Abstract

Automatic summarization is one of the basic tasks both in Natural Language Processing – text summarization – and in Computer Vision – video summarization. Multimodal summarization builds a bridge between those two fields. The idea of multimodal summary is a rather flexible one. Depending on the types of input (and output) modalities, custom modeling and evaluation techniques are required. In this thesis we focus on a text-centric multimodal summarization, requiring the textual modality to be present both in the input and in the summary. We present our results regarding the multi-modal feature extraction, task-specific pre-training and intra-video similarities. We propose a human evaluation framework for assessing the multimodal summary quality. We sketch our future research plans regarding the automatic evaluation techniques for multimodal output.

1 Introduction

As per Oxford English Dictionary:\footnote{https://www.oxfordlearnersdictionaries.com}

- **modality** (pl. modalities) – the particular way in which something exists, is experienced, or is done
- **summary** (pl. summaries) – a short statement that gives only the main points of something, not the details

Deep learning based methods have become a de facto go-to solution for a variety of machine learning applications. In 2015 the ResNet-152 model by He et al. (2016a) achieved a 4.49% classification error rate on ImageNet (Deng et al., 2009) validation set, outperforming human performance of 5.1% (Russakovsky et al., 2015). In 2018 a machine translation system by Popel et al. (2020) outperformed professional translators on isolated sentences in WMT2018 News Translation Task (Bojar et al., 2018).

With such progress both in vision and language, a significant interest of the research community is now directed towards multimodal challenges that combine linguistic and visual information. In this thesis we focus on a task of Multimodal Summarization that aims at fusing disjoint information from several sources (modalities) and distilling them into a concise and precise summary.

The task of automatic creation of multimodal summaries can also be motivated by real needs in today’s digital world. According to Eurostat (Eurostat), in 2021 80% of individuals in the EU accessed the internet on a daily basis. A recent study (Uswitch) reported that the average UK citizen spends up to 6.4 hours a day on the internet, of which 1.8 hours are spend on social media. With hundreds of thousands of hours of video content and millions of articles uploaded to the internet every day, methods that automatically filter, summarize and recommend the content are necessary. Automatic multimodal data processing methods, such as multimodal summarization, are thus beneficial for everyone. This applies both from the perspective of an internet publisher – limiting the manual annotation required – and from the perspective of a final user consuming the content – helping to decide where to spend the most valuable resource, their time.

The remainder of the thesis is structured as follows: in Section 2 we formally introduce the task of Multimodal Summarization, describe previous works (Section 2.2), formulate the task of Multimodal Summarization with Multimodal Output (Section 2.3) and present the most commonly used modeling (Section 2.4) and evaluation (Section 2.5) techniques. In Section 3 we frame our research plan, describing our completed and in-progress work and sketch our future plans. In Section 4 we briefly introduce our works not directly connected to the multimodal summarization and finally conclude this thesis proposal in Section 5.
2 Multimodal Summarization

2.1 Task formulation

Following Jangra et al. (2021) we define a multimodal summarization task as follows:

“A summarization task that takes more than one mode of information representation (termed as modality) as input, and depends on information sharing across different modalities to generate the final summary.”

Formally, let’s define a multimodal document $D_i$ as a tuple:

$$D_i = (M_{i1}, M_{i2}, \ldots, M_{ik})$$ (1)

where $M_{ij}$ denotes a disjoint information from a particular modality $M_j$, such as video (movie clip), text (textual document) or audio (voice recording) in document $D_i$. While using this notation, we always assume that a particular document $D_i$ is aligned. By that we mean that all modalities are coming from the same source and the document is supposed to be presented as a whole, see Figure 1. It might be the case that some modalities are aligned on an even finer granularity – e.g. video subtitles (text) corresponding to particular timestamps in a video clip (video), but we don’t require it to say that the document as a whole is aligned. Therefore, the task of multimodal summarization (MS) can be formalized with the following formula:

$$\text{MS} : \{D_i\}_{i=1}^k \overset{\sigma}{\rightarrow} D_j$$ (2)

by which we mean the task of creating a (multimodal) summary $D_j$, based on collection of input documents $\{D_i\}_{i=1}^k$ using the $\sigma$ to denote a summarization function. If $D_j$ consists of a single modality, we talk about multimodal summarization with uni-modal output. Otherwise, the task is called multimodal summarization with multimodal output (MSMO).

For a reminder of this thesis we will focus on a text-centric multimodal summarization – we assume that the textual modality is always present both in the input document and in the summary. Most of our attention is directed towards the MSMO variant with a single multimodal document in the input.

2.2 Overview

Early works on multimodal summarization explored the usage of the secondary modalities as an auxiliary source of information to guide the refinement process of the main modality. Tjondronegoro et al. (2011) conducted sentiment analysis of web and social media articles to annotate the key events in sport videos. Li et al. (2017) collected videos and news articles covering a hand-crafted list of recent significant world events and trained a model to mimic the reference summaries written by human annotators. Those early approaches operated on collections of unaligned documents. Data used in the experiments was created by manually querying a search engine for a particular phrase and collecting resources from available outputs. From a modeling point of view, summaries were created in an extractive manner – non-textual features were not used directly in the generation process, but rather distilled to a set of weights.

Li et al. (2018) introduced the multimodal sentence summarization task that generates a short textual summary from a pair of sentence and image. The authors argue that the visual clues are useful for identifying the event highlights which should help to produce better summaries. In their experiments they use the (sentence, headline) tuples from the Gigaword corpus (Rush et al., 2015) and use the search engine to crawl matching images. Human annotators are used to select the best-match image for each sentence. The authors identified the need for a filtering mechanism. The proposed model should be able to filter out noises from the visual modality, in case of e.g. the image failing to represent some abstract concepts. Compared to the previous works, the input documents are still unaligned but the non-textual features are directly incorporated in the representations used for decoding. Li et al. (2020b) proposed an alternative approach for solving this task. Instead of fusing the visual and textual features into a cross-modal representation, they use the image features to train visual selective gates that control flow from the textual encoder.

Besides the news domain, multimodal summarization was applied to the e-commerce data. Li et al. (2020a) presented an abstractive summarization system that produces textual summary for Chinese e-commerce products. They curated a dataset of (product information, product summary) pairs – the product information contains an image, a title and a variable amount of textual descriptions. Product summaries were written by professionals to provide customers with valuable information that
is supposed to convince them to buy the product. The provided images represent products from three categories: **Home Appliances**, **Clothing** and **Cases & Bags**. Due to the rather narrow domain and similar style of the images, the authors propose a novel approach for extracting visual features. Instead of using activations from the pre-softmax dense layer of a CNN model trained for image classification, activations from the Region of Interest (ROI) pooling layer (Girshick, 2015) of a model trained for object recognition are explored. Im et al. (2021) approached a similar problem, opinion summarization, in a self-supervised manner. Each instance in their dataset consists of a collection of reviews ($R$) describing a particular product, user-supplied product images and additional tabulated metadata. A Transformer-based model (Vaswani et al., 2017) is trained to generate a textual summary, using one of the reviews $r_j$ as a target and the remaining ones $R_{-j}$ as input.

The How2 Dataset (Sanabria et al., 2018; Palaskar et al., 2019) provides an example of multimodal summarization applied to yet another domain – instructional, open domain videos. The dataset was created by scraping a popular multimedia hosting platform, collecting video, audio, subtitles and textual „descriptions” that play the role of summaries. Videos were chosen based on the search engine output when queried with a manually created list of key-words. Thanks to its impressive size (over 13,000 instances) and the fact that the authors released the dataset as an easy to download package, this resource was extensively used by other researchers (Khullar and Arora, 2020; Liu et al., 2020; Yu et al., 2021). It is however worth noting that the authors released the pre-computed features and not the raw data, making the study involving e.g. vision-language pre-trained models such as VisualBERT (Li et al., 2019), HERO (Li et al., 2020c) or CLIP (Radford et al., 2021) unfeasible.

### 2.3 MSMO

The task of multimodal summarization with multimodal output (MSMO) was first introduced by Zhu et al. (2018). The authors argue that the multimodal output is crucial from the user perspective – it is helpful both to clarify a generic...
statement such as „four-legged creature” in the context of a summary, but also to get an initial grasp of the key information provided, see Figure 2. They collect a large-scale multimodal corpus by web-scrapping a popular news website. From each article, the main textual body, a set of images and human written highlights are collected. Human annotators are employed to create test set annotations, by selecting up to three most relevant images to play the role of pictorial summary. During training, the learning signal is provided only by the gold-standard textual summaries, while the (visual) coverage mechanism (See et al., 2017) is employed to learn the text-image alignment. The coverage vector is used during inference to sort the input images and select the highest scoring one as a cover picture. A novel method is proposed to evaluate the quality of multimodal output, see Section 2.5. In a follow-up work (Zhu et al., 2020a) the authors propose a method to extend the text reference to a multimodal one. They sort the images based on either the order in which they appear in the original news or the lexical similarity between the image caption and the text reference. Thanks to this, an additional learning signal is provided to a model.

Building upon this work, Li et al. (2020e) propose the task of Video-based Multimodal Summarization with Multimodal Output (VMSMO), see Figure 2. Li et al. (2020e) argue that in real-world applications a text article is usually accompanied by a video consisting of hundreds of frames rather than a few images. Therefore, they propose to choose a single frame to act as a pictorial summary that should represent the salient point of the whole video. To facilitate their research, they collect a dataset from the largest social network website in China. Besides individuals, China’s mainstream media also have accounts on that platform, which they use to post short, lively videos and articles. Each instance in the curated dataset contains a textual article, textual summary and a video with a reference cover picture. In their experiments, the cover picture is not used directly. Instead, they regard the frame that has the maximum cosine similarity with the reference cover picture as the positive sample and all the others as negative samples. In a similar work, Fu et al. (2021) present a full-scale multimodal dataset comprehensively gathering documents, summaries, images, captions, videos, audios, transcripts, and titles. The dataset was collected from well-known English news websites. Compared to Li et al. (2020e) the proposed dataset does not include a single reference picture, but instead utilizes unsupervised methods during training.

2.4 Modeling techniques

The generic architecture used for multimodal summarization modeling is presented in Figure 4. Three universal components can be identified:

- **Feature Encoder** used to obtain the numerical representations of input modalities,
- **Cross-modal Interaction Module** fusing the representations,
- **Multimodal Decoder** responsible for summary generation.

In the following paragraphs we will address each component separately.

2.4.1 Feature Encoder

To obtain the numerical representations, each modality is processed by a separate encoder.
Figure 4: An overview of a generic architecture used for multimodal summarization modeling. For a detailed discussion see Section 2.4.

The text modality is first tokenized into subwords (Sennrich et al., 2016; Kudo and Richardson, 2018) and then contextualized with either the LSTM (Hochreiter and Schmidhuber, 1997), e.g. (Zhu et al., 2018, 2020a; Li et al., 2020e; Fu et al., 2021) or Transformer (Vaswani et al., 2017) encoder, e.g. (Yu et al., 2021; Im et al., 2021).

To encode images, most of the previous works (Li et al., 2018; Zhu et al., 2018; Li et al., 2020e; Im et al., 2021) extracted the activations from the pre-softmax dense layer of a CNN model trained for image classification. Variants of ResNet (He et al., 2016b) and VGG (Simonyan and Zisserman, 2014) are the most commonly used ones. Li et al. (2020a) proposed instead to use activations from the ROI pooling layer of model trained for object detection. This kind of features was shown to improve performance in a related task of Visual Question Answering (Teney et al., 2018; Wu et al., 2019). Since no specific order can be imposed on the images, it is uncommon to contextualize the image representations with a dedicated encoder.

In terms of video encoding, previous works extracted the frame-level features using the same methods that were used to encode images. Fu et al. (2021) model the sequential pattern with a single BiLSTM encoder. Li et al. (2020a) argue that video can be divided into meaningful segments (scenes). To capture this phenomenon they propose a hierarchical encoder, using a low-level frame encoder in parallel with a segment-level encoder that encodes equally distributed sequences of frames. Liu et al. (2020) utilize the temporal dependencies directly, by extracting features from a 3D ResNet (Hara et al., 2018) trained for action recognition.

2.4.2 Cross-modal Interaction Module

The Interaction Module is used to fuse the sequences representing disjoint modalities into a common subspace. A variety of architectures have been proposed for the fusion task, as such joint representations are also a key component of other applications e.g. video question answering (Jang et al., 2017; Tapaswi et al., 2016) or video captioning (Zhou et al., 2018b; Yao et al., 2015).

One of the simpler methods that were proposed is to just concatenate the sequences along the temporal dimension and process them as a whole with a dedicated encoder (Li et al., 2020d). This solution enables merging an arbitrary number of sequences, at the increased computational cost due to the quadratic complexity.

Most of the modern cross-model modules are based instead on the attention mechanism (Bahdanau et al., 2015). In the models based on Recurrent Neural Networks (RNN) sequences can attend to one another similarly how the decoder can attend to the source in traditional seq2seq models. In the
models based on Transformer the cross-attention block from decoder is used, by utilizing the fact that the queries can be computed from a sequence of different length. To enable even deeper integration sophisticated tangled, hierarchical modules are used (Zhu and Yang, 2020; Yu et al., 2021).

2.4.3 Multimodal Decoder
To generate the textual summary RNN (Zhu et al., 2018; Li et al., 2020e; Fu et al., 2021) or Transformer (Yu et al., 2021; Im et al., 2021) based decoders are used, operating on the fused representations. Cross-entropy loss is applied to compute gradients.

To choose the image as a pictorial summary, Zhu et al. (2018, 2020b) select the image with the largest (visual) coverage score. Li et al. (2020e) compute a matching score for each video frame based on the original and conditional representations, and during inference choose the frame with the highest score. They use the pairwise hinge loss to compute the learning signal.

2.5 Evaluation methods
Since manual annotation for any generative task is costly and time consuming, automatic metrics are commonly used to evaluate the model performance.

2.5.1 Automatic metrics
To evaluate the textual summary, most works (Zhu et al., 2018; Li et al., 2020e; Fu et al., 2021) keep relying solely on ROUGE (Lin, 2004), a string-overlap metric measuring the n-gram correspondence with the reference summary. Liu et al. (2020) and Yu et al. (2021) report also several other metrics such as BLEU (Papineni et al., 2002) or CIDERr (Vedantam et al., 2015). Im et al. (2021) follow the recent trends in summary evaluation and report the BERT-score (Zhang et al., 2020b) metric. Palaskar et al. (2019) introduce the Content F1 metric that is designed to fit the template-like structure of multimodal summaries. This metric computes the monolingual alignment between the model output and reference summary. After removing function words and task-specific stop words that appear in a majority of summaries, a F1 score is computed over the alignment, treating the hypothesis and reference as two bags of words.

The first work (Zhu et al., 2018) to introduce multimodal (pictorial) summary reports image precision to measure the salience of image, framing the problem as an image recommendation task. Since their training data does not include the gold-standard (reference) images, human annotators are employed to select relevant images (\{ref\_img\}) from the articles in the subset of the test set. Those are then compared to the top-n images as scored by model (\{rec\_img\}):

$$IP = \frac{|\{\text{ref\_img}\} \cap \{\text{rec\_img}\}|}{|\{\text{rec\_img}\}|}.$$  

In this work authors notice that a prerequisite for a pictorial summary to help users accurately acquire information is that the image must be related to the text – and this is not measured with the IP metric. Therefore, they train a model for image-caption retrieval on the Flickr30K dataset (Young et al., 2014) and use it to calculate the similarity between the selected images and sentences in the textual summary. An attempt is made to measure the quality of a multimodal summary when perceived as a whole (text + image(s)). The proposed MMAE method combines scores from several metrics (comparing text-to-text, image-to-image and text-to-image) by fitting a regression model over the metric outputs. Human judgments measuring „user satisfaction” are used to fit weights. In their follow-up work (Zhu et al., 2020a) an additional metric from a cross-modal retrieval model is considered as an input for the regression, introducing the MMAE++ method.

Li et al. (2020e) and Fu et al. (2021), both of which present work on the VMSMO task, operate in different settings. Li et al. (2020e) sample frames (one of every 120) from the video to obtain candidates for the pictorial summary. Their reference is a single image that was representing the article on the website. They regard the frame that has the maximum cosine similarity with the ground truth cover as the positive sample, and the others as negative samples, during both training and testing. Therefore, they report the mean average precision (MAP) and recall at position \(k\) (\(R_n@k\)). \(R_n@k\) measures if the positive sample is ranked in the top \(k\) positions of \(n\) candidates. Fu et al. (2021) also sample frames from the video to obtain candidates for the pictorial summary. However, they do not have access to a single reference picture, but rather to a set of images co-located with the article. Thus, they train the frame selector in an unsupervised manner following Zhou et al. (2018a) and report an
average cosine similarity of the top-1 frame with the set of reference images.

2.5.2 Human Evaluation

Previous works applied human evaluation to assess the quality of multimodal summarization. Zhu et al. (2018) conducted an experiment to investigate whether a pictorial summary can improve the user perception of the informativeness of the summary. Annotators were given a collection of source news pages with corresponding textual summaries and pictorial summaries. They were requested to independently evaluate the text summaries and the pictorial summaries, according to the input news. Scoring was done on a scale of 1 to 5. To obtain a single score for a multimodal summary, the two scores were averaged.

Li et al. (2020e) measured to what extent the system (textual) summaries were sufficient to answer questions generated from the reference summary and ranked them based on Informativeness, Coherence and Succinctness; Fu et al. (2021) scored the system (textual) summaries based on Informativeness and Satisfaction. Neither of these works judged the quality of the chosen cover frame (pictorial summary).

3 Our contribution

Having described the current state of research, in this Section we will discuss what we believe to be the main challenges of multimodal summarization (focusing on the VMSMO variant, see Section 2.3) that we would like to approach.

1. Lack of data To the best of our knowledge only two datasets have been introduced for the VMSMO task – Li et al. (2020e)\(^2\) and Fu et al. (2021)\(^3\), see Table 1. The dataset by Fu et al. (2021) is shared directly with the public, but Li et al. (2020e) shared only URLs and instruction on how to download the data. In our attempt to re-create the dataset, only less than 10% of the URLs were active.

2. Cross-modal feature extraction Previous works used separate feature encoders to obtain the numerical representations for each modality, which are then fused into the contextualized representation (Section 2.4). We believe that directly using multi-modal embeddings (Miech et al., 2019; Li et al., 2019, 2020c; Radford et al., 2021; Xu et al., 2021) should enable even deeper fusion.

3. Task-specific pre-training Yu et al. (2021) studied different methods for injecting visual information into pre-trained generative language models (Lewis et al., 2020; Raffel et al., 2020; Zhang et al., 2020a), in the context of multimodal summarization. They did not however explore a dedicated, task-specific pre-training.

4. Multimodal evaluation As discussed in Section 2.5, existing works evaluate each output modality independently. Zhu et al. (2018, 2020b) are the only ones to propose a method that would measure the quality of multimodal output as a whole. Their solution however requires human annotated data to determine key parameters and is thus not applicable for evaluating summaries from different domains/languages. In addition, even the unimodal evaluation metrics that are commonly used do not follow recent guidelines. In Section 2.5 we notice that most works relay solely on ROUGE, which is highly discouraged by e.g. Fabbri et al. (2021).

3.1 MLASK

In this Section we will reference our unpublished work, currently under review for the COLING 2022 conference.

To enable our research and extend available resources for the VMSMO task, we collected a multimodal summarization dataset (Challenge 1) in Czech (MLASK – Multimodal Article Summarization Kit). Each instance in our dataset includes the article’s text, title, abstract, video and a single cover picture. For comparison with previous works see Table 1. The dataset was obtained by automatically crawling several Czech news websites.

In our experiments, a video-based news article was represented by a pair \((V, X)\). \(V\) corresponds to the video input – a sequence of frames: \(V = (v_1, v_2, \ldots , v_N)\). \(X\) is the news article presented as a sequence of tokens: \(X = (x_1, x_2, \ldots , x_M)\). We assumed that for each article there is a ground-truth textual summary \(Y = (y_1, y_2, \ldots , y_L)\) and a ground-truth cover picture \(P\). Our goal was to generate a textual summary \(\hat{Y}\) that includes the
main points of the article and to choose a frame $v$ to act as a cover picture (pictorial summary).

We proposed a multimodal summarization model that was structured into three parts: Feature Encoder composed of a text, video, and frame encoder, Cross-modal Interaction Module fusing the visual and textual representations, and Multimodal Decoder responsible for summary generation and frame selection, see Figure 5. We have used the pre-trained mt5 model (Xue et al., 2021) to initialize both text encoder and decoder weights. We experimented with both CNN (Tan and Le, 2019) and Transformer (Dosovitskiy et al., 2021) based visual feature extractors. Following Liu et al. (2020) we implemented the forget gate mechanism so that the model can filter out low-level cross-modal adaptation information.

Having access to the raw videos, we were able to show that using the multi-modal embeddings is beneficial to the final performance (Challenge 2). By incorporating the visual representations from the model trained on text-video pairs (Miech et al., 2020) we were able to improve the ROUGE-L score (12.93 → 13.26) as compared to the variant using feature extractor trained solely on video data (Ghadiyaram et al., 2019).

We showed that by pre-training the textual encoder and decoder on a simpler task of text-to-text summarization we can effectively take advantage of larger, text-only resources available (Challenge 3). By pre-training the text encoder and decoder on the Czech news summarization corpus (Straka et al., 2018) we have managed to improve the quality of textual summary (ROUGE-L: 13.26 → 14.32, ROUGE-1: 18.34 → 19.64).

Previous works on VMSMO did not use the cover picture directly, but rather regarded the frame that has the maximum cosine similarity with the reference cover picture as the positive sample and all the others as negative samples, during both training and testing (Section 2.3). After examining the cosine similarity patterns (Figure 6), we noticed that the per-video similarity often either has more than one peak (capturing a recurring scene) or includes consecutive sequences of frames with very similar scores (capturing a still scene). Our intuition was that this may harm the model performance – we may label very similar frames as both positive and negative examples. To overcome this issue, we are the first to propose the smooth labels, by directly assigning the cosine similarity score as targets in (cross-entropy) loss computation. Results of our experiments support our hypothesis: we have managed to improve on average both the Recall@10 (0.318 → 0.330) and the cosine similarity (0.541 → 0.551) between the top-1 frame chosen by the model and the reference picture.

As mentioned in Section 2.5 previous works on VMSMO performed human evaluation only to assess the quality of the textual output. In our work we propose a framework (Challenge 4) to judge the quality of a chosen cover frame (pictorial summary). Figure 7 displays a screenshot of the annotation tool that we used. For each instance considered, the annotators were asked to rate 3 images on the scale of 0 to 4 (the higher the better) in the context of the article’s title and the reference summary. Four methods were considered for annotation: the reference picture, a random frame from the video and the outputs of two test models that we propose\(^4\). We have designed the annotation process in a way that allowed us to control the inter-annotator agreement – our test data was split into batches and each annotator was asked to score the control batch. Cohen’s $\kappa$ value of 0.217 indicated a “fair” agreement. The aggregated results allowed us to conclude that the reference picture is assigned the highest score and our proposed multimodal summarization models performs better than the random baseline.

3.2 Future plans

One negative result that came up from our experiments concerns the usefulness of video features. Although the video is crucial for the task (we follow previous works and frame the cover picture choice as a frame selection problem) in our experiments (with text encoder an decoder pre-trained on text summarization) the quality of textual summary did not change if we masked the video features with a random noise. Human evaluation confirmed the findings based on automatic metrics. We plan to further investigate this issue by considering auxiliary training objectives that encourage the model to effectively incorporate the visual features.

Hessel et al. (2021) showed recently that CLIP (Radford et al., 2021), a cross-modal model pre-trained on 400M image-caption pairs from the web,\(^4\)For each instance we sample 3 out of 4 images to be displayed during annotation.
Figure 5: An overview of the multimodal summarization model that we proposed, see Section 3.1. MHA stands for Multi-Head Attention.

Table 1: Comparison of the datasets used for the VMSMO task. The concrete statistics are reported as an average computed over the whole corpus. For the textual part we report the average number of tokens.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>#Articles</th>
<th>Article Length</th>
<th>Summary Length</th>
<th>Video Length</th>
<th>Language</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLASK (ours)</td>
<td>41,243</td>
<td>277</td>
<td>33</td>
<td>86s</td>
<td>cs</td>
</tr>
<tr>
<td>MM-AVS (Fu et al., 2021)</td>
<td>2,173</td>
<td>685</td>
<td>57</td>
<td>109s</td>
<td>en</td>
</tr>
<tr>
<td>VMSMO (Li et al., 2020e)</td>
<td>184,920</td>
<td>97</td>
<td>11</td>
<td>60s</td>
<td>zh</td>
</tr>
</tbody>
</table>

Figure 6: Three examples of cosine similarity plots between CNN features of the reference cover picture and candidate frames from the video. The examples were chosen manually to present three different video similarity patterns: with a single peak (red), with more than one peak (blue, capturing a recurring scene), and with consecutive sequence of frames having very similar scores (violet, capturing a still scene). For a detailed discussion see Section 3.1.

4 Other Work

Besides the multimodal summarization, we have approached two text-only tasks: machine translation evaluation and text summarization evaluation.

4.1 Machine Translation Evaluation

Several works identified recently that the abstractive text summarization models are vulnerable to hallucinations (Wiseman et al., 2017; Dhingra et al., 2019; Kryscinski et al., 2020). Due to the noise in training data, models tend to output fluent text with reasonably high log-likelihood, that is however not consistent with the input, see Table 2. To address this issue, a series of works (Eyal et al., 2019; Scialom et al., 2019; Durmus et al., 2020; Wang et al., 2020) proposed evaluation methods capable of identifying the factual inconsistencies can be used for robust automatic evaluation of image captioning without the need for references. The authors proposed also a simple way of incorporating the reference caption(s). In the future research we plan to adapt this solution for the MSMO evaluation. The framework we developed and the annotations that we collected (Section 3.1) make this research feasible. This approach would answer the issue that we raised previously (Challenge 4) – lack of methods that measure the quality of multimodal output as a whole.
(hallucinations). The common idea was that if we identify a piece of information in model output and (automatically) generate a question asking about this information nugget, then such question should be answerable based on the original document/reference summary.

In Krubiński et al. (2021a) we argue that neural MT models have similar issues – they produce fluent output that is not consistent with the source. We examine the usefulness of the question-answer framework for the MT evaluation by proposing a new metric – MTEQA. We show that the system-level correlations with human judgments obtained by our metric are on pair with other state of the art solutions, while considering only a certain amount of information from the whole translation output. To further evaluate our finding, we participated (Krubiński et al., 2021b) in the WMT Metrics Shared Task (Freitag et al., 2021). The metric that we proposed achieved the highest system-level correlation with human judgments on the Chinese→English direction when scoring the TED talks test-set (Table 13 in Freitag et al. (2021)).

4.2 Textual Summary Evaluation

In our unpublished work (to be submitted to the Eval4NLP Workshop co-located at the AACL-IJCNLP 2022 conference) we look at the problem from a reverse perspective. Thanks to the WMT News Translation Shared Task (Barrault et al., 2019, 2020; Akhabardeh et al., 2021) a large collection of roughly 800k annotated (source, hypothesis, reference) triplets is available. Using these resources for training, several trainable neural-based evaluation metrics capable of directly regressing quality score were developed (Lo, 2019; Kepler et al., 2019; Rei et al., 2020; Sellam et al., 2020).

In our work we focus on COMET (Rei et al., 2020), a metric that was chosen as the best performing one for the MT quality evaluation in a recent study (Kocmi et al., 2021). The question we ask is whether we can use the recent advances in MT evaluation – in particular, trainable neural-based metrics – to improve summary evaluation. We believe this may be a way towards addressing two of the issues making research on summary evaluation metrics difficult – lack of a standardized framework for collecting human judgments (Table 3) and relatively modest size of available annotated data. Our results indicate that the metrics trained on multilingual MT outputs perform surprisingly well in the mono-lingual settings, when evaluating summarization output quality. We also show that pre-training on the large collection of annotated MT outputs and then fine-tuning on the much smaller collection of annotated summary outputs is a promising research direction that enables predicting several aspects of summary quality. We apply our finding by proposing a new metric for textual summary evaluation – COMES. We further explore this idea by reporting performance in the quality estimation settings (without access to the reference summary) and using several datasets with human judgments collected for different notions of summary quality.

5 Conclusion

In this thesis proposal we introduced the task of Multimodal Summarization, presented previous
Jerusalem (CNN)The flame of remembrance burns in Jerusalem, and a song of memory haunts Valerie Braham as it never has before. This year, Israel’s Memorial Day commemoration is for bereaved family members such as Braham. “Now I truly understand everyone who has lost a loved one,” Braham said. Her husband, Philippe Braham, was one of 17 people killed in January’s terror attacks in Paris. He was in a kosher supermarket when a gunman stormed in, killing four people, all of them Jewish.

France’s memorial day commemoration is for bereaved family members as Braham. Valerie Braham was one of 17 people killed in January’s terror attacks in Paris.

Table 2: Example of hallucinations in abstractive text summarization from Kryscinski et al. (2020). We emphasise the matching phrases, using colors to indicate hallucinations, see Section 4.

| SummEval (Fabbri et al., 2021) | Coherence | Consistency | Fluency | Relevance | SCU | Accuracy | Coverage | Focus | Overall |
| RealSumm (Bhandari et al., 2020) | ✓ | ✓ | ✓ | ✓ | | ✓ | ✓ | ✓ | ✓ |
| Human Feedback (Stiennon et al., 2020) | ✓ | | | | | ✓ | ✓ | ✓ | ✓ |
| Multi_SummEval (Koto et al., 2021) | | | | | | | | | |

Table 3: Comparison of the types (dimensions) of human annotations in the summary evaluation datasets used in our experiments, see Section 4. Unlike other generative tasks such as Machine Translation, it is a custom in Text Summarization to grade the model output along several independent dimensions.

works and discussed possible variants of the task (multimodal summarization with uni-modal output, multimodal summarization with multimodal output). We identified and described what we believe to be the main challenges that current approaches need to solve. We briefly introduced our work – collection of the MLASK dataset and our results regarding multi-modal embeddings and task-specific pre-training. We also proposed a human evaluation framework for assessing the quality of pictorial summary. We sketched our future research plans and described our results concerning machine translation evaluation and text summarization evaluation.

References


Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. **HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips.** In *ICCV*.


