Multi-Modal Supplementary-Complementary Summarization using Multi-Objective Optimization

Anubhav Jangra
Indian Institute of Technology Patna, India
anubhav0603@gmail.com

Sriparna Saha
Indian Institute of Technology Patna, India
sriparna.saha@gmail.com

Adam Jatowt
University of Innsbruck, Austria
adam.jatowt@uibk.ac.at

Mohammad Hasanuzzaman
Munster Technological University, Ireland
hasanuzzaman.im@gmail.com

ABSTRACT

Large amounts of multi-modal information online make it difficult for users to obtain proper insights. In this paper, we introduce and formally define the concepts of supplementary and complementary multi-modal summaries in the context of the overlap of information covered by different modalities in the summary output. A new problem statement of combined complementary and supplementary multi-modal summarization (CCS-MMS) is formulated. The problem is then solved in several steps by utilizing the concepts of multi-objective optimization by devising a novel unsupervised framework. An existing multi-modal summarization data set is further extended by adding outputs in different modalities to establish the efficacy of the proposed technique. The results obtained by the proposed approach are compared with several strong baselines; ablation experiments are also conducted to empirically justify the proposed techniques. Furthermore, the proposed model is evaluated separately for different modalities quantitatively and qualitatively, demonstrating the superiority of our approach.

CCS CONCEPTS

• **Information systems** → **Summarization**; **Multimedia and multimodal retrieval**.

KEYWORDS

multi-modal summarization; multi-objective optimization; data driven summarization; grey wolf optimizer

ACM Reference Format:

Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammad Hasanuzzaman. 2021. Multi-Modal Supplementary-Complementary Summarization using Multi-Objective Optimization. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3404835.3462877

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462877

1 INTRODUCTION

Sharing information via the Internet has become the most popular means of transferring information. Thanks to technological advancements, people have effective means to share content in multi-media formats. However, the vast plethora of public opinion and facts on a topic makes it difficult to access the crux of a topic. Motivated by this, we propose an approach to summarize multi-modal information on a certain topic and to output summary in a multi-modal format.

Prior studies have shown that multi-modal output containing images and text increases user satisfaction by 12.4% as compared to single modality output such as text only summary [48]. We also conduct a human evaluation based experiment (Sec. 6.2), where we find that having visual cues in the summary helps improve the overall satisfaction by 22%, and makes the topic 19% more fascinating, as well as helps users better understand the overall information by a factor of 14.5%. Having information in multiple modalities can cater to the needs of a more diverse community, including users that are less proficient in textual language (e.g., non-native speakers), users having problems comprehending textual information, or adroit users trying to skim through the topic [42].

Multi-modal content makes the information representation more appealing, and may help captivate users' interests. Having a well-designed multi-modal summary becomes necessary, because of constraints of each modality to express some ideas. For example, abstract concepts such as freedom, affection, depression, gravity etc. cannot be well expressed with visual media alone, whereas describing an object such as a tarsier with words is also difficult if the user is unaware of what a tarsier actually looks like. Hence there is a need of coherence across modalities, with some constraints, depending on the themes that a modality can describe.

When dealing with multi-modal information retrieval tasks, the extent to which a particular modality contributes to the final output might differ from other modalities. Amongst the modalities, there is often a preferable mode of representation based on the significance and ability to fulfill the task. We denote these preferred modalities as *key modalities* or *central modalities* (will be referred to as *central modalities* from here onwards). The other modalities help assist the central modalities in fulfilling the desired task, and are known as *adjacent modalities*. The adjacent modalities can enhance the user experience by either supplementing or by complementing the information represented via the central modality. When these

adjacent modalities reinforce the facts and ideas presented in central modality, the enhancement is known as *supplementary* enhancement. On the other hand, when these adjacent modalities complete the central modality, by providing additional or alternate information that is relevant, albeit not covered by the central modality, the enhancement is known as *complementary* enhancement (refer to Section 3 for the formal definitions).

Figure 1: An illustration of supplementary (top) and complementary (middle) enhancements for a sample text (bottom). The complementary enhancement in the middle contains the key-frames of a video and its speech transcription.

In the case of multi-modal summarization, we consider text as the central modality, and visual form of information (images and videos) as adjacent modalities[1]. The ability of text to express events in details is the reason why text has been the preferred mode of conveying information, be it newspapers, legal documents, bills, or books. This is why in our research we use text as the central modality, and we enhance the output text with complementary and supplementary visual modalities.

In this paper we propose a novel combined complementary and supplementary multi-modal summarization (CCS-MMS) problem, where the task is to output a multi-modal summary, such that the adjacent modalities have a mixture of both complementary and supplementary type enhancements. We believe that a multi-modal summary is incomplete without both the complementary and supplementary enhancements (as illustrated by our experiments in Section 6.2). For instance, the sample multi-modal summary presented in Figure 1 contains information that helps the user better relate to the overall incident along with information that improves his/her overall understanding of the situation, fulfilling the agenda of a summary of news event. Similarly, consider a multi-modal summary of a soccer game; the text would cover the major events including injuries, players scoring goals etc. but the video highlights of the event would also contain some great chances, passes and dribbles, that the textual summary would not include, necessitating a multi-modal summary comprising both complementary and supplementary enhancements. Another example could be given for movie summarization, where the textual part of the final summary focuses on the movie plot, while the visual enhancements bring forth the aesthetics in the summary.

Even though there is a need of both kinds of enhancements in a summary, it is possible that in some cases the users would appreciate one kind over the other. For instance, in case of summarization of a forensic report, supplementary information supporting the central summary covering the major incidents would be the ideal multi-modal summary, since it will contain the evidence as part of the visual enhancements. Therefore, in this work, we propose a generic framework capable of generating different ratios of enhancements in the output by tweaking the hyper-parameter ζ (Sec. 4.3). In order to generalize the model, we use asynchronous data, i.e., data having no alignment amongst different modalities. We work in an unsupervised setting, since data annotation is costly as it requires a close reading of many documents and watching large multi-media input (images and videos).

We propose a generic population-based system to solve the CCS-MMS problem. The motivation of using population-based techniques instead of other single-point optimization strategies include - a) creating a multi-modal summary involves optimization for conflicting objectives, and population-based techniques can handle that by utilizing a multi-objective optimization framework, whereas the single-point optimization strategies involve combining the objectives, which leads to sub-optimal solutions; b) population-based strategies are capable of generating diverse set of solutions in a single run, which generates multiple unique solutions for the users to select from; c) population-based techniques are able to generate satisfactory output due to their meta-heuristic behavior. We later verified through our experiments that the average scores of the generated populations were higher than ones of the existing state-of-the-art approaches, and the best solutions were able to improve these scores by a great margin (refer to Section 6.1).

The major contributions of this work are:

- To the best of our knowledge, we are the first to introduce and formally define the concepts of complementary and supplementary enhanced summaries containing text, images and videos.
- We introduce a novel problem of combined complementary and supplementary multi-modal summarization (CCS-MMS), which takes text, images and videos as input, and outputs text, complementary and supplementary images, as well as supplementary videos as summary.
- We create an extension of a multi-modal summarization dataset [22] by augmenting the output summary with supplementary and complementary images and videos.
- We propose a novel multi-objective optimization (MOO) framework to solve the CCS-MMS task. The framework is kept generic and any MOO technique can be used as the underlying optimization strategy. In this work, we illustrate the strength of this framework by using the Grey Wolf Optimizer [28][2].

[1]Text modality can also be considered as an adjacent modality (for instance, in case of micro-blog summarization). We leave this as a future work.

[2]The proposed model is flexible, and can output a variable size summary, i.e. having a variable number of sentences and multiple image and video enhancements.

2 RELATED WORK

A lot of work have been done in the area of text summarization, including extractive [20, 33] and abstractive [11, 12, 30, 34] summarization techniques. Various methods have been used to solve the problem of extractive text summarization, including integer linear programming [10], genetic algorithms [25, 32], graph-based techniques [7, 26, 27], deep learning [9, 29, 31], etc. Thanks to the recent development of sequence-to-sequence RNNs [39], there has been a lot of research in the area of abstractive summarization [4, 5, 15, 30]. These text summarization techniques have been used in diverse tasks such as micro-blog summarization [35], query-based summarization [13], timeline summarization [38, 41], comparative summarization [6], medical report summarization [1, 46], etc. Other than text summarization, researchers have also explored the areas of image summarization [36, 45], and video summarization [44, 47]. Recent years have also shown growth in the area of multi-modal summarization. Research work has been done in unsupervised [16, 22], and supervised [21, 48, 49] multi-modal summarization approaches. All the recent research in supervised multi-modal summarization [3, 8, 21, 48, 49] involve deep learning frameworks. Our approach belongs to the unsupervised type of models[3]. These models include a Joint Integer Linear Programming framework by Jangra et al. [16] that optimizes weighted average of uni-modal salience and cross-modal correspondence. A linear combination of submodular functions (salience of text, redundancy and visual coverage in this case) under a budget constraint to obtain near-optimal solutions at a sentence level was proposed by Li et al. [22], and a multi-objective optimization framework that uses Genetic Algorithms as the optimization strategy was introduced by Jangra et al. [17].

Even though significant work has been done in the domain of multi-modal summarization, to the best of our knowledge, no one has explored the area of complementary enhanced multi-modal summarization, neither combined complementary and supplementary multi-modal summarization styles.

3 PROBLEM DEFINITION

Given a topic T = $\{D_1, D_2, ..., D_{|D|}\} \bigcup I \bigcup V$, where D_j is the j^{th} document such that $D_i = \{s_1^i, s_2^i, ..., s_{|D_i|}^i\}$ where s_n^m is the n^{th} sentence of the m^{th} document, $I = \{i_1, i_2, ..., i_{|I|}\}$ is the set of input images such that i_k is the k^{th} image, and $V = \{v_1, v_2, ..., v_{|V|}\}$ is the set of input videos such that v_k is the k^{th} video[4], the task is to generate a combined complementary and supplementary multi-modal summary (CCS-MMS).

Defining complementary and supplementary enhancements: Taking text as the central modality, and images and videos as adjacent modalities, we define complementary and supplementary enhancements as follows:

Complementary visual enhancement: When the adjacent modality completes the central modality, by contributing alternate, but relevant information, that adjacent modality is a complementary visual enhancement.

Supplementary visual enhancement: When the adjacent modality reinforces the facts and ideas presented in the central modality that adjacent modality is a supplementary visual enhancement.

We use two helper functions to help formally define these enhancements:

Quality estimation function: $Q : (X_{mod}, T_{sum}) \mapsto [0, 1]$ takes as input an entity to be evaluated for enhancement (from the set X_{mod}) subject to a central modality based summary (text summary in our case, T_{sum}), and gives a quality score ranging from 0 to 1 based on relevance to the text summary; 0 being no relevance, and 1 being highly relevant.

Diversity estimation function: $Dvr : (X_{mod}, T_{sum}) \mapsto [0, 1]$ takes as input an entity to be evaluated for enhancement subject (from the set D_{mod}) to a central modality based summary (text summary in our case), and gives a diversity score ranging from 0 to 1 based on information overlap with the text summary; 0 denoting high overlap (i.e., not diverse), and 1 denoting no overlap (highly diverse).

Using the quality estimation function (Q) we first define an *enhancement candidate*, and then we categorize it into supplementary-type or complementary type enhancements.

Enhancement candidate: An item itm in the adjacent modality input that satisfies the condition $Q(itm, T_{sum}) > \psi$ is considered as an *enhancement candidate*, where ψ is the quality screening threshold.

Categorizing enhancement candidates: Given an *enhancement candidate* C, we can classify it into supplementary or complementary using Eq. 1.

$$Categorization(C, T_{sum}) = \begin{cases} supplementary & \text{if } Dvr(C, T_{sum}) < \phi_s \\ complementary & \text{if } Dvr(C, T_{sum}) > \phi_c \end{cases} \quad (1)$$

where ϕ_c and ϕ_s are the diversity thresholds[5].

In CCS-MMS task, the output must contain all three modalities, i.e., text, image and videos, where text covers the entire input, while the images and the videos together enhance the central modality, such that the output has both supplementary and complementary enhancements. The proposed task is extractive, and the output is of the format $Summary = T_{sum} \bigcup I_{sum} \bigcup V_{sum}$, where $T_{sum} = \{s_1^{sum}, s_2^{sum}, ..., s_{|T_{sum}|}^{sum}\}$ is the textual part of final summary comprising of sentences from input documents, $I_{sum} = \{i_1^{sum}, i_2^{sum}, ..., i_{|I_{sum}|}^{sum}\}$ is a subset of input images, and $V_{sum} = \{v_1^{sum}, v_2^{sum}, ..., v_{|V_{sum}|}^{sum}\}$[6] is the video enhancement of the text summary.

The image enhancement, I_{sum} comprises of both complementary and supplementary visual enhancements, whereas the video is only selected as the supplementary enhancement. The intuition behind keeping video as supplementary is that the role of a video in multi-modal information is to give users a wholesome idea of the topic. A user might skip the text and just prefer to watch video to get a grasp on the topic, but that is rather not likely to happen in case of images. One can even argue that text and images

[3] To make fair comparison with our work, we will evaluate the performance of our system juxtaposed to the other unsupervised models.

[4] |.| denotes the cardinality of a set

[5] For simplicity, we set in this work $\phi_c = \phi_s$. However, different values can be used depending on the task requirements.

[6] We set $|V_{sum}| = 1$ in our case, because we assume that in most situations one video would be enough.

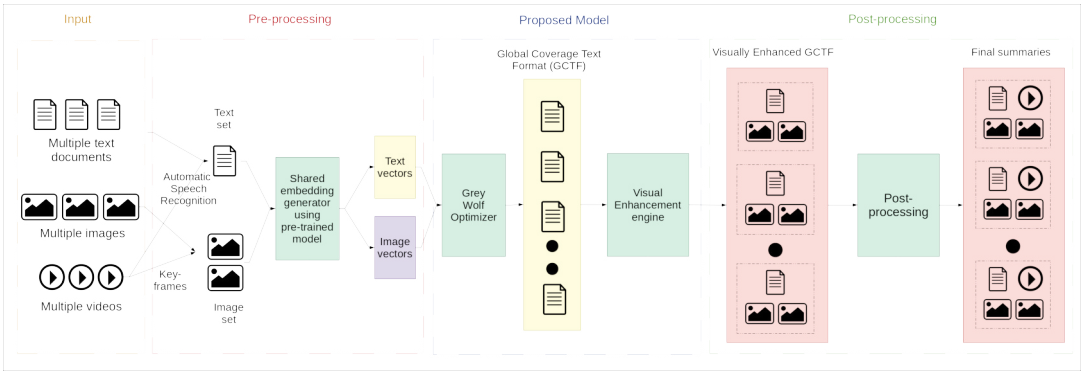

Figure 2: Proposed model architecture.

are the most relevant in a multi-modal summary, while the video based enhancements are more like an add-on. Keeping the video enhancement only supplementary is based on our intuition and is a debatable point; our model however is flexible enough to perform both kinds of video-based enhancements.

4 PROPOSED MODEL

We propose a multi-objective optimization based technique to tackle the CCS-MMS problem. We divide our proposed model into two sections[7]: (1) Global Coverage Text Format (GCTF) and (2) Visual Enhancement of Text Summaries (VETS). For GCTF, we use Grey Wolf Optimizer to create multiple text summaries by optimizing multiple objectives. These text summaries are then visually enhanced before the post-processing step, where some images and a video are selected based on these enhancements. Our model outputs multiple multi-modal summaries, containing text, images and videos. Algorithm 2 is provided to give an overview of our method and to facilitate its understanding.

4.1 Pre-processing

The proposed model takes as input multiple text documents, images, videos and audio. The audio from videos are transcribed into text[8], and the key-frames are extracted from videos using the shot boundary detection algorithm [50]. The speech transcriptions together with the input text document constitute our *text-set*, and the key-frames along with input images form the *image-set*. In order to quantify the similarity amongst different modalities, we apply a shared embedding model to get a joint representation of the information present in *image-set* and *text-set*. We encode the sentences in *text-set* using Hybrid Gaussian-Laplacian mixture model (HGLMM) [18], and then apply PCA to reduce the dimensions to get 6,000-dimensional sentence-level embedding. We use a VGG-19 [37] model pre-trained on the ImageNet dataset [19] to encode the images into 4096-dimensional vectors[9]. The multi-layered two-branch neural network [43] is then used to project these high dimensional uni-modal embeddings to a shared 512-dimensional vector space. These projection vectors are then used to compute the similarity overlap amongst sentence-image, sentence-sentence and image-image pairs in our main model[10].

4.2 Global Coverage Text Format (GCTF)

In this step, multiple text summaries are produced using a multi-objective optimizer. Note that the proposed method is very generic in nature and any optimizer can be utilized for solving this. In the recent literature it was shown that GW converges faster as compared to other meta-heuristic optimization techniques like GSA, DE, PSO, EP, ES [28]. Therefore, it is utilized as the underlying optimization strategy in the current framework. In order to improve the interaction between text and image, and to overcome the shortcomings and limitations of the shared embedding generator (Fig. 2), we propose a significance weighting factor to compute the cross-modal similarities[11].

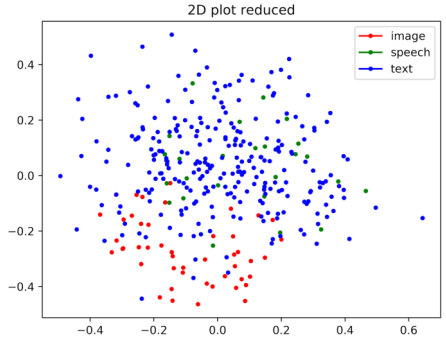

Figure 3: Projection of joint representation of text sentences, speech transcriptions and images from the image-set.

Significance weighting factor (SWF).

We observed that there is an innate affinity of similar modalities to stick together in the shared embedding space, because of the relative closeness of similar kinds of input representation formats leading to unsatisfactory mixing (see in Fig. 3). Furthermore, even though all the sentences and image vectors participate in the clustering

[7]We note that even though the model is divided into two sections, it is still an end-to-end system, viz. it produces the desired output given the corresponding multi-modal input in one shot. Since the techniques we use are unsupervised, there is no training phase.

[8]Using IBM Watson Speech-to-Text Service: www.ibm.com/watson/developercloud/speech-to-text.html

[9]More recent CNN based models [14, 40] could be used to encode the images, but for a fair comparison with other baselines, we used VGG model instead.

[10]The choice of pre-processing is orthogonal to the proposed task; the mentioned techniques were selected to mimic the existing work so as to assure an impartial model comparison.

[11]Cosine similarity of embeddings in shared-embeddings space is used for the intra-modal similarity as will be shown in Eq. 3.

process, not all images are equally capable of covering the overall content of the topic. Thus, we propose a *significance weighting factor* (W_j (Eq. 2) to give higher weights to the images that are more relevant to the overall theme. The weighting factor is considered when computing the sentence-image semantic overlap, otherwise plain cosine similarity is used to compute the overlap amongst images or sentences.

$$CMS(i_j, s_k) = sim_{cos}(i_j, s_k) \times W_j \tag{2}$$

$$W_j = \frac{\sum_{x \in \{RSS\}} sim_{cos}(i_j, x)}{\sum_{x \in \{RSS\}} sim_{cos}(O_{txt}, x)} \tag{3}$$

where CMS is Cross-Modal Similarity, i_j is the j^{th} image, s_k is the k^{th} sentence, O_{txt} is the central text vector computed by averaging all the text vectors, and *RSS* is the set of randomly selected sentences, which is a collection of σ randomly selected sentence vectors. σ is called the *computation vs quality trade-off factor*, since the higher the σ value, the better the coverage of the entire text domain, but with higher computation requirements. This weighting factor is unbiased to any sub-part of the topic, is flexible in terms of computational requirement, and is normalized using an impartial normalizing factor ($\sum_{x \in \{RSS\}} sim_{cos}(O_{txt}, x)$). We also define a *Balancing Factor (BF)* in Eq. 4 to balance out the minority modalities (i.e., the modality with lesser data inputs) in order to give equal attention to all modalities while performing the global coverage[12].

$$BF = 1 + (\tau - 1) \times \frac{(|I| + |S|)}{(2 \times |I| \times |S|)} \tag{4}$$

where $|I|$ is the number of images in *image-set*, $|S|$ is the number of sentences in *text-set*, and τ is a hyper-parameter for choosing the degree of emphasis to minority modality. The proposed *BF* is directly proportional to the gap between different modalities, and thus is able to balance out the difference in the number of elements in each modality.

We incorporate the balancing factor to our cross-modal similarity function as below:

$$CMS(i_j, s_k) = sim_{cos}(i_j, s_k) \times W_j \times BF \tag{5}$$

Grey Wolf Optimizer.

Grey wolf optimizer is a nature inspired optimization technique, based on the strategy wolves use while hunting for prey. Grey wolves hunt in packs, and each wolf pack has a leader, called the $\alpha - wolf$, closely aided by the $\beta - wolf$, an advisor to the $\alpha - wolf$ who is most likely to be the successor of the $\alpha - wolf$. To manage a large pack, there are also $\delta - wolves$ responsible for managing small groups of the $\omega - wolves$, the lowest in the wolf hierarchy.

The initial population (P) is produced by applying k-medoid clustering using *kmeans++* seeding [2] over the entire shared embedding space to generate $|P|$ solutions with diverse cluster sizes (lines 6-14 in Algorithm 2). The objective functions for each solution are calculated, and non-dominated sorting is applied to attain multiple fronts (line 16 in Algorithm 2). These diverse solutions are then fed into our grey-wolf optimizer [28], where the *Archive* is initialized with the Pareto-optimal solutions (rank-1 solutions, that are non-dominated to each other and dominate every other solution). α, β, δ wolves are selected from the Archive using the Roulette Wheel Selection mechanism [24] such that $\alpha \neq \beta \neq \delta$ (line 17 in Algorithm 2). If the number of Pareto optimal solutions is less than three, then solutions from the next front are considered, until distinct α, β, δ wolves get selected. Eqs. 6 to 9 are used to update the positions of each wolf under the guidance of α, β, δ wolves (lines 18-26 in Algorithm 2).

$$\vec{x}(t+1) = \frac{\vec{x_1}(t) + \vec{x_2}(t) + \vec{x_3}(t)}{3} \tag{6}$$

$$\vec{x_1}(t) = \vec{x_\alpha}(t) - \vec{A_1} \cdot \vec{D_\alpha} \quad \& \quad \vec{x_2}(t) = \vec{x_\beta}(t) - \vec{A_2} \cdot \vec{D_\beta} \quad \& \quad \vec{x_3}(t) = \vec{x_\delta}(t) - \vec{A_3} \cdot \vec{D_\delta} \tag{7}$$

$$\vec{D_\alpha} = |\vec{C_1} \cdot \vec{x_\alpha}(t) - \vec{x}(t)| \quad \& \quad \vec{D_\beta} = |\vec{C_1} \cdot \vec{x_\beta}(t) - \vec{x}(t)| \quad \& \quad \vec{D_\delta} = |\vec{C_1} \cdot \vec{x_\delta}(t) - \vec{x}(t)| \tag{8}$$

$$\vec{A_i} = 2\vec{a} \cdot \vec{r_1} - \vec{a} \quad \& \quad \vec{C_i} = 2 \cdot \vec{r_2}; \quad i \in \{1, 2, 3\} \qquad \vec{a} = 2\left(1 - \frac{g}{g_{max}}\right) \tag{9}$$

where $\vec{x}(t+1)$ is the new solution position for solution $\vec{x}(t)$, $\vec{x_\alpha}(t)$, $\vec{x_\beta}(t)$, and $\vec{x_\delta}(t)$ are the position vectors for α, β, and δ wolves, respectively, $\vec{r_1}$ and $\vec{r_2}$ are random vectors in $[0, 1]$, g is the current generation number, and g_{max} is the maximum generation length.

In Eq. 9, $\vec{A_i}$ is the factor responsible for handling exploration and exploitation. If $|\vec{A_i}| > 1$, wolf diverges from the prey (optimal solution) leading to exploration, whereas if $|\vec{A_i}| < 1$, then the wolf converges to the prey, exploiting the information gathered so far. Term $\vec{a}$ in Eq. 9 ensures that exploration takes place in the earlier stages of training phase while exploitation in the later half of the training, ensuring the convergence of the model.

Objective Functions

We use three different objective functions, global salience (Eq. 10), global redundancy (Eq. 11) and heterogeneity (Eqs. 12 and Eqs. 13 [13]) ('global' here means covering all the modalities). We propose a separate objective of heterogeneity to ensure the formation of well-mixed heterogeneous clusters.

$$Sal = \underset{max}{\text{Arg}} \sum_{c_j} \sum_{x_k^{mod} \in cluster(j)} sim_{cos}(c_j, x_k) \tag{10}$$

$$Red = \underset{min}{\text{Arg}} \sum_{c_j} \sum_{c_i \, i \neq j} sim_{cos}(c_j, c_i) \tag{11}$$

$$Het = \underset{min}{\text{Arg}} \sum_{c_j} f(c_j) \tag{12}$$

$$f(c_j) = \begin{cases} \frac{max(I_{c_j}^{num}, S_{c_j}^{num})^p}{min(I_{c_j}^{num}, S_{c_j}^{num})} & \text{if } min(I_{c_j}^{num}, S_{c_j}^{num}) \neq 0 \\ \infty & \text{otherwise} \end{cases} \tag{13}$$

where c_j is the j^{th} cluster, and $cluster(i)$ returns the elements in i^{th} cluster, $I_{c_j}^{num}, S_{c_j}^{num}$ are the number of images and sentences in cluster c_j, respectively, and p is a hyper-parameter to enforce the degree of heterogeneity in clusters[14].

[12] In our dataset, images are the minority modality as the number of sentences in *text-set* are 7-10 times the number of images in the *image-set*.

[13] For implementation purpose, ∞ is considered as a very large number.

[14] For simplicity, the value of p is set to 1.

After each generation, non-dominated sorting is applied, and *Archive* is updated using the rank-1 non-dominated solutions. After the completion of this step, we get multiple global coverage text summaries (line 27 in Algorithm 2).

Algorithm 1: Enhancement Allocation Algorithm.

Input: text_summary, image clusters
Output: List of supplementary and complementary images.

```
for c_j in image_cluster_centers do
    nearestSentence =
      getNearestElement(c_j,text_summary);
    sentConnection[nearest-sentence] += 1;
end
ghostSentences = list of s_i s.t. 'sentConnection[s_i] == 0';
for s_i in ghostSentences do
    c_j = getNearestElement(s_i, image_cluster_centers);
    imageConnections[c_j] += 1;
end
for c_j in image_cluster_centers do
    if imageConnections[c_j] > 0 then
        supplementary_list.append(c_j);
    end
    else
        complementary_list.append(c_j);
    end
end
```

4.3 Visual Enhancement of Text Summaries (VETS)

In this step, the GCTF summaries are visually enhanced by selecting a few complementary and supplementary images from the *image-set*. We propose a one-shot population based technique to enhance a text summary. In this approach, we initialize $|Q|$ different solutions for each text summary, and then apply k-medoid clustering to each solution belonging to the *image-set*, with random cluster sizes (lines 31-36 in Algorithm 2). This is done to maintain a data-driven approach, such that the most effective clustering is able to enhance the text summary. Since one image can either be complementary or supplementary with respect to the text summary's content, we apply the Algorithm 1 to assign an *enhancement tag* to each image cluster (line 36 in Algorithm 2). Two assumptions are made as the basis of this algorithm: 1) all the cluster members would have the same kind of *enhancement tag* as the cluster representative, 2) the cosine similarity is able to capture the semantic overlap, as well as the capability of a data point to convey the information in other data point. The algorithm first detects the *ghost sentences*, the sentences that are not very well expressible in the visual format. Then for every *non-ghost sentence*, the nearest images are deemed to be of the supplementary kind, since the likelihood of information overlap between these image-sentence pairs is high. The rest of the images are determined to be of the complementary kind.

To evaluate the quality of a cluster center, we use Eqs. 14 and 15. By including the factor of cluster size, we are able to include the intra-relevance of an image, since if a cluster center belongs to a larger cluster, multiple images would endorse that image, and hence it would have a higher relevance.

$$Score_{comp}(c_j) = (1 - sim_{cos}(c_j, NS(c_j))) \times CS(c_j) \quad (14)$$

$$Score_{supp}(c_j) = sim_{cos}(c_j, NS(c_j)) \times CS(c_j) \quad (15)$$

where $NS(C_j)$ is a function that returns the nearest sentence to cluster center c_j, and $CS(c_j)$ returns the cluster size whose cluster center is c_j.

To select a single enhancement from these $|Q|$ solutions, we define an *Enhancement Selection Quotient (ESQ)* in Eq. 16, and select the solution with the highest ESQ value (lines 39-40 in Algorithm 2).

$$ESQ = \zeta \sum_{c_j} Score_{comp}(c_j) + (1 - \zeta) \sum_{c_j} Score_{supp}(c_j) \quad (16)$$

where ζ is a hyper-parameter to manage the degree of complementary enhancement in final summary.

4.4 Post-processing

In order to select a few supplementary and complementary images from the visually enhanced text summaries, we first sort each cluster on the basis of $score_{supp}$ and $score_{comp}$. Then for each enhancement type, if the cluster center is not a key-frame, then that image is given a higher priority of getting selected as a part of the final enhancement. If the minimum number of images for an enhancement is not met, then the images closer to the key-frame cluster centers are selected in the order of enhancement scores until sufficient images are selected. The video with the best average score of *visualScore* (Eq. 17) and *verbalScore* (Eq. 18) is chosen to be a part of the final summary.

$$visualScore(V_i) = \sum_{kf_j^i \in V_i} sim_{cos}(kf_j^i, clusterCenter(kf_j^i)) \quad (17)$$

$$verbalScore(V_i) = \sum_{x_j \in ST(V_i)} sim_{cos}(x_i, NS(x_i)) \quad (18)$$

where V_i is the i^{th} video, kf_j^i is the j^{th} key-frame of i^{th} video, $clusterCenter(kf_j^i)$ returns the cluster center that kf_j^i belongs to, $ST(V_i)$ provides the speech transcriptions for video V_i, and $NS(x_i)$ returns the nearest sentence to vector x_i.

5 EXPERIMENTS

5.1 Dataset

There is no benchmark dataset for the proposed CCS-MMS task. Therefore, we create our own complementary and supplementary enhanced multi-modal dataset[15], by extending the multi-modal summarization dataset introduced by Li et al. [22]. There are 25 topics in the dataset, all in the news domain. Each topic contains 20 text documents (all from different news sources), 3 text summaries, from 3 to 9 images, and from 3 to 8 videos, with an average video length of 197 seconds. Each text summary is of an extractive type, and has an upper word limit of 300 words. We had to limit our dataset size to 25 topics because the extension of existing dataset was a laborious task taking approximately 60 hours for each annotator to complete the annotation process. Other than that, the size of the dataset is similar to the recent relevant works [16, 17, 22]. Three annotators were hired to annotate the images

[15] https://github.com/anubhav-jangra/CCS-MMS-dataset

Algorithm 2: Pseudo-code for the entire process.

```
// Step-1: pre-processing (refer to Section 4.1)
txt_vecs, img_vecs = preProcess(vets); // returns text and image sets in shared embedding space
// Step-2: GCTF (refer to Section 4.2)
// population initialization
P = [];
for i in population_size do
    c_size = rand();
    curr_sol = special_kmedoid(txt_vecs, img_vecs, c_size);
    // special_kmediod uses Eq. 5 as the cross-modal similarity function, and ensures that the text sentences become the cluster centers
    evaluate_objective_values(curr_sol);
    // using Eqs. 10 to 13
    P.append(curr_solution);
end
// Grey-Wolf Optimizer
Apply non-dominated sorting on P and initialize Archive with Pareto optimal solutions;
Select leaders α, β, δ from archive such that α ≠ β ≠ δ;
for epoch = 1 to max_generation do
    for a_solution in P do
        // update the position using Eqs. 6 to 9
        update_position(a_solution);
        evaluate_objective_values(a_solution);
    end
    Apply non-dominated sorting on new wolves and update the Archive with Rank-1 solutions;
    Select leaders α, β, δ from archive such that α ≠ β ≠ δ;
end
text_sums = get_pareto_sols(Archive);
// Step-3: Visual enhancement (refer to Section 4.3)
vets = [];
for a_sum in text_sums do
    visual_enhanced_sums = [];
    // Q: no of visual enhancements for i = 1 to |Q| do
        c_size = rand();
        // perform clustering on image
        img_clstr = kmediod(img_vecs, c_size);
        assign tags to each cluster center in img_clstr using Algorithm 1;
        visual_enhanced_sums.append(img_clstr);
    end
    // select the best enhancement using Eq. 16
    best_sum = max(ESQ(visual_enhanced_sums));
    vets.append(best_sum);
end
// Step-4: Post-processing (refer to Section 4.4)
// postProcess selects the images and video for each solution using Eqs. 14 to 18
return postProcess(vets);
```

and videos, and thereby to extend the dataset to be used for our task. Inter-annotator agreement score of 0.8 was observed for the task of annotating the images and videos. In order to create the supplementary enhancement for these text summaries, each annotator was required to give a score to every image and video in the topic ranging from 1 to 5, based on their similarity with the text references. Subsequently, average annotation scores were calculated, and the top-2 images that contribute to the supplementary image output were selected, along with the top video as the video part of the multi-modal summary. Similarly, the annotators were asked to assign a relevance score (1 to 5) and a difference score (1 to 5) to each image, and then a combined *relevance-difference score* was calculated as the product of individual scores. The top-2 images with the highest *relevance-difference score* were selected to form the complementary image output. In case of disagreement, we looked at individual difference scores.

5.2 Experimental Settings

In this section, we discuss the hyper-parameters used in the experiments. For most of the hyper-parameters, standard values have been used to assure fair comparison. The values of the hyper-parameters were selected after thorough empirical testing. When computing the individual image weights for *Significance weighting factor*, the value for *Computation vs quality trade-off factor (σ)* was set to be $|text - set|/5$, keeping in mind that a larger value leads to increased computation, whereas a lower value would have an impact on the quality of weights assigned. The value of τ was set to 2, to avoid giving a very high boost to noisy images, which leads to an inferior global coverage. For all the population based techniques, we create a population of 20 solutions, and train it for 20 generations. In order to explore the degree of complementary and supplementary mix in our dataset, we ran our proposed model with different ζ values to provide better insight about the dataset. We finally selected ζ to be 0.5 to generate a well-balanced summary. It took approximately 110 minutes per topic to train the model[16].

5.3 Compared Methods

In this section we briefly explain the compared methods.

Visually Enhanced TextRank (VE-TextRank): We use the graph-based technique proposed by [26] to evaluate the quality of text summary[17]. To compare image and video scores, we apply the VETS (Sec. 4.3) to get a visual enhancement to the summary produced by TextRank algorithm[18].

Image Match: A greedy summarization technique [22] is proposed to generate the text summary considering all the multi-modal aspects. Of the multiple variations of multi-modal summarization techniques proposed in [22], image match seems to be the most promising, and is thus chosen to compare our results with[19].

Joint integer linear programming based multi-modal summarization (JILP-MMS): In [16], an integer linear programming

[16]The training was performed on a HP Spectre-x360 system integrated with Intel i7 7th Gen processor and 16 GB DDR3 RAM

[17]The Gensim library's implementation is used: https://radimrehurek.com/gensim/summarization/summariser.html

[18]Since TextRank is not a population based framework, it returns a single summary. We only use it to compare the average visual enhancement scores of our population based techniques.

[19]Due to the unavailability of ROUGE R-L score in [22], we only report the ROUGE R-1 and R-2 scores.

based multi-modal summarization model is proposed to solve the text-image-video summary (TIVS) generation. Even though the visual enhancement of TIVS task differs from CCS-MMS task, the text summaries in both the tasks deal with global coverage of entire multi-modal input.

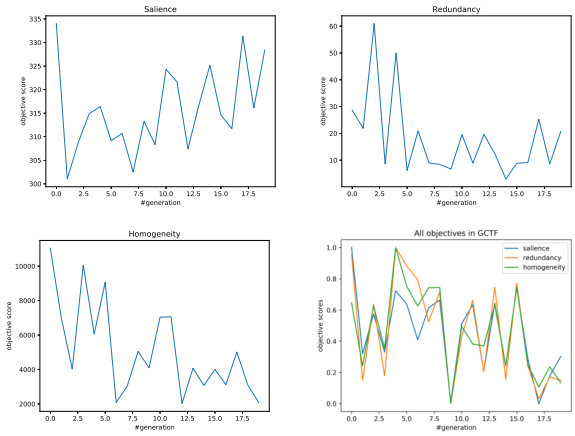

Figure 4: Plots of objectives for sample topics over multiple iterations, Salience (top left), Redundancy (top right), and Heterogeneity (bottom left). X-axis denotes number of generations, and Y-axis denotes the value of objective functions. A plot of all the objectives normalized to same scale for a particular text summary over the course of multiple iterations (bottom right).

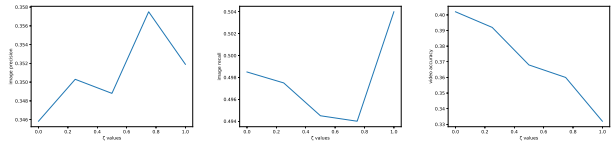

Figure 5: Plots of image precision (left), image recall (middle) and video accuracy (right) scores for different ζ values.

Summarization-based multi-modal multi-objective optimization (Summ MMS-MOO): Multi-modal summarization is solved using a multi-objective optimization based differential evolution framework, where multiple summarization objectives like intra-modal salience, intra-modal redundancy and inter-modal correspondence are optimized simultaneously to solve TIVS task [17].

Clustering-based multi-modal multi-objective optimization (Clus MMS-MOO): To tackle the TIVS task, [17] proposed a clustering-based multi-modal summarization technique using multi-objective optimization model, where a Differential Evolution is used to optimize various conflicting cluster-validity indices and cross-modal correspondence.

Similarity-based visual enhancement of text summaries (Sim-VETS): This is the only baseline that is used to evaluate the visual scores. We apply this enhancement strategy to the main model, i.e. *Sal-Red-Het MM-SWF*, as it achieves the highest scores in text evaluation. In this baseline, we rank each image on the basis of its summation with each sentence in text summary, and select the top images to be the supplementary enhancement images, and the bottom images to be complementary enhancement images in the final summary. In order to select the best video, visual scores for a video are calculated as the summation of similarity of all key-frames with sentences in the final text summary.

Salience-Redundancy based uni-modal summarization (Sal-Red UMS): To illustrate the benefit of using multiple modalities over plain text summarization, we perform the GCTF (Sec. 4.2) using the sentences from multiple documents in the input, with text salience and text redundancy as the objectives. *Significance weighting factor (SWF)* is not used.

Salience-Redundancy based multi-modal summarization (Sal-Red MMS): In this model, we use the global salience and global redundancy objectives to generate the GCTF (Sec. 4.2) and then apply the visual enhancement using the VETS approach (Sec. 4.3). *Significance weighting factor (SWF)* was not used in in this baseline.

Salience-Redundancy-Heterogeneity based multi-modal summarization (Sal-Red-Het MMS): In this model, we use the global salience, global redundancy and heterogeneity as the objectives to generate the GCTF (Sec. 4.2) and then apply the visual enhancement using the VETS approach (Sec. 4.3). *Significance weighting factor (SWF)* was not used here.

6 RESULTS

6.1 Quantitative results

To evaluate the quality of our summaries, we compare the quality of each modality of the generated summary separately. For text, we use the ROUGE R-1, R-2, and R-l precision scores [23]; for image we calculate the image precision (IP) and image recall (IR); and since we have a single video as output, we use accuracy (here it is abbreviated as the video accuracy (VA) score)[20]. Since we use a population based technique, we report both the average and the best scores. Table 1 and Table 2 illustrate the best and average ROUGE scores, and Table 3 and Table 4 give the best and average visual enhancement scores. It is important to note that although all the baselines are used to compare our textual scores, only the scores of *Sim-VETS, VE-TextRank, Sal-Red UMS, Sal-Red MMS, Sal-Red-Het MMS* are reported to compare the capability of visual enhancement of our proposed model, since the other models were not formulated for the *combined complementary and supplementary multi-modal summarization (CCS-MMS)* task. The proposed model *Salience-Redundancy-Heterogeneity based multi-modal summarization using Salience Weighting Factor (Sal-Red-Het MMS-SWF)* achieves higher average score than the existing state-of-the-art models in all the three modalities, namely text, images and videos (Tables 3 & 4). Since the grey wolf optimizer is a meta-heuristic technique, the average scores attained by different solutions on final Pareto optimal front give a better insight as to the quality of the model than the best scores. Even if ROUGE R-1 scores for *Sal-Red UMS* are better as compared to *Sal-Red MMS*, the ROUGE R-2 and R-l scores for *Sal-Red MMS* are better than the *Sal-Red UMS* as shown in Table 2[21], which supports the use of multiple modalities instead of text only input as a guidance strategy when constructing the GCTF (Section 4.2). Increasing the level of the

[20]Note that we choose to adhere to the standard metrics used in document summarization as well as ones applied in the previous works on multi-modal summarization. The proposal of a new evaluation metric for the CCS-MMS task is out of scope for this work.

[21]Meaning a higher overlap of larger sections with gold standard summaries, making R-2 and R-L better metrics to evaluate the performance

Table 1: ROUGE scores for evaluation of the text summaries.

Model	Rouge R-1	Rouge R-2	Rouge R-l
VE-TextRank	0.312	0.117	0.273
Image Match	0.422	0.133	-
JILP-MMS	0.260	0.074	0.226
Summ MMS-MOO	0.405	0.194	0.370
Clus MMS-MOO	0.420	0.167	0.390
Sal-Red UMS	0.519	0.214	0.404
Sal-Red MMS	0.503	**0.256**	**0.473**
Sal-Red-Het MMS	0.544	0.229	0.433
Sal-Red-Het MMS-SWF	**0.556**	0.228	0.452

Table 2: Avg. ROUGE scores for pop. based techniques.

Model	Rouge R-1	Rouge R-2	Rouge R-l
Summ DE-MMS-MOO	0.296	0.088	0.264
Clus MMS-MOO	0.299	0.081	0.266
Sal-Red UMS	0.423	0.095	0.291
Sal-Red MMS	0.348	0.101	0.309
Sal-Red-Het MMS	0.427	0.157	0.322
Sal-Red-Het MMS-SWF	**0.434**	**0.185**	**0.338**

Table 3: Results for visual enhancement of text summary.

Model	Image Precision	Image Recall	Video Accuracy
Sim-VETS	0.590	0.610	0.400
Sal-Red UMS	0.604	0.700	0.60
Sal-Red MMS	0.614	0.690	**0.640**
Sal-Red-Het MMS	0.571	0.710	**0.640**
Sal-Red-Het MMS-SWF	**0.620**	**0.720**	**0.640**

Table 4: Average results for visual enhancements.

Model	Image Precision	Image Recall	Video Accuracy
Sim-VETS	0.343	0.401	0.342
VE-TextRank	0.309	0.430	0.360
Sal-Red UMS	0.301	0.405	0.326
Sal-Red MMS	0.339	0.473	0.346
Sal-Red-Het MMS	0.329	0.470	0.344
Sal-Red-Het MMS-SWF	**0.348**	**0.495**	**0.368**

heterogeneity in the clusters while generating the GCTF turned out to be fruitful, as indicated by the fact that the *Sal-Red-Het MMS* baseline performs better than the *Sal-Red MMS* baseline (Table 2). The use of the proposed *Significance Weighting Factor (SWF)* brings around a significant improvement in the quality of text summaries, too (Table 2).

Fig. 4 (bottom row) illustrates the values of different objective functions normalized to the same scale obtained over the course of the entire training. The figure illustrates the conflict present in the different objective functions, since salience is a maximization objective, and redundancy and heterogeneity are minimization objectives. However, the plots show that decreasing the redundancy decreases the heterogeneity, yet it also decreases the salience of the clusters. Fig. 5 illustrates the (IP), (IR) and (VA) scores for multiple ζ values. We observe that the video score constantly decreases with the increase in ζ, which supports the idea that along with the increase in the degree of complementary enhancement, we have a lower accuracy for the supplementary videos in the final summary. We can also see a classic precision-recall trade-off as we approach the center of the ζ-curve from either side - precision increases whereas recall decreases. Since our overall model is designed to enhance a text summary in both supplementary and complementary ways, we get a higher precision for ζ in the middle, achieving a higher precision score. A central value of ζ is chosen as precision is preferred over recall in summarization tasks. Fig. 4 (top row) shows the learning of different objectives over the training course for sample topics.

6.2 Qualitative evaluation of generated summaries

In order to evaluate the quality of summaries prepared by our main model, we also conduct a human evaluation. We hire two human annotators, and report the average scores in Tab. 5. We conduct an experiment for 10 different topics, selecting the text summary with the highest ROUGE R-2 scores along with its complementary and supplementary image enhancement for $\zeta = 0.5$. The quality of each kind of enhancement is evaluated separately, and to remove any biases, we split our complementary and supplementary enhanced summaries into two equal parts, each containing 5 complementary and 5 supplementary enhanced summaries, namely *Part-1* and *Part-2*. We divide this experiment into three phases, with phase 1 being the scoring of text summaries, and phases 2 and 3 being the scoring of *Part-1* and *Part-2*, respectively. For phase 1, we ask the users to score the following aspects on a scale of 1 to 5: understanding of information present in text summary (Score-A), degree of satisfaction (Score-B), overall score (Score-C). For phases 2 and 3, we also ask the user to rate the richness which the visual enhancements bring to the textual summary (Score-D).

Table 5: Human evaluation of multi-modal summaries.

Format	Score-A	Score-B	Score-C	Score-D
Text summary	3.60	3.66	3.60	
Complementary enhanced text summary	4.20	4.35	4.50	4.30
Supplementary enhanced text summary	4.05	4.15	4.30	4.10

We observe that having any kind of visual enhancement improves the overall score by 22% (25% in case of complementary and 19.4% in case of supplementary) when compared with a plain text summary, while improving the understanding by 14%[22]. Users are 19% more satisfied by the content in summaries when there is any visual enhancement to the summary, 22% in the case of complementary and 16.9% in the case of supplementary kind of enhancement. We also observe that the users find complementary enhancement more enriching than the supplementary enhancement by 4.8%. This experiment shows that having both kinds of enhancements is important to produce a wholesome summary, and thus confirms our motivation for the CCS-MMS task.

7 CONCLUSION

In this paper, we formally defined complementary and supplementary enhanced summaries, and we proposed a novel combined complementary and supplementary multi-modal summarization (CCS-MMS) task. To solve this task, we created a dataset, and proposed a multi-objective optimization based model, that surpasses state-of-the-art unsupervised multi-modal frameworks. Although we explored the framework's potential using only the grey wolf based optimizer, we note that the proposed framework is generic and hence adaptable to different settings. We also compared our proposed model with several baselines, and conducted a human evaluation on the prepared summaries.

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.

[22]We use the average scores of both kinds of enhancements when reporting the scores of visually enhanced summaries in comparison with text only scores.

REFERENCES

[1] Stergos Afantenos, Vangelis Karkaletsis, and Panagiotis Stamatopoulos. 2005. Summarization from medical documents: a survey. *Artificial intelligence in medicine* 33, 2 (2005), 157–177.

[2] David Arthur and Sergei Vassilvitskii. 2006. *k-means++: The advantages of careful seeding*. Technical Report. Stanford.

[3] Jingqiang Chen and Hai Zhuge. 2018. Abstractive Text-Image Summarization Using Multi-Modal Attentional Hierarchical RNN. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4046–4056.

[4] Yen-Chun Chen and Mohit Bansal. 2018. Fast Abstractive Summarization with Reinforce-Selected Sentence Rewriting. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 675–686. https://doi.org/10.18653/v1/P18-1063

[5] Sumit Chopra, Michael Auli, and Alexander M Rush. 2016. 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*. 93–98.

[6] Yijun Duan and Adam Jatowt. 2019. Across-Time Comparative Summarization of News Articles. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. ACM, 735–743.

[7] Günes Erkan and Dragomir R Radev. 2004. Lexrank: Graph-based lexical centrality as salience in text summarization. *Jour. of artif. intel. res.* 22 (2004), 457–479.

[8] Xiyan Fu, Jun Wang, and Zhenglu Yang. 2020. Multi-modal Summarization for Video-containing Documents. *arXiv preprint arXiv:2009.08018* (2020).

[9] Xiyan Fu, Jun Wang, Jinghan Zhang, Jinmao Wei, and Zhenglu Yang. 2020. Document summarization with vhtm: Variational hierarchical topic-aware mechanism. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 7740–7747.

[10] Dimitrios Galanis, Gerasimos Lampouras, and Ion Androutsopoulos. 2012. Extractive multi-document summarization with integer linear programming and support vector regression. In *Proceedings of COLING 2012*. 911–926.

[11] Min Gui, Junfeng Tian, Rui Wang, and Zhenglu Yang. 2019. Attention Optimization for Abstractive Document Summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 1222–1228. https://doi.org/10.18653/v1/D19-1117

[12] Min Gui, Zhengkun Zhang, Zhenglu Yang, Yanhui Gu, and Guandong Xu. 2018. An Effective Joint Framework for Document Summarization. In *Companion Proceedings of the The Web Conference 2018*. 121–122.

[13] Johan Hasselqvist, Niklas Helmertz, and Mikael Kågebäck. 2017. Query-based abstractive summarization using neural networks. *arXiv preprint arXiv:1712.06100* (2017).

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[15] Anubhav Jangra, Raghav Jain, Vaibhav Mavi, Sriparna Saha, and Pushpak Bhattacharyya. 2021. Semantic Extractor-Paraphraser based Abstractive Summarization. *arXiv preprint arXiv:2105.01296* (2021).

[16] Anubhav Jangra, Adam Jatowt, Mohammad Hasanuzzaman, and Sriparna Saha. 2020. Text-Image-Video Summary Generation Using Joint Integer Linear Programming. In *European Conference on Information Retrieval*. Springer, 190–198.

[17] Anubhav Jangra, Sriparna Saha, Adam Jatowt, and Mohammad Hasanuzzaman. 2020. Multi-Modal Summary Generation using Multi-Objective Optimization. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1745–1748.

[18] Benjamin Klein, Guy Lev, Gil Sadeh, and Lior Wolf. 2014. Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation. *arXiv preprint arXiv:1411.7399* (2014).

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*. 1097–1105.

[20] Julian Kupiec, Jan Pedersen, and Francine Chen. 1995. A trainable document summarizer. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 68–73.

[21] Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 2018. Multi-modal Sentence Summarization with Modality Attention and Image Filtering.. In *IJCAI*. 4152–4158.

[22] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2017. Multi-modal summarization for asynchronous collection of text, image, audio and video. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 1092–1102.

[23] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. https://www.aclweb.org/anthology/W04-1013

[24] Adam Lipowski and Dorota Lipowska. 2012. Roulette-wheel selection via stochastic acceptance. *Physica A: Statistical Mechanics and its Applications* 391, 6 (2012), 2193–2196.

[25] Marina Litvak, Mark Last, and Menahem Friedman. 2010. A new approach to improving multilingual summarization using a genetic algorithm. In *Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics, 927–936.

[26] Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*. 170–173.

[27] Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*. 404–411.

[28] Seyedali Mirjalili, Seyed Mohammad Mirjalili, and Andrew Lewis. 2014. Grey wolf optimizer. *Advances in engineering software* 69 (2014), 46–61.

[29] Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[30] Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Guİ‡lçehre, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*. Association for Computational Linguistics, Berlin, Germany, 280–290. https://doi.org/10.18653/v1/K16-1028

[31] Ramesh Nallapati, Bowen Zhou, and Mingbo Ma. 2016. Classify or select: Neural architectures for extractive document summarization. *arXiv preprint arXiv:1611.04244* (2016).

[32] Maxime Peyrard and Judith Eckle-Kohler. 2016. A general optimization framework for multi-document summarization using genetic algorithms and swarm intelligence. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. 247–257.

[33] Naveen Saini, Sriparna Saha, Anubhav Jangra, and Pushpak Bhattacharyya. 2019. Extractive single document summarization using multi-objective optimization: Exploring self-organized differential evolution, grey wolf optimizer and water cycle algorithm. *Knowledge-Based Systems* 164 (2019), 45–67.

[34] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 1073–1083. https://doi.org/10.18653/v1/P17-1099

[35] Beaux P Sharifi, David I Inouye, and Jugal K Kalita. 2014. Summarization of twitter microblogs. *The computer journal* 57, 3 (2014), 378–402.

[36] Vasu Sharma, Akshay Kumar, Nishant Agrawal, Puneet Singh, and Rajat Kulshreshtha. 2015. Image summarization using topic modelling. In *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 226–231.

[37] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[38] Julius Steen and Katja Markert. 2019. Abstractive Timeline Summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. 21–31.

[39] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*. 3104–3112.

[40] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1–9.

[41] Giang Tran, Mohammad Alrifai, and Eelco Herder. 2015. Timeline summarization from relevant headlines. In *European Conference on Information Retrieval*. Springer, 245–256.

[42] Naushad UzZaman, Jeffrey P Bigham, and James F Allen. 2011. Multimodal summarization of complex sentences. In *Proceedings of the 16th international conference on Intelligent user interfaces*. ACM, 43–52.

[43] Liwei Wang, Yin Li, and Svetlana Lazebnik. 2016. Learning deep structure-preserving image-text embeddings. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5005–5013.

[44] Huawei Wei, Bingbing Ni, Yichao Yan, Huanyu Yu, Xiaokang Yang, and Chen Yao. 2018. Video summarization via semantic attended networks. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[45] Hongliang Yu, Zhi-Hong Deng, Yunlun Yang, and Tao Xiong. 2014. A joint optimization model for image summarization based on image content and tags. In *Twenty-eighth AAAI conference on artificial intelligence*.

[46] Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to Summarize Radiology Findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*. 204–213.

[47] Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[48] Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal Summarization with Multimodal Output. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4154–4164.

[49] Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9749–9756.

[50] Yueting Zhuang, Yong Rui, Thomas S Huang, and Sharad Mehrotra. 1998. Adaptive key frame extraction using unsupervised clustering. In *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, Vol. 1. IEEE, 866–870.