Multimodal Summarization
Ph.D. Thesis Proposal

Mateusz Krubiński
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
krubinski@ufal.mff.cuni.cz

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[1]:

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 became 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.

[1] https://www.oxfordlearnersdictionaries.com

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:

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

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 granulation – 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:

$$\text{MS} : \{D_i\}_1^k \xrightarrow{\sigma} D_j \tag{2}$$

by which we mean the task of creating a (multimodal) summary D_j, based on collection of input documents $\{D_i\}_1^k$ using the σ 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).

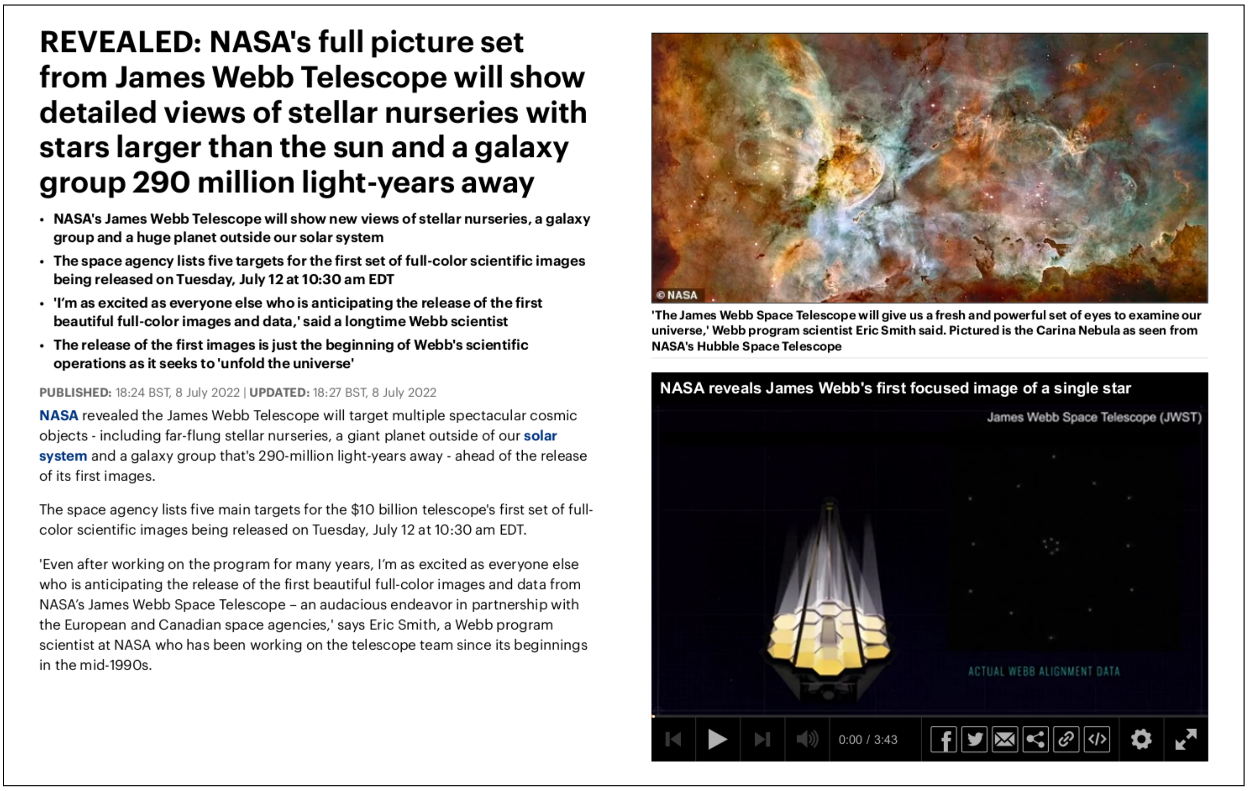
For a reminder of this thesis we will focus on a *text-centric* multimodal summarization – we assume that the textual modality is always present both in the input document and in the summary. Most of our attention is directed towards the MSMO variant with a single multimodal document in the input.

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



REVEALED: NASA's full picture set from James Webb Telescope will show detailed views of stellar nurseries with stars larger than the sun and a galaxy group 290 million light-years away

- NASA's James Webb Telescope will show new views of stellar nurseries, a galaxy group and a huge planet outside our solar system
- The space agency lists five targets for the first set of full-color scientific images being released on Tuesday, July 12 at 10:30 am EDT
- 'I'm as excited as everyone else who is anticipating the release of the first beautiful full-color images and data,' said a longtime Webb scientist
- The release of the first images is just the beginning of Webb's scientific operations as it seeks to 'unfold the universe'

PUBLISHED: 18:24 BST, 8 July 2022 | UPDATED: 18:27 BST, 8 July 2022

NASA revealed the James Webb Telescope will target multiple spectacular cosmic objects - including far-flung stellar nurseries, a giant planet outside of our solar system and a galaxy group that's 290-million light-years away - ahead of the release of its first images.

The space agency lists five main targets for the $10 billion telescope's first set of full-color scientific images being released on Tuesday, July 12 at 10:30 am EDT.

'Even after working on the program for many years, I'm as excited as everyone else who is anticipating the release of the first beautiful full-color images and data from NASA's James Webb Space Telescope – an audacious endeavor in partnership with the European and Canadian space agencies,' says Eric Smith, a Webb program scientist at NASA who has been working on the telescope team since its beginnings in the mid-1990s.

© NASA

'The James Webb Space Telescope will give us a fresh and powerful set of eyes to examine our universe,' Webb program scientist Eric Smith said. Pictured is the Carina Nebula as seen from NASA's Hubble Space Telescope

Figure 1: Example of a multimodal news article from an online publisher (Daily Mail). Three modalities: text, image(s) and video are presented to a user. Each of them brings a new, unique piece of information. While particular modalities may have an inner structure – text can be split into *Title*, *Abstract* and *Story*, in general, no specific order can be imposed on objects from different modalities.

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 scrapping 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,

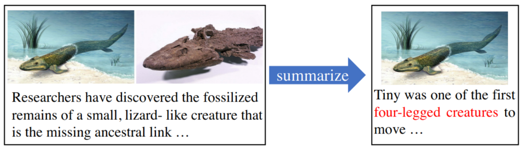


Figure 2: Illustration of the Multimodal Summarization with Multimodal Output (MSMO) task proposed by Zhu et al. (2018). Figure reprint from Zhu et al. (2018).

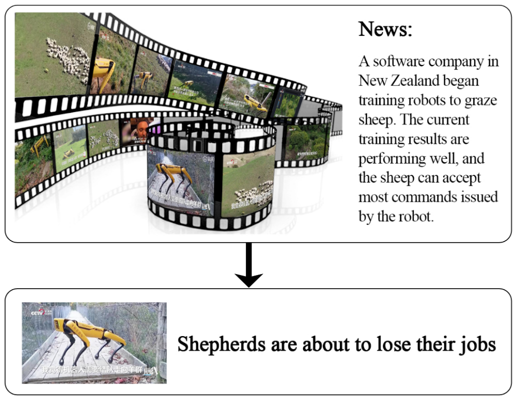


Figure 3: Illustration of the Video-based Multimodal Summarization with Multimodal Output (VMSMO) task proposed by Li et al. (2020e). Figure reprint from Li et al. (2020e).

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.

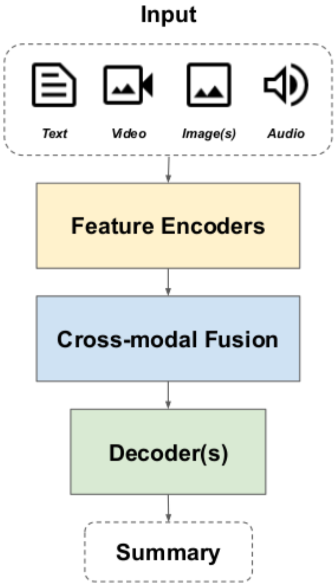


Figure 4: An overview of a generic architecture used for multimodal summarization modeling. For a detailed discussion see Section 2.4.

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. (2020e) 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 ($\{\text{ref}_{img}\}$) from the articles in the subset of the test set. Those are then compared to the *top-n* images as scored by model ($\{\text{rec}_{img}\}$):

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

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 give 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 – **M**ultimoda**L** **A**rticle **S**ummarization **K**it). 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: $V = (v_1, v_2, \dots, v_N)$. X is the news article presented as a sequence of tokens: $X = (x_1, x_2, \dots, x_M)$. We assumed that for each article there is a ground-truth textual summary $Y = (y_1, y_2, \dots, y_L)$ and a ground-truth cover picture P. Our goal was to generate a textual summary $\widehat{Y}$ that includes the

[2] https://github.com/iriscxy/VMSMO
[3] https://github.com/xiyan524/MM-AVS

main points of the article and to choose a frame $\widehat{v}$ 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 ($12.93 \rightarrow 13.26$) 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: $13.26 \rightarrow 14.32$, ROUGE-1: $18.34 \rightarrow 19.64$).

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 ($0.318 \rightarrow 0.330$) and the cosine similarity ($0.541 \rightarrow 0.551$) 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 asses 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 κ 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.

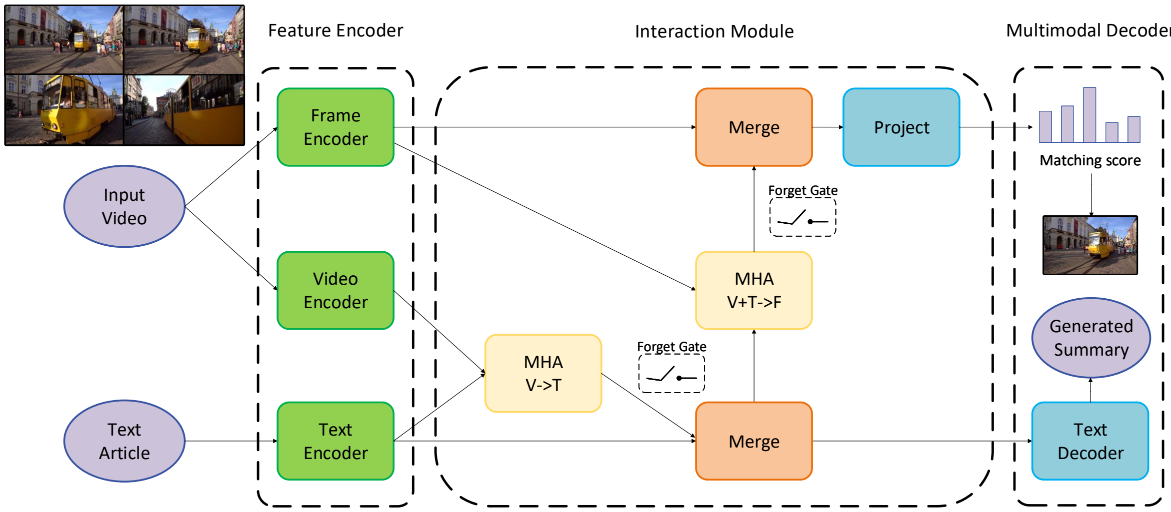


Figure 5: An overview of the multimodal summarization model that we proposed, see Section 3.1. MHA stands for Multi-Head Attention.

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

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.

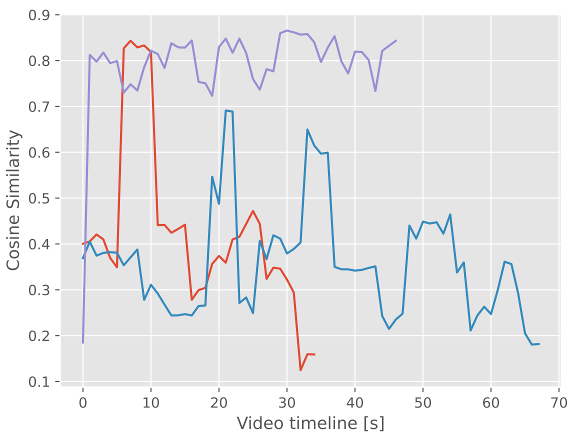


Figure 6: Three examples of cosine similarity plots between CNN features of the reference cover picture and candidate frames from the video. The examples were chosen manually to present three different video similarity patterns: with a single peak (red), with more than one peak (blue, capturing a recurring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

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.

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

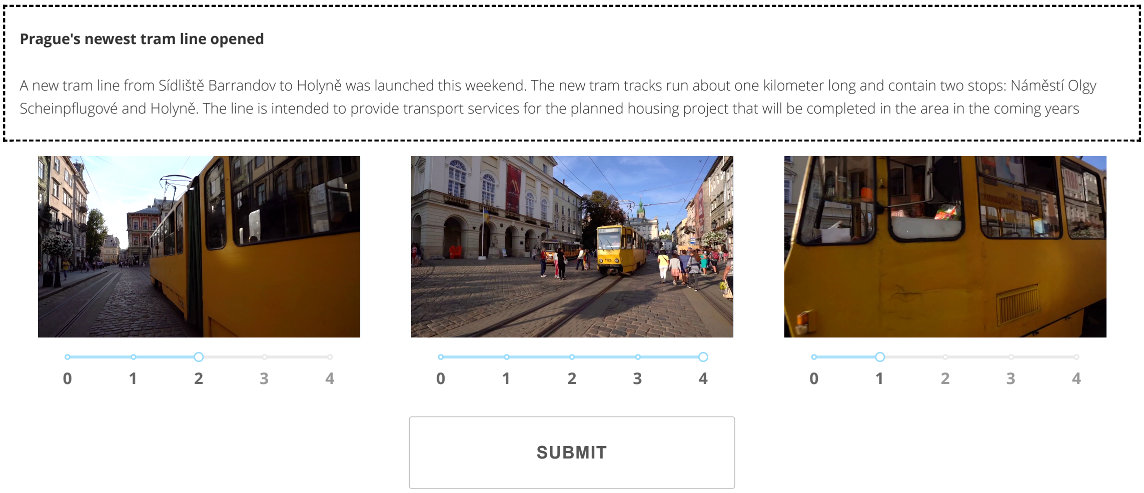


Figure 7: Screenshot of the annotation tool that we used to collect human judgments about the quality and usefulness of selected cover frame, see Section 3.1.

(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; Akhbardeh 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

CNN/DailyMail Source	Model output
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.

	Coherence	Consistency	Fluency	Relevance	SCU	Accuracy	Coverage	Focus	Overall
SummEval (Fabbri et al., 2021)	✓	✓	✓	✓					
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

Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyung Hyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 ; Conference date: 07-05-2015 Through 09-05-2015.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 9347–9359, Online. Association for Computational Linguistics.

Ondřej Bojar, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Belgium, Brussels. Association for Computational Linguistics.

Daily Mail. `www.dailymail.co.uk/sciencetech/article-10995713/NASAs-James-Webb-Telescope-targets-Carina-Nebula-Stephans-Quintet-Southern-Ring-Nebula-more.html`. [Online; accessed 9-July-2022].

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Bhuwan Dhingra, Manaal Faruqui, Ankur Parikh, Ming-Wei Chang, Dipanjan Das, and William Cohen. 2019. Handling divergent reference texts when evaluating table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4884–4895, Florence, Italy. Association for Computational Linguistics.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5055–5070, Online. Association for Computational Linguistics.

Eurostat. `https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Digital_economy_and_society_statistics_-_households_and_individuals`. [Online; accessed 17-July-2022].

Matan Eyal, Tal Baumel, and Michael Elhadad. 2019. Question answering as an automatic evaluation metric for news article summarization. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3938–3948, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Markus Freitag, Ricardo Rei, Nitika Mathur, Chi-kiu Lo, Craig Stewart, George Foster, Alon Lavie, and Ondřej Bojar. 2021. Results of the WMT21 metrics shared task: Evaluating metrics with expert-based human evaluations on TED and news domain. In *Proceedings of the Sixth Conference on Machine Translation*, pages 733–774, Online. Association for Computational Linguistics.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. MM-AVS: A full-scale dataset for multi-modal summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5922–5926, Online. Association for Computational Linguistics.

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. 2019. Large-scale weakly-supervised pre-training for video action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12038–12047.

Ross Girshick. 2015. Fast r-cnn. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1440–1448.

Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016a. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016b. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '16, pages 770–778. IEEE.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational*

Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 388–403, Online. Association for Computational Linguistics.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1359–1367.

Anubhav Jangra, Adam Jatowt, Sriparna Saha, and Md. Hasanuzzaman. 2021. A survey on multi-modal summarization. *ArXiv*, abs/2109.05199.

Fabio Kepler, Jonay Trénous, Marcos Treviso, Miguel Vera, and André F. T. Martins. 2019. OpenKiwi: An open source framework for quality estimation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 117–122, Florence, Italy. Association for Computational Linguistics.

Aman Khullar and Udit Arora. 2020. MAST: Multimodal abstractive summarization with trimodal hierarchical attention. In *Proceedings of the First International Workshop on Natural Language Processing Beyond Text*, pages 60–69, Online. Association for Computational Linguistics.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Evaluating the efficacy of summarization evaluation across languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 801–812, Online. Association for Computational Linguistics.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021a. Just ask! evaluating machine translation by asking and answering questions. In *Proceedings of the Sixth Conference on Machine Translation*, pages 495–506, Online. Association for Computational Linguistics.

Mateusz Krubiński, Erfan Ghadery, Marie-Francine Moens, and Pavel Pecina. 2021b. MTEQA at WMT21 metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation*, pages 1024–1029, Online. Association for Computational Linguistics.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the factual consistency of abstractive text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346, Online. Association for Computational Linguistics.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8188–8195.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4152–4158. International Joint Conferences on Artificial Intelligence Organization.

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*, pages 1092–1102, Copenhagen, Denmark. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Jiajun Zhang, Xiaodong He, and Chengqing Zong. 2020b. Multimodal sentence summarization via multimodal selective encoding. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5655–5667, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020c. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065, Online. Association for Computational Linguistics.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2020d. What does bert with vision look at? In *ACL (short)*.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020e. VMSMO: Learning to generate multimodal summary for video-based news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9360–9369, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1834–1845, Online. Association for Computational Linguistics.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In *CVPR*.

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*.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6587–6596, Florence, Italy. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. 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)*, pages 3246–3256, Hong Kong, China. Association for Computational Linguistics.

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)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages

7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

Milan Straka, Nikita Mediankin, Tom Kocmi, Zdeněk Žabokrtský, Vojtěch Hudeček, and Jan Hajič. 2018. SumeCzech: Large Czech news-based summarization dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mingxing Tan and Quoc Le. 2019. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding Stories in Movies through Question-Answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Damien Teney, Peter Anderson, Xiaodong He, and Anton van den Hengel. 2018. Tips and tricks for visual question answering: Learnings from the 2017 challenge. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4223–4232.

Dian Tjondronegoro, Xiaohui Tao, Johannes Sasongko, and Cher Han Lau. 2011. Multi-modal summarization of key events and top players in sports tournament videos. In *2011 IEEE Workshop on Applications of Computer Vision (WACV)*, pages 471–478.

Uswitch. `https://www.uswitch.com/mobiles/screentime-report/`. [Online; accessed 17-July-2022].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575. IEEE Computer Society.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5008–5020, Online. Association for Computational Linguistics.

Sam Wiseman, Stuart Shieber, and Alexander Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2253–2263, Copenhagen, Denmark. Association for Computational Linguistics.

Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Ruifan Li. 2019. Differential networks for visual question answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):8997–9004.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. 2021. VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6787–6800, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3995–4007, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2020a. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. Bertscore: Evaluating text generation with bert. In *ICLR*. OpenReview.net.

Kaiyang Zhou, Yu Qiao, and Tao Xiang. 2018a. Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7582–7589. AAAI Press.

Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. 2018b. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8739–8748.

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*, pages 4154–4164, Brussels, Belgium. Association for Computational Linguistics.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020a. Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020b. Multimodal summarization with guidance of multimodal reference. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):9749–9756.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8743–8752.