Summarizing Multimedia Content

Natwar Modani[1(✉)], Pranav Maneriker[1], Gaurush Hiranandani[1], Atanu R. Sinha[1], Utpal[1], Vaishnavi Subramanian[1], and Shivani Gupta[2]

[1] BigData Experience Lab, Adobe Research, Bangalore, India
{nmodani,pmanerik,ghiranan,atr,utp,vasubram}@adobe.com
[2] Adobe India, Noida, India
shivanig@adobe.com

Abstract. Today multimedia content comprising both text and images is growing at a rapid pace. There has been a body of work to summarize text content, but to the best of our knowledge, no method has been developed to summarize multimedia content. We propose two methods for summarizing multimedia content. Our novel approach explicitly recognizes two desirable, normative characteristics of a summary - good coverage and diversity of the respective text and images, and that text and images should be coherent with each other. Two methods are examined - graph based and a modification to the submodular approach. Moreover, we propose a metric to measure the quality of a multimedia summary which captures coverage and diversity of text and images as well as coherence between the text and images in the summary. We experimentally demonstrate that the proposed methods achieve good quality multimedia summaries.

Keywords: Summarization · Text and images · Multimedia content · Algorithms

1 Introduction

Today multimedia content is growing at a rapid pace on the web. To cater to readers, publishers such as The New York Times and The Wall Street Journal offer briefings of content with text and images. The growing shift to mobile devices calls for summarizing multimedia content. Text summarization has been addressed. Our research fills a void by examining summarization of multimedia content - text and images.

The first of two formulations we propose is graph based, inspired by [13]. Each fragment of either content type is a node, the edge weight within a content type is the similarity between two fragments, and the edge weight between fragments of two different types is the coherence between them. The node weight signifies the amount of information in the fragment. The objective function includes all three properties. The second approach uses sub-modular functions, inspired by [9]. The objective function models coverage and diversity of both content types in the summary, but introduces an additional term for the coherence between these types.

W. Cellary et al. (Eds.): WISE 2016, Part II, LNCS 10042, pp. 340–348, 2016.
DOI: 10.1007/978-3-319-48743-4_27

In the absense of ground truth, information coverage and diversity of a summary are used to measure its quality. We extend this notion to images, while also incorporating the coherence of the images and text to define quality for the new concept of multimedia summary. A quality metric, labeled MuSQ (Multimedia Summary Quality) is introduced. With a small manually annotated data set, we demonstrate that the proposed metric shows better agreement with human judgement when compared to traditional metrics such as retention/compression rate and KL divergence. We then evaluate the proposed algorithms using this metric, and the experimental results show that our proposed algorithms perform better compared to the baseline methods.

2 Related Work

While text summarization has been an active area of research for several years, summarizing multimedia content is relatively unexplored. Recent work has presented a multimedia summarizer system for retrieving relevant information from web repositories based on the extraction of semantic descriptors of documents [1]. In this approach, images are not treated as primary objects, but are chosen secondarily based on the selected text summary. Notably, the content of the images is not leveraged, instead only its metadata is used, making the summary potentially less accurate.

The literature on summarization of multimedia data [3] focuses largely on video summarization. Other works [2], based on video/audio features, exploit natural language engines to create textual summaries.

For text summarization, the two broad approaches are: abstractive and extractive. In this research paper, we will be focusing on extractive summarization only.

Starting with Luhn [10] automated (text-only) extractive document summarization has been examined by researchers in Information Retrieval and Computational Linguistics [14]. Algorithms such as support vector machine (SVM) and regression models have been used. However, Wu et al. [17] found that certain graph-based algorithms (for example, TextRank [11]) perform better than SVM and regression methods.

Solving the summarization problem for product reviews, [13] proposed a graph based formulation which uses a fast and scalable greedy algorithm. They considered the informativeness and diversity of the sentences to select the summary of the reviews.

The papers mentioned above follow the bag of words approach, which rely on frequency of words in documents. In a different approach, [5] used continuous vector representations for semantically aware representations of sentences as a basis for measuring similarity. Our technology extends the approach presented in [5,13] to incorporate images in the summary.

With our goal of multimedia summary it is necessary to associate segments of text with segments of images. Approaches that describe contents of images are formulated either by mapping images to a fixed set of human-constructed

sentences [4,15], or by automatically generating novel captions [8,12]. Other approaches use Kernel Canonical Correlation Analysis [16] to align images and sentences; however their reliance on computing kernels, quadratic in number of images and sentences, make them not easily scalable. We use the framework developed by [6] to map the text and images onto a common vector space in our work.

3 Problem Definition

First, we present five desirable qualities of multimedia summary qualitatively, by extending well-established concepts in text summarization.

- The text (image) part of the summary should provide good coverage of the text (image) part of the document.
- The text (image) part of the summary should be diverse.
- The text and image part of the summary should be coherent.

We start by defining the content fragment which is either a text unit (typically, a sentence), or an image segment. The desired size of the summary images is a configuration parameter of our system. The image segments are generated as follows. First, we apply an image segmentation algorithm [7] to identify informative objects in an image. Then, each image segment is bounded by a box. This is achieved by finding the smallest rectangle parallel to boundaries that completely encloses the informative object as identified above. If the rectangle is smaller than desired size, it is merged with other image segments that overlap with it. Eventually, when the bounding rectangle is at least of the desired size, we re-size it (by zoom out) to fit the desired size. Now, each such rectangle is an image segment.

The similarity between a pair of text units (sentences) is determined by first applying a recursive auto-encoder based vector representation to both the text units and then taking the cosine similarity between the two vectors. For finding the similarity between a pair of image segments, we apply the deep learning based CNN (convolutional neural network) technique [6] to transform images into a vector of size 4096, and then assess the cosine similarity between these two vectors. To find the similarity between a text unit and an image segment, we apply the transformation to project them into a common vector space [6] and then we compute the cosine similarity between the vectors representing the image and the text.

3.1 Graph Based Approach

In this approach, (inspired by [13]), we construct a graph to represent the document. Each node represents a content fragment. We draw an edge between two nodes, representing two content fragments, with the edge weight as their similarity. We also assign a reward to each content fragment. A text unit is assigned the reward score as the number of nouns, adverbs, adjectives, verbs and half of

the number of pronouns. An image fragment is assigned the reward score based on the information content. We take the image segment reward as the average level of similarity with all other image segments.

We attach a cost to each content fragment. The cost of a text fragment is taken in units of sentences, word or characters, and the cost of an image segment is taken as one unit, as all image segments are resized to the desired level. The user also specifies the upper limit on the size of summary for the text and image parts separately, called as budget for the text and image parts, respectively, and represented as b_T and b_I.

We follow an iterative greedy strategy [13] to select the content fragments to include in the summary. In particular, we find the gain G_i of including an available content fragments i in the summary, given by:

$$G_i = \sum_{j=1}^{n} w_{ij} * R_j + \sum_{k=1}^{m} \hat{w}_{ik} * \hat{R}_k \tag{1}$$

Here, w_{ij} is the edge weight between the i^{th} content fragment and j^{th} text unit, and $\hat{w}_{ik}$ is the edge weight between the i^{th} content fragment and k^{th} image segment. Further, R_j is the reward of the j^{th} text unit, and $\hat{R}_k$ is the reward for the k^{th} image segment.

Then we find the content fragment, with the maximum gain to cost ratio, and include it in the summary. Note that we do not impose any order while choosing the text and image fragments for the summary, although the number of text units and image segments selected are controlled by the individual budgets for those two parts of the summary. When a content fragment is included in the summary, the rewards for all other content fragments are updated, per following rules. If a content fragment is the same type as the selected content fragment, its rewards is multiplied by $(1 - w_{ij})$, and if the content fragment in question is of a different type compared to the selected content fragment, its reward is multiplied by $(1 + w_{ij})$. This ensures diversity because the value of including another content fragment that is similar and of the same type as the summary is reduced. At the same time, coherence is achieved since the value of including a content fragment that is similar but of a different type is increased.

3.2 Coverage-Diversity Based Approach

In this approach, inspired by the sub-modular approach to text summarization [5], we have a five part objective function. We have a text coverage term, and a text diversity reward term. Along similar lines, we define the image coverage term, and an image diversity reward term. Finally, we define a coherence term which captures the similarity between text and image(s) selected in the summary. For document D, we denote the summary of the text T as S and of the images V as I. The objective function is defined as

$$F(S, I) = \alpha_1 * C_T(S) + \alpha_2 * R_T(S) + \alpha_3 * C_V(I) + \alpha_4 * R_V(I) + \alpha_5 * H(S, I) \tag{2}$$

Here, α's represent the weights which can be tuned by the user.

The term $C_T(S)$ represents the coverage of the text T of the document by the summary text S, defined in the same way as [5]

$$C_T(S) = \sum_{i \in T} min\{\sum_{j \in S} w_{ij},\ \alpha \sum_{j \in T}\{w_{ij}\}\} \tag{3}$$

The term $R_T(S)$ is the reward for diversity of the text summary S with respect to the text of the document, defined in the same way as [9]

$$R_T(S) = \sum_{i \in S} \sqrt{\sum_{j \in P_i \cap S} r_j} \quad \text{where} \quad r_j = \frac{1}{n} \sum_{i \in T} w_{ij} \tag{4}$$

where P_i is a partition of the ground set T into separate clusters and r_j is the singleton reward of including sentence j in the empty summary. The clustering is done using CLUTO with the 4096 sized vector representation of the sentences derived from [5] with number of clusters as 0.2 times the number of sentences (so, on average, each cluster would have 5 sentences), a direct K-mean clustering algorithm is used following the same choice as made in [9]. The term r_j is defined again in the same manner as [9] where n is the number of sentences in T and w_{ij} is the similarity between sentences i and j. By replacing T with V and S with I, we can define the corresponding terms for images and their summary.

The term $H(S, I)$ represents the coherence between the summary text and summary images. It is defined as the sum of all pairs of text units and image fragments, i.e.,

$$H(S, I) = \sum_{i \in S} \sum_{j \in I} \hat{w}_{ij}$$

here, $\hat{w}_{ij}$ represents the similarity between the text fragment i in the text part of the summary and image fragment j in the image part of the summary.

4 Multimedia Summary Quality

Measuring quality of a summary relative to its original source is important. Since the problem of multimedia summarization has not been addressed, no quality metrics have been proposed. We propose *MuSQ*, or *Multimedia Summary Quality*, which includes the desirable characteristics stated in Sect. 3. This metric does not require ground truth.

Let the similarity between a content fragment (text or image) u and another content fragment v be given by $Sim(u, v)$. Consider a text sentence v present in the document text T and a sentence u in the summary text S.

Now consider a metric μ_T defined as

$$\mu_T = \sum_{v \in T} R_v * \max_{u \in S} \{Sim(u, v)\} \tag{5}$$

The term $\max_{u \in S} Sim(u, v)$ represents the maximum level of similarity between a sentence v in the document text and any sentence in the summary S.

Recall that the term R_v is the reward value of the sentence v, and contribution of the sentence v towards the quality of the summary is accordingly $R_v * \max_{u \in S} Sim(u, v)$. Note that due to the *max* function, if there are two sentences which are similar to the given sentence v, it will not lead to enhanced contribution of the sentence to the quality of the summary. On the other hand, if the summary is having a sentence similar to a sentence in the document, it leads to increase in the metric value for the summary quality. In this way, the function μ_T is able to simultaneously capture the diversity and the information content of the summary with respect to the text of the original document T.

We define the overall quality metric *MuSQ* as:

$$\mu_M = \mu_T + \mu_I + \sigma_{T,I} \tag{6}$$

$$\mu_I = \sum_{w \in V} \hat{R}_w * \max_{x \in I} \{Sim(w, x)\} \tag{7}$$

$$\sigma_{T,I} = \sum_{v \in S} \sum_{w \in I} \{Sim(v, w) * R_v * \hat{R}_w\} \tag{8}$$

The terms μ_T and μ_I are diversity aware information coverage measure for the text part and the image part of the summary, respectively. The third term $\sigma_{T,I}$ measures the degree of cohesion between the text and the image part of the summary, as the sum of similarities between the sentences and the images in the summary, across all pairs.

5 Experimental Results

Now, we describe experimental results to validate our algorithms, as well as the proposed quality metric. First, on a small dataset we check whether the quality metric *MuSQ* correlates well with human judgment about the quality of multimedia summary, since obtaining human input for a large dataset is very expensive. Once *MuSQ* is validated, it is used to evaluate the proposed summarization algorithms on a larger dataset.

The small dataset comprised ten articles from the *New York Times* for each of which we created two summaries. In a survey, participants were shown the original article, the two summaries and were asked which one of the two summaries was better, or whether they were almost of similar quality. To control the order effect, the summaries were randomly placed first or second (without regard to their *MuSQ* scores), and the participants were not given any information about how the summaries were generated.

We define agreement level in three different ways. The first definition treats the 'Equal' option as half agreement and half disagreement, i.e., $AL1 = 100 * (A+0.5E)/(A+E+D)$, where $AL1$ is the agreement level according to definition 1, A is number of agreements (i.e., the human judge preferred the summary which had higher *MuSQ* score), E is number of times both summaries were deemed to be of same quality by the human judge, and D is the number of disagreements (i.e., the human judge preferred the summary which had lower *MuSQ* score).

Second definition treats the 'Equal' option as disagreement, i.e., $AL2 = 100 * A/(A + E + D)$. Third definition ignores the 'Equal' option completely, i.e., $AL3 = 100 * A/(A + D)$.

In total, 22 human judges provided 128 responses. Out of these, 87 responses favoured summaries with higher *MuSQ* scores, whereas 14 responses found the summaries to be almost equal in quality. The remaining 27 responses disagreed with the ranking based on the *MuSQ* scores. This translates to 68 % agreement for the *MuSQ* scores (where, as a conservative approach 'equal' is classified as a disagreement), and 76 % agreement ignoring the votes for 'equal'. The Pearson correlation coefficient between the agreement levels (AL1, AL2 and AL3) and the fractional difference in the *MuSQ* scores is approximately 0.51 for all the three definitions of agreement levels, which shows that our proposed quality metric correlates well with human judgment.

Now, we describe the experiments performed on a larger dataset, considering *MuSQ* as the quality metric. We collected 1,000 articles from *New York Times*, which typically have text and images, both. We kept only those articles which had at least 20 and at most 100 sentences, and at least 1 image. This resulted in selecting 703 articles for the experiment. Further, the size of the summary was specified as 3 sentences and 1 image of size $200 * 200$ pixels.

The image segmentation algorithm takes the number of objects to be identified as input. We choose to identify 20 objects, with a further constraint that each class of objects does not occur more than 10 times. This ensures that the objects from a general class, such as background, do not end up as the only objects in the segments. Also, we used [9] to compute the similarity between two sentences. The similarity between two images, as well as, between a text sentence and an image was computed in the same way as [6].

We evaluated the two approaches proposed in this paper using the *MuSQ* score. As a baseline, we used the text only version of these two algorithms for finding the three summary sentences, and augmented this summary with the first (whole) image from the article (hitherto only known method). The graph based approach we propose achieves the highest score 539 times, and the coverage-diversity based approach achieves the highest score 90 times. Only 103 times out of 703 articles, one of the two baseline approaches outperform our proposed approaches, and 587 times our proposed approaches outperform the baseline approaches. This means that our proposed approaches are better 83.5 % of the times and equally good another 1.5 % of the times. As the *MuSQ* scores are dependent on the size of the original document, it is not appropriate to compare them across articles.

We also report the traditional text only performance metric for the summary quality for the four algorithms in Table 1, as well as the newly proposed metric *MuSQ*. As expected, the *MuSQ* score is higher for the enhanced versions compared to the baseline methods. One finds that both for retention rate and KL-Divergence, the baseline approaches perform better than the enhanced approaches, which are to be expected. However, note that the performance degradation is fairly small and less severe for the graph based approach. Hence, the algorithms proposed by us provide significant value for summarizing multimedia content.

Table 1. Quality metric for the four approaches (retention rate and KL-Divergence are measured only for the text part of the summary)

Metric	Enhanced approaches		Baseline approaches	
	Submodular	Graph based	Submodular	Graph based
MuSQ	1528.37	1592.18	1519.95	1564.78
Retention rate	0.3704	0.4608	0.3896	0.4652
KL-Divergence	1.2052	0.8980	1.0822	0.8725

6 Conclusion

Today multimedia content in the form of text and images are commonplace across publishing sites and devices. The need for the summarization of such content to comprise both text and images is stronger than ever before. The results provide strong evidence in support of our proposed methods and validate the new quality metric. These summaries are better than the summaries generated only using text part and then adding the first image, which is the only known multimedia summary method. We hope that future work will advance our understanding and knowledge in multimedia summarization to parallel that of text summarization.

References

1. dAcierno, A., Gargiulo, F., Moscato, V., Penta, A., Persia, F., Picariello, A., Sansone, C., Sperl, G.: A multimedia summarizer integrating text and images. In: Intelligent Interactive Multimedia Systems and Services, pp. 21–33. Smart Innovation, Systems and Technologies (2014)
2. Ding, D., Metze, F., Rawat, S., Schulam, P.F., Burger, S.: Generating natural language summaries for multimedia. In: Proceedings of the Seventh International Natural Language Generation Conference, pp. 128–130. Association for Computational Linguistics (2012)
3. Ding, D., Metze, F., Rawat, S., Schulam, P.F., Burger, S., Younessian, E., Bao, L., Christel, M.G., Hauptmann, A.: Beyond audio and video retrieval: towards multimedia summarization. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, p. 2. ACM (2012)
4. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: generating sentences from images. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 15–29. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15561-1_2
5. Kageback, M., Mogren, O., Tahmasebi, N., Dubhashi, D.: Extractive summarization using continuous vector space models. In: Proceedings of the 2nd Workshop on Continuous Vector Space Models and their Compositionality (CVSC), pp. 31–39. EACL (2014)
6. Karpathy, A., Joulin, A., Fei-Fei, L.: Deep fragment embeddings for bidirectional image sentence mapping. Archive, Cornell University Library (2014). http://arXiv.org/abs/1406.5679

7. Krähenbühl, P., Koltun, V.: Geodesic object proposals. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 725–739. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10602-1_47
8. Kulkarni, G., Premraj, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Baby talk: understanding and generating simple image descriptions. In: CVPR (2011)
9. Lin, H., Bilmes, J.: A class of submodular functions for document summarization. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, HLT 2011, Stroudsburg, PA, USA, vol. 1, pp. 510–520 (2011)
10. Luhn, H.: The automatic creation of literature abstracts. IBM J. Res. Dev. **2**(2), 159–165 (1958)
11. Mihalcea, R.: Language independent extractive summarization. In: ACLdemo, pp. 49–52 (2005)
12. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., Hal Daum, I.: Midge: generating image descriptions from computer vision detections. In: EACL (2012)
13. Modani, N., Khabiri, E., Srinivasan, H., Caverlee, J.: Graph based modeling for product review summarization. In: WISE (2015)
14. Nenkova, A., McKeown, K.: A survey of text summarization techniques. In: Aggarwal, C.C., Zhai, C.X. (eds.) Mining Text Data, pp. 43–76. Springer, New York (2012)
15. Ordonez, V., Kulkarni, G., Berg, T.L.: Im2text: describing images using 1 million captioned photographs. In: NIPS (2011)
16. Socher, R., Fei-Fei, L.: Connecting modalities: semi-supervised segmentation and annotation of images using unaligned text corpora. In: CVPR (2010)
17. Wu, J., Xu, B., Li, S.: An unsupervised approach to rank product reviews. In: FSKD, pp. 1769–1772 (2011)