.049	0.039
ETS	0.207	0.047	0.042
Tran et al.	0.230	0.053	0.050
Regression	0.303	0.078	0.081
Our approach			
Text	0.312	0.089	0.112
Text+Vision	**0.331**	**0.091**	**0.115**

Table 1: Comparing the timeline summarization performance to various baselines on the 17 Timelines dataset. The best-performing results are highlighted in **bold**.

and MEAD (Radev et al., 2004a) are not specifically designed for the timeline summarization task, they perform relatively well against the ETS system for timeline summarization (Yan et al., 2011b). Tran et al. (2013b) was previously the state-of-the-art method on the 17 timelines dataset. The ROUGE regression method is shown as a strong supervised baseline. Our matrix factorization approach outperforms all of these methods, achieving the best results in all three ROUGE metrics. We also see that there is an extra boost in the performance when considering visual features for timeline summarization. Figure 3 shows an example of the retrieved images we used. In general, images retrieved by using more important sentences (measured by ROUGE) include objects, as well a more vivid and detailed scene.

4.2 Comparative Evaluation Results for Headline Based Timeline Summarization

To evaluate the robustness of our approach, we show the performance of our method on the recently released *crisis* dataset (Tran et al., 2015). The main difference between the crisis dataset and the 17 timelines dataset is that here we focus on a headline based timeline summarization task, rather than using sentences from the news documents. The crisis dataset includes four topics: Egypt revolution, Libya war, Syria war, and Yemen crisis. There are a total of 15,534 news documents in the dataset, and each topic has around 4K documents. There are 25 manually created timelines for these topics, collected from major news agencies such as BBC, CNN, and Reuters. We perform standard cross-validation on this dataset: we train on three topics, and test on the other. Here k is set to 300, and the vocabulary is 10K words for all systems. Table 3 shows the performance of our system. Our system is significantly better than the strong supervised regression baseline. When considering joint learning of text and vision, we see that there is a further improvement.

Methods	ROUGE-1	ROUGE-2	ROUGE-S
Regression	0.207	0.045	0.039
Our approach			
Text	0.211	0.046	0.040
Text+Vision	**0.232**	**0.052**	**0.044**

Table 3: Comparing the timeline summarization performance to the state-of-the-art supervised sentence regression approach on the crisis dataset. The best-performing results are highlighted in **bold**.

4.3 Headline Based Timeline Summarization: A Qualitative Analysis

In this section, we perform a qualitative analysis of the output of our system for the headline based timeline summarization task. We train the system on three topics, and show a sample of the output on the Syria war. Table 2 shows a subset of the timeline for the Syria war generated by our system. We see that most of the daily summaries are relevant to the topic, except the one generated on 2011-11-24. When evaluating the quality, we notice that most of them are of high quality: after the initial hypothesis of the Syria war on 2011-11-18, the following daily summaries concern the world's response to the crisis. We show that most of the relevant summaries are also providing specific information, with an exception on 2011-12-02. We suspect that this is because this headline contains three keywords "syria", "civil", "war", and also the key date information: the model was trained partly on the Libya war timeline, and therefore many features and parameters were activated in the matrix factorization framework to give a high recommendation in this testing scenario. In contrast, when evaluating the output of the joint text and vision system, we see that this error is eliminated: the selected sentence on 2011-12-02 is "*Eleven killed after weekly prayers in Syria on eve of Arab League deadline*".

Date	Summary	Relevant?	Good?
2011-11-18	Syria is heading inexorably for a civil war and an appalling bloodbath	✓	✓
2011-11-19	David Ignatius Sorting out the rebel forces in Syria	✓	✓
2011-11-20	Syria committed crimes against humanity, U.N. panel finds	✓	✓
2011-11-21	Iraq joins Syria civil war warnings	✓	✓
2011-11-22	The Path to a Civil War in Syria	✓	✓
2011-11-23	Report Iran, Hezbollah setting up militias to prepare for post-Assad Syria	✓	✓
2011-11-24	Q&A Syria's daring actress Features Al Jazeera English	×	×
2011-11-25	Syria conflict How residents of Aleppo struggle for survival	✓	✓
2011-11-27	Syrian jets bomb rebel areas near Damascus as troops battle	✓	✓
2011-11-28	Is the Regional Showdown in Syria Rekindling Iraqs Civil War?	✓	✓
2011-11-29	Syria Crisis Army Drops Leaflets Over Damascus	✓	✓
2011-11-30	Russia says West's Syria push "path to civil war"	✓	✓
2011-12-01	UN extends Syria war crimes investigation despite opposition from China	✓	✓
2011-12-02	Un syria civil war 12 2 2011	✓	×
2011-12-03	Israel says fires into Syria after Golan attack on troops	✓	✓

Table 2: A timeline example for Syria war generated by our text-only system.

5 Conclusions

In this paper, we introduce a low-rank approximation based approach for learning joint embeddings of news stories and images for timeline summarization. We leverage the success of matrix factorization techniques in recommender systems, and cast the multi-document extractive summarization task as a sentence recommendation problem. For each sentence in the corpus, we compute its similarity to a human-generated abstract, and extract lexical, event, and temporal features. We use a convolutional neural architecture to extract vision features. We demonstrate the effectiveness of this joint learning method by comparison with several strong baselines on the 17 timelines dataset and a headline based timeline summarization dataset. We show that image features improve the performance of our model significantly. This further motivates investment in joint multimodal learning for NLP tasks.

Acknowledgments

The authors would like to thank Kapil Thadani and the anonymous reviewers for their thoughtful comments.

References

Amr Ahmed, Qirong Ho, Choon H Teo, Jacob Eisenstein, Eric Xing, and Alex Smola. 2011. Online inference for the infinite topic-cluster model: Storylines from streaming text. In *Proceedings of AISTATS*.

James Allan, Rahul Gupta, and Vikas Khandelwal. 2001. Temporal summaries of new topics. In *Proceedings of SIGIR*.

Tim Althoff, Xin Luna Dong, Kevin Murphy, Safa Alai, Van Dang, and Wei Zhang. 2015. TimeMachine: Timeline generation for knowledge-base entities. In *Proceedings of KDD*.

Robert Bell and Yehuda Koren. 2007. Lessons from the Netflix prize challenge. *ACM SIGKDD Explorations Newsletter*, 9(2).

James Bennett and Stan Lanning. 2007. The Netflix prize. In *Proceedings of the KDD Cup and Workshop*.

Hai Leong Chieu and Yoong Keok Lee. 2004. Query based event extraction along a timeline. In *Proceedings of SIGIR*.

Wei-Sheng Chin, Yong Zhuang, Yu-Chin Juan, and Chih-Jen Lin. 2015. A fast parallel stochastic gradient method for matrix factorization in shared memory systems. *ACM Transactions on Intelligent Systems and Technology*, 6(1).

James Conroy, Judith Schlesinger, Diane O'Leary, and Mary Okurowski. 2001. Using HMM and logistic regression to generate extract summaries for DUC. In *Proceedings of DUC*.

Nazanin Dehghani and Masoud Asadpour. 2015. Graph-based method for summarized storyline generation in Twitter. *arXiv preprint arXiv:1504.07361*.

Günes Erkan and Dragomir Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1).

Anthony Fader, Dragomir Radev, Michael Crespin, Burt Monroe, Kevin Quinn, and Michael Colaresi. 2007. Mavenrank: Identifying influential members of the

US senate using lexical centrality. In *Proceedings of EMNLP-CoNLL*.

Goran Glavaš and Jan Šnajder. 2014. Event graphs for information retrieval and multi-document summarization. *Expert Systems with Applications*, 41(15).

Ahmed Hassan, Anthony Fader, Michael Crespin, Kevin Quinn, Burt Monroe, Michael Colaresi, and Dragomir Radev. 2008. Tracking the dynamic evolution of participant salience in a discussion. In *Proceedings of COLING*.

Lifu Huang and Lian'en Huang. 2013. Optimized event storyline generation based on mixture-event-aspect model. In *Proceedings of EMNLP*.

Chris Kedzie, Kathleen McKeown, and Fernando Diaz. 2014. Summarizing disasters over time. In *Proceedings of the Bloomberg Workshop on Social Good at KDD*.

Jerry Kiernan and Evimaria Terzi. 2009. Constructing comprehensive summaries of large event sequences. *ACM Transactions on Knowledge Discovery from Data*, 3(4).

Gunhee Kim and Eric Xing. 2014. Reconstructing storyline graphs for image recommendation from web community photos. In *Proceedings of CVPR*.

Gunhee Kim, Leonid Sigal, and Eric P Xing. 2014. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Proceedings of CVPR*.

Yehuda Koren, Robert Bell, and Chris Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 8.

Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of SIGIR*.

Ju-Hong Lee, Sun Park, Chan-Min Ahn, and Daeho Kim. 2009. Automatic generic document summarization based on non-negative matrix factorization. *Information Processing & Management*, 45(1).

Omer Levy and Yoav Goldberg. 2014. Neural word embedding as implicit matrix factorization. In *Proceedings of NIPS*.

Chen Lin, Chun Lin, Jingxuan Li, Dingding Wang, Yang Chen, and Tao Li. 2012. Generating event storylines from microblogs. In *Proceedings of CIKM*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Proceedings of the ACL workshop "Text summarization branches out"*.

Qiaozhu Mei, Jian Guo, and Dragomir Radev. 2010. Divrank: the interplay of prestige and diversity in information networks. In *Proceedings of KDD*.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into texts. In *Proceedings of EMNLP*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Borja Navarro-Colorado and Estela Saquete. 2015. GPLSIUA: Combining temporal information and topic modeling for cross-document event ordering. In *Proceedings of SemEval*.

Jun-Ping Ng, Yan Chen, Min-Yen Kan, and Zhoujun Li. 2014. Exploiting timelines to enhance multi-document summarization. In *Proceedings of the ACL*.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. MaltParser: A data-driven parser-generator for dependency parsing. In *Proceedings of LREC*.

Dragomir Radev, Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Celebi, Stanko Dimitrov, Elliott Drabek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004a. MEAD – a platform for multidocument multilingual text summarization. In *Proceedings of LREC*.

Dragomir R Radev, Hongyan Jing, Małgorzata Styś, and Daniel Tam. 2004b. Centroid-based summarization of multiple documents. *Information Processing & Management*.

Benjamin Recht, Christopher Re, Stephen Wright, and Feng Niu. 2011. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Proceedings of NIPS*.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3).

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR*.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of NAACL-HLT*.

Giang Binh Tran, Mohammad Alrifai, and Dat Quoc Nguyen. 2013a. Predicting relevant news events for timeline summaries. In *Proceedings of WWW*.

Giang Binh Tran, Tuan A Tran, Nam-Khanh Tran, Mohammad Alrifai, and Nattiya Kanhabua. 2013b. Leveraging learning to rank in an optimization framework for timeline summarization. In *Proceedings of the SIGIR Workshop on Time-Aware Information Access*.

Giang Tran, Mohammad Alrifai, and Eelco Herder. 2015. Timeline summarization from relevant headlines. In *Proceedings of ECIR*.

Dingding Wang, Tao Li, Shenghuo Zhu, and Chris Ding. 2008. Multi-document summarization via sentence-level semantic analysis and symmetric matrix factorization. In *Proceedings of SIGIR*.

Dingding Wang, Tao Li, and Mitsunori Ogihara. 2012. Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs. In *Proceedings of AAAI*.

Lu Wang, Claire Cardie, and Galen Marchetti. 2015. Socially-informed timeline generation for complex events. In *Proceedings of NAACL-HLT*.

Rui Yan, Liang Kong, Congrui Huang, Xiaojun Wan, Xiaoming Li, and Yan Zhang. 2011a. Timeline generation through evolutionary trans-temporal summarization. In *Proceedings of EMNLP*.

Rui Yan, Xiaojun Wan, Jahna Otterbacher, Liang Kong, Xiaoming Li, and Yan Zhang. 2011b. Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In *Proceedings of SIGIR*.

Xin Wayne Zhao, Yanwei Guo, Rui Yan, Yulan He, and Xiaoming Li. 2013. Timeline generation with social attention. In *Proceedings of SIGIR*.

Wubai Zhou, Chao Shen, Tao Li, Shu-Ching Chen, and Ning Xie. 2014. Generating textual storyline to improve situation awareness in disaster management. In *Proceedings of the IEEE International Conference on Information Reuse and Integration*.

Weizhong Zhu and Chaomei Chen. 2007. Storylines: Visual exploration and analysis in latent semantic spaces. *Computers & Graphics*, 31(3).