Text-Image-Video Summary Generation Using Joint Integer Linear Programming

Anubhav Jangra[1], Adam Jatowt[2(✉)], Mohammad Hasanuzzaman[3], and Sriparna Saha[1]

[1] Department of Computer Science, Indian Institute of Technology Patna, Patna, India
anubhav0603@gmail.com, sriparna.saha@gmail.com
[2] Department of Social Informatics, Kyoto University, Kyoto, Japan
jatowt@gmail.com
[3] Department of Computer Science, Cork Institute of Technology, Cork, Ireland
hasanuzzaman.im@gmail.com

Abstract. Automatically generating a summary for asynchronous data can help users to keep up with the rapid growth of multi-modal information on the Internet. However, the current multi-modal systems usually generate summaries composed of text and images. In this paper, we propose a novel research problem of text-image-video summary generation (TIVS). We first develop a multi-modal dataset containing text documents, images and videos. We then propose a novel joint integer linear programming multi-modal summarization (JILP-MMS) framework. We report the performance of our model on the developed dataset.

Keywords: Multi-modal summarization · Integer Linear Programming

1 Introduction

Advancement in technology has led to rapid growth of multimedia data on the Internet, which prevent users from obtaining important information efficiently. Summarization can help tackle this problem by distilling the most significant information from the plethora of available content. Recent research in summarization [2,11,31] has proven that having multi-modal data can improve the quality of summary in comparison to uni-modal summaries. Multi-modal information can help users gain deeper insights. Including supportive representation of text can reach out to a larger set of people including those who have reading disabilities, users who have less proficiency in the language of text and skilled readers who are looking to skim the information quickly [26]. Although visual representation of information is more expressive and comprehensive in comparison to textual description of the same information, it is still not a thorough model of representation. Encoding abstract concepts like guilt or freedom [11], geographical locations or environmental features like temperature, humidity etc. via images is impractical. Also images are a static medium and cannot represent

J. M. Jose et al. (Eds.): ECIR 2020, LNCS 12036, pp. 190–198, 2020.
https://doi.org/10.1007/978-3-030-45442-5_24

dynamic and sequential information efficiently. Including videos could then help overcome these barriers since video contains both visual and verbal information. To the best of our knowledge, all the previous works have focused on creating text or text-image summaries, and the task of generating an extractive multi-modal output containing text, images and videos from a multi-modal input has not been done before. We thus focus on a novel research problem of text-image-video summary generation (TIVS). To tackle the TIVS task, we design a novel Integer Linear Programming (ILP) framework that extracts the most relevant information from the multimodal input. We set up three objectives for this task, (1) *salience within modality*, (2) *diversity within modality* and (3) *correspondence across modalities*. For preprocessing the input, we convert the audio into text using an Automatic Speech Recognition (ASR) system, and we extract the key-frames from video. The most relevant images and videos are then selected in accordance with the output generated by our ILP model.

To sum up, we make the following contributions: (1) We present a novel multi-modal summarization task which takes news with images and videos as input, and outputs text, images and video as summary. (2) We create an extension of the multi-modal summarization dataset [12] by constructing multi-modal references containing text, images and video for each topic. (3) We design a joint ILP framework to address the proposed multi-modal summarization task.

2 Related Work

Text summarization techniques are used to extract important information from textual data. A lot of research has been done in the area of extractive [10,21] and abstractive [3,4,19,23] summarization. Various techniques like graph-based methods [6,15,16], artificial neural networks [22] and deep learning based approaches [18,20,29] have been developed for text summarization. Integer linear programming (ILP) has also shown promising results in extractive document summarization [1,9]. Duan et al. [5] proposed a joint-ILP framework that produces summaries from temporally separate text documents.

Recent years have shown great promise in the emerging field of multi-modal summarization. Multi-modal summarization has various applications ranging from meeting recordings summarization [7], sports video summarization [25], movie summarization [8] to tutorial summarization [13]. Video summarization [17,28,30] is also a major sub-domain of multi-modal summarization. A few deep learning frameworks [2,11,31] show promising results, too. Li et al. [12] uses an asynchronous dataset containing text, images and videos to generate a textual summary. Although some work on document summarization has been done using ILP, to the best of our knowledge no one has ever used an ILP framework in the area of multi-modal summarization.

3 Problem Definition

Let M: $\{D_1, D_2, ..., D_{|D|}\} \bigcup \{I_1, I_2, ..., I_{|I|}\} \bigcup \{V_1, V_2, ..., V_{|V|}\}$ be a multi-modal dataset related to a topic T, where D_i is a text document, I_i is an image and V_i is

a video. |.| denotes the cardinality of a set. Each document D_i contains sentences x_j such that $D_i : \{x_{i,1}, x_{i,2}, ..., x_{i,|D_i|}\}$. Our objective is to generate a multi-modal summary $\mathrm{S} = X_{sum} \bigcup I_{sum} \bigcup V_{sum}$ such that the final summary S covers up all the important information in the original data while minimizing the length of summary, where $X_{sum} \subseteq \{x_{i,j} | x_{i,j} \in D_i \wedge D_i \in M\}$, $I_{sum} \subseteq \{I_i, I_j, ..., I_k\}$, where $|I_{sum}| \leq |I|$ and $V_{sum} \subseteq \{V_l, V_m, ..., V_{|n|}\}$, where $|V_{sum}| \leq |V|$[1].

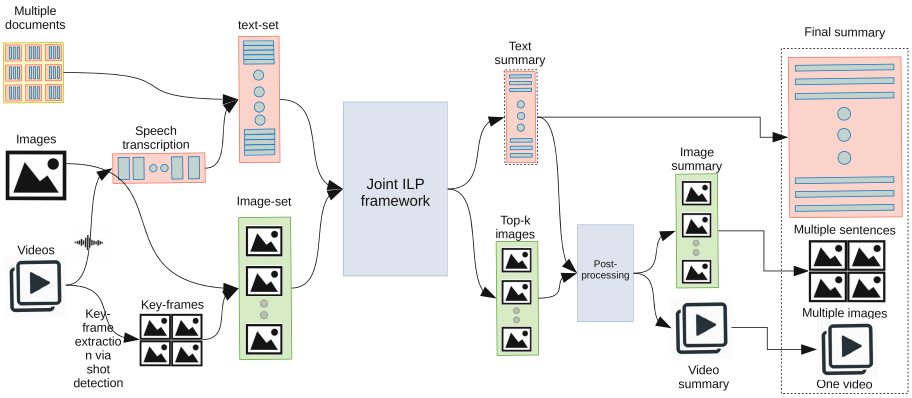

Fig. 1. The framework of our proposed model.

4 Proposed Method

4.1 Pre-processing

Each topic in our dataset comprises of text documents, images, audio and videos. As shown in Fig. 1, we firstlextract key-frames from the videos [32]. These key-frames together with images from the original data form the *image-set*. The audio is transcribed into text (IBM Watson Speech-to-Text Service: www.ibm.com/watson/developercloud/speech-to-text.html), which contributes to the *text-set* together with the sentences from text-documents. The images from then *image-set* are encoded by the VGG model [24] and the 4,096-dimensional vector from the pre-softmax layer is used as the image representation. Every sentence from the *text-set* is encoded using the Hybrid Gaussian-Laplacian Mixture Model (HGLMM) into a 6,000-dimensional vector. For text-image matching, these image and sentence vectors are fed into a two-branch neural network [27] to have a 512-dimensional vector for images and sentences in a shared space.

4.2 Joint-ILP Framework

ILP is a global optimization technique, used to maximize or minimize an objective function subject to some constraints. In this paper, we propose a joint-ILP

[1] We set $|V_{sum}| = 1$ for simplicity assuming that in many cases one video would be enough.

technique to optimize the output to have high salience, diversity and cross-modal correlation. The idea of joint-ILP is similar to the one applied in the field of across-time comparative summarization [5]. However, to the best of our knowledge, an ILP framework was not used to solve multi-modal summarization (Gurobi optimizer is used for ILP optimization: https://www.gurobi.com/).

Decision Variables. M_{txt} is a $n \times n$ binary matrix such that $m^{txt}_{i,i}$ indicates whether sentence s_i is selected as an exemplar or not and $m^{txt}_{i,j \neq i}$ indicates whether sentence s_i votes for s_j as its representative. Similarly, M_{img} is a $p \times p$ binary matrix that indicates the exemplars chosen in the image set. M_c is $n \times p$ binary matrix that indicates the cross-modal correlation. $m^c_{i,j}$ is true when there is some correlation between sentence s_i and image I_j.

Objective Function

$$\underset{max}{\text{Arg}}[\lambda \cdot m \cdot \{Sal(M_{txt}) + Sal(M_{img})\} + (1-\lambda) \cdot (k_{txt} + k_{img}) \cdot MCorr(M_c)] \quad (1)$$

$$Sal(M_{mod}) = \sum_{i=1}^{t} m^{mod}_{i,i} \cdot Imp(item_i, G(item_i)); \; i \in \{1, 2, \ldots, t\} \quad (2)$$

$$Imp(item_i, G(item_i)) = \sum_{item_j \in G(item_i)} Sim_{cosine}(item_i, item_j); \; i, j \in \{1, 2, \ldots, t\} \quad (3)$$

$$G(item_i) = \{item_j | m^{mod}_{ji} = 1; j \in \{1, 2, \ldots, t\}\}; i \in \{1, 2, \ldots, t\} \quad (4)$$

where $\langle mod, t, item \rangle \in \{\langle text, n, s \rangle, \langle img, p, I \rangle\}$ is used to represent multiple modalities together in a simple way.

$$MCorr = \sum_{i=1}^{n} \sum_{j=1}^{p} m^c_{i,j} \cdot Sim_{cosine}(s_i, I_j) \quad (5)$$

We need to maximize the objective function in Eq. 1, containing salience of text, images and cross-modal correlation. Similar to the joint-ILP formulation in [5] the diversity objective is implicit in this model. Equation 4 generates the set of entities that are a part of the cluster whose exemplar is $item_i$. The salience is calculated by Eqs. 2 and 3 by taking cosine similarity over all the exemplars with the items belonging to their representative clusters separately for each modality. The cross-modal correlation score is calculated in Eq. 5.

Constraints

$$m^{mod}_{i,j} \in \{0, 1\}; \; mod \in \{txt, img, c\} \quad (6)$$

$$\sum_{i=1}^{n} m^{txt}_{i,j} = k_{txt} \; and \sum_{i=1}^{p} m^{img}_{i,j} = k_{img} \quad (7)$$

$$\sum_{j=1}^{n} m^{txt}_{i,j} = 1; \; i \in \{1, 2, \ldots, n\} \; and \sum_{j=1}^{p} m^{img}_{i,j} = 1; \; i \in \{1, 2, \ldots, p\} \quad (8)$$

$$m_{j,j}^{mod} - m_{i,j}^{mod} \geq 0;\ mod \in \{txt, img\} \tag{9}$$

Equation 7 ensures that exactly k_{txt} and k_{img} clusters are formed in their respective uni-modal vector space. Equation 8 guarantees that an entity can either be an exemplar or be part of a single cluster. According to Eq. 9, a sentence or image must be exemplar in their respective vector space to be included in the sentence-image summary pairs. Values of m, k_{txt} and k_{img} are set to be 10, same as in [5].

4.3 Post-processing

The Joint-ILP framework outputs the text summary (X_{sum}) and *top–m* images from the *image-set*. This output is used to prepare the image and video summary.

Extracting Images

$$I_{sum} = I_{sum1} \cup I_{sum2} \tag{10}$$

$$I_{sum1} = \{I_z |\, I_z \in \ top\ m \wedge I_z \neq keyframes\} \tag{11}$$

$$I_{sum2} = \{I_z |\, \alpha \leq Sim_{cosine}(I_z, I_y) \leq \beta \wedge I_y \in I_{sum1}\} \tag{12}$$

Equation 11 selects all those images from *top10* images that are not key-frames. Assuming that images which look similar would have similar annotation scores and would help users gain more insight, the images relevant to the images in I_{sum1} (at least with α cosine similarity) but not too similar (at max with β cosine similarity) to avoid redundancy are also selected to be a part of the final image summary I_{sum} (Eq. 12). α is set to 0.4 and β is 0.8 in our experiments.

Extracting Video. For each video, weighted sum of visual (Eq. 13) and verbal (Eq. 14) scores is computed. The video with the highest score is selected as our video summary.

$$visual - score = \sum_{I_k \in KF} \sum_{I_f \in I_{sum}} Sim_{cosine}(I_f, I_k) \tag{13}$$

$$verbal - score = \sum_{x_j \in ST} \sum_{x_i \in X_{sum}} Sim_{cosine}(x_i, x_j) \tag{14}$$

where KF is the set of all key-frames and ST is the set of speech transcriptions.

5 Dataset Preparation

There is no benchmark dataset for the TIVS task. Therefore, we created our own text-image-video dataset by extending and manually annotating the multi-modal summarization dataset introduced by Li et al. [12]. Their dataset comprised of 25 new topics. Each topic was composed of 20 text documents, 3 to 9 images, and 3 to 8 videos. The final summary however was unimodal, that is, in the form of only a textual summary containing around 300 words. We then extended it by

selecting some images and a video for each topic that summarize the topic well. Three undergraduate students were employed to score the images and videos with respect to the benchmark text references. All annotators scored each image and video on a scale of 1 to 5, on the basis of similarity between the image/video and the text references (1 indicating no similarity and 5 denoting the highest level of similarity). Average annotation scores (AAS) were calculated for each image and video. The value of the minimum average annotation score for images was kept as a hyper-parameter to evaluate the performance of our model in various settings[2]. The video with the highest score is chosen to be the video component of the multi-modal summary[3].

Table 1. Overall performance comparison for textual summary using ROUGE.

System	ROUGE-1	ROUGE-2	ROUGE-l
Baseline-1	0.254	0.065	0.216
Baseline-2	**0.261**	0.068	0.225
Baseline-3	0.249	0.068	0.212
JILP-MMS	0.260	**0.074**	**0.226**

Table 2. Overall performance of image summary using precision and recall. AAS denotes here the threshold value of image summary generation.

AAS	Average precision	Average recall	Variance precision	Variance recall
5	0.016	0.060	0.003	0.048
4.5	0.099	0.084	0.067	0.041
4	0.258	0.313	0.105	**0.151**
3.5	0.335	0.332	0.111	0.125
3	**0.599**	**0.383**	**0.139**	0.076

6 Experimental Settings and Results

We evaluate the performance of our model using the dataset as described above. We use the ROUGE scores [14] to evaluate the textual summary, and based on them we compare our results with the ones of three baselines.

We use the multi-document summarization model proposed in [1]. For **Baseline-1** we feed the model with embedded sentences from all the original documents together. The central vector is calculated as the average of all the sentence vectors. The model is given vectors for sentences from the *text-set* and images from the *image-set* in the joint space for other baselines. For **Baseline-2**, the

[2] Every topic had at least one image when we set threshold for average annotation score to 3.
[3] In case two videos have the same score, the video with shorter length was chosen.

average of all the vectors is taken as the central vector. For **Baseline-3**, the central vector is calculated as the weighted average of all the sentence and image vectors. We give equal weights to text, speech and images for simplicity.

As shown in Table 1, our model produces better results than the prepared baselines in terms of ROUGE-2 and ROUGE-l scores. Table 2 shows the average precision and recall scores as well as the variance. We set various threshold values for the annotation scores to generate multiple image test sets in order to evaluate the performance of our model. We get a higher precision score for low AAS value, because the number of images in the final solution increases on decreasing the threshold values. The proposed model gave **44%** accuracy in extracting the most appropriate video (whereas random selection of images for 10 different iterations gives an average 16% accuracy).

7 Conclusion

Unlike other problems that focus on text-image summarization, we propose to generate a truly multi-modal summary comprising of text, images and video. We also develop a dataset for this task, and propose a novel joint ILP framework to tackle this problem.

Acknowledgement. Dr. Sriparna Saha would like to acknowledge the support of Early Career Research Award of Science and Engineering Research Board (SERB) of Department of Science and Technology, India to carry out this research. Mohammed Hasanuzzaman would like to acknowledge ADAPT Centre for Digital Content Technology which is funded under the SFI Research Centres Programme (Grant 13/RC/2106).

References

1. Alguliev, R., Aliguliyev, R., Hajirahimova, M.: Multi-document summarization model based on integer linear programming. Intell. Control Autom. **1**(02), 105 (2010)
2. Chen, J., Zhuge, H.: Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4046–4056 (2018)
3. Chen, Y.C., Bansal, M.: Fast abstractive summarization with reinforce-selected sentence rewriting. arXiv preprint arXiv:1805.11080 (2018)
4. Chopra, S., Auli, M., Rush, A.M.: Abstractive sentence summarization with attentive recurrent neural networks. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 93–98 (2016)
5. Duan, Y., Jatowt, A.: Across-time comparative summarization of news articles. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 735–743. ACM (2019)
6. Erkan, G., Radev, D.R.: LexRank: graph-based lexical centrality as salience in text summarization. J. Artif. Intell. Res. **22**, 457–479 (2004)

7. Erol, B., Lee, D.S., Hull, J.: Multimodal summarization of meeting recordings. In: Proceedings of the 2003 International Conference on Multimedia and Expo, ICME 2003, (Cat. No. 03TH8698), vol. 3, pp. III–25. IEEE (2003)
8. Evangelopoulos, G., et al.: Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Trans. Multimed. **15**(7), 1553–1568 (2013)
9. Galanis, D., Lampouras, G., Androutsopoulos, I.: Extractive multi-document summarization with integer linear programming and support vector regression. In: Proceedings of COLING 2012, pp. 911–926 (2012)
10. Kupiec, J., Pedersen, J., Chen, F.: A trainable document summarizer. In: Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 68–73. ACM (1995)
11. Li, H., Zhu, J., Liu, T., Zhang, J., Zong, C., et al.: Multi-modal sentence summarization with modality attention and image filtering (2018)
12. Li, H., Zhu, J., Ma, C., Zhang, J., Zong, C., et al.: Multi-modal summarization for asynchronous collection of text, image, audio and video (2017)
13. Libovickỳ, J., Palaskar, S., Gella, S., Metze, F.: Multimodal abstractive summarization for open-domain videos. In: Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL), NIPS (2018)
14. Lin, C.Y.: ROUGE: a package for automatic evaluation of summaries. In: Text Summarization Branches Out, pp. 74–81. Association for Computational Linguistics, Barcelona, July 2004. https://www.aclweb.org/anthology/W04-1013
15. Mihalcea, R.: Graph-based ranking algorithms for sentence extraction, applied to text summarization. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 170–173 (2004)
16. Mihalcea, R., Tarau, P.: TextRank: bringing order into text. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing, pp. 404–411 (2004)
17. Mirzasoleiman, B., Jegelka, S., Krause, A.: Streaming non-monotone submodular maximization: personalized video summarization on the fly. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
18. Nallapati, R., Zhai, F., Zhou, B.: Summarunner: a recurrent neural network based sequence model for extractive summarization of documents. In: Thirty-First AAAI Conference on Artificial Intelligence (2017)
19. Nallapati, R., Zhou, B., Gulcehre, C., Xiang, B., et al.: Abstractive text summarization using sequence-to-sequence RNNs and beyond. arXiv preprint arXiv:1602.06023 (2016)
20. Nallapati, R., Zhou, B., Ma, M.: Classify or select: neural architectures for extractive document summarization. arXiv preprint arXiv:1611.04244 (2016)
21. Paice, C.D.: Constructing literature abstracts by computer: techniques and prospects. Inf. Process. Manag. **26**(1), 171–186 (1990)
22. Saini, N., Saha, S., Jangra, A., Bhattacharyya, P.: Extractive single document summarization using multi-objective optimization: exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm. Knowl.-Based Syst. **164**, 45–67 (2019)
23. See, A., Liu, P.J., Manning, C.D.: Get to the point: summarization with pointer-generator networks. CoRR abs/1704.04368 (2017). http://arxiv.org/abs/1704.04368
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

25. Tjondronegoro, D., Tao, X., Sasongko, J., Lau, C.H.: Multi-modal summarization of key events and top players in sports tournament videos. In: 2011 IEEE Workshop on Applications of Computer Vision (WACV), pp. 471–478. IEEE (2011)
26. UzZaman, N., Bigham, J.P., Allen, J.F.: Multimodal summarization of complex sentences. In: Proceedings of the 16th International Conference on Intelligent User Interfaces, pp. 43–52. ACM (2011)
27. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5005–5013 (2016)
28. Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X., Yao, C.: Video summarization via semantic attended networks. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
29. Zhang, Y., Er, M.J., Zhao, R., Pratama, M.: Multiview convolutional neural networks for multidocument extractive summarization. IEEE Trans. Cybern. **47**(10), 3230–3242 (2016)
30. Zhou, K., Qiao, Y., Xiang, T.: Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
31. Zhu, J., Li, H., Liu, T., Zhou, Y., Zhang, J., Zong, C.: MSMO: multimodal summarization with multimodal output. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4154–4164 (2018)
32. Zhuang, Y., Rui, Y., Huang, T.S., Mehrotra, S.: Adaptive key frame extraction using unsupervised clustering. In: Proceedings 1998 International Conference on Image Processing, ICIP98 (Cat. No. 98CB36269), vol. 1, pp. 866–870. IEEE (1998)