MCLS: A Large-Scale Multimodal Cross-Lingual Summarization Dataset

Xiaorui Shi
School of Information, Renmin University of China, Beijing, China
xiaorshi@gmail.com

Abstract

Multimodal summarization which aims to generate summaries with multimodal inputs, *e.g.*, text and visual features, has attracted much attention in the research community. However, previous studies only focus on monolingual multimodal summarization and neglect the non-native reader to understand the cross-lingual news in practical applications. It inspires us to present a new task, named Multimodal Cross-Lingual Summarization for news (MCLS), which generates cross-lingual summaries from multi-source information. To this end, we present a large-scale multimodal cross-lingual summarization dataset, which consists of 1.1 million article-summary pairs with 3.4 million images in 44 * 43 language pairs. To generate a summary in any language, we propose a unified framework that jointly trains the multimodal monolingual and cross-lingual summarization tasks, where a bi-directional knowledge distillation approach is designed to transfer knowledge between both tasks. Extensive experiments on many-to-many settings show the effectiveness of the proposed model.

1 Introduction

The goal of multimodal summarization is to produce a summary with the help of multi-source inputs, *e.g.*, text and visual features. With the rapid growth of multimedia content on the Internet, this task has received increasing attention from the research communities and has shown its potential in recent years. It benefits users from better understanding and accessing verbose and obscure news, and thus can help people quickly master the core ideas of a multimodal article.

In the literature, many efforts have been devoted to the multimodal summarization fields, *e.g.*, SportsSum (Tjondronegoro et al., 2011), MovieSum (Evangelopoulos et al., 2013), MSMR (Erol et al., 2003), MMSS (Li et al., 2017), MSS (Li et al., 2018a), How2 (Sanabria et al., 2018), MSMO (Zhu et al., 2018), E-DailyMail (Chen and Zhuge, 2018), EC-product (Li et al., 2020a), MM-AVS (Fu et al., 2021), and MM-Sum (Liang et al., 2022b). All these datasets cover video summarization, movie summarization, meeting records summarization, sentence summarization, product summarization, and news summarization. With the predefined task, former state-of-the-art multimodal summarization models have achieved great outcomes. For instance, Palaskar et al. (2019) and Zhang et al. (2021a) explore the hierarchy between the textual article and visual features, and integrate them into the MAS model. Liu et al. (2020) design a multistage fusion network to model the fine-grained interactions between the two modalities. And Yu et al. (2021a) study multiple multimodal fusion methods to infuse the visual features into generative pre-trained language models, *e.g.*, BART (Lewis et al., 2020). Despite their efforts and effectiveness, existing methods are all conducted in monolingual scenarios. In practical applications, for non-native news viewers, they desire some native language summaries to better understand the contents of the news in other languages. To our knowledge, little research work has been devoted to multimodal cross-lingual summarization. One important season is the lack of a large-scale multimodal cross-lingual benchmark.

To assist those non-native readers, we propose a new task: Multimodal Cross-Lingual Summarization for news (MCLS). As shown in Figure 1, the inputs consist of two parts: the image sequence and textual article in the source language (*e.g.*, English), and the summary outputs can be in any target language (*e.g.*,

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Figure 1: An example of our MM-CLS dataset. Inputs: an article and image sequence pair; Output: summaries in different language directions.

English, Chinese, Japanese, and French). Therefore, the MCLS seeks to generate summaries in any target language to reflect the salient new contents based on the image sequence and the article in the source language. To this end, based on CrossSum (Bhattacharjee et al., 2022), we first construct a large-scale multimodal cross-lingual summarization dataset (MM-CLS) for news. The MM-CLS includes over 1.1 million article-summary pairs with 3.4 million images in 44 * 43 language pairs.

Based on the constructed MM-CLS, we benchmark the MCLS task by establishing multiple Transformer-based (Vaswani et al., 2017) systems adapted from the advanced representative multimodal monolingual models (Yu et al., 2021a), based on mT5 (Xue et al., 2021). Specifically, we incorporate multimodal features into the models for a suitable summarization in any language. Furthermore, to transfer the knowledge between monolingual summarization and cross-lingual summarization, we design a bidirectional knowledge distillation (BKD) method. Extensive experiments on many-to-many settings in terms of ROUGE scores (Lin, 2004), demonstrate the effectiveness of multimodal information fusion and the proposed BKD.

In summary, our main contributions are:

- We propose a new task: multimodal cross-lingual summarization for news named MCLS, to advance multimodal cross-lingual summarization research.
- We are the first that contributes the large-scale multimodal cross-lingual summarization dataset (MM-CLS), which contains 1.1 million article-summary pairs with 3.4 million images, in total 44 * 43 language pairs.
- We implement multiple Transformer-based baselines and provide benchmarks for the new task. Extensive experiments show that our model achieves state-of-the-art performance on the benchmark. We also conduct a comprehensive analysis and ablation study to offer more insights.

2 Related Work

2.1 Abstractive Text Summarization (ATS)

Given the input textual article, the goal of ATS is to generate a concise summary (Hermann et al., 2015; Wang et al., 2022c). Thanks to the generative pre-trained language models (Lewis et al., 2020), the ATS has achieved remarkable performance (Paulus et al., 2018; Liu and Lapata, 2019; Zhang et al., 2020; Goodwin et al., 2020; Rothe et al., 2021; Xiao et al., 2022; Xu et al., 2020b; Yu et al., 2021b; Wang et

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

al., 2023b). Different form them, this work mainly focuses on benchmarking multimodal cross-lingual summarization.

2.2 Multimodal Abstractive Summarization (MAS)

With the rapid growth of multimedia, many MAS datasets have been built such as SportsSum (Tjondronegoro et al., 2011), MovieSum (Evangelopoulos et al., 2013), MSMR (Erol et al., 2003), MMSS (Li et al., 2017), MSS (Li et al., 2018a), How2 (Sanabria et al., 2018; Liu et al., 2022), MSMO (Zhu et al., 2018), E-DailyMail (Chen and Zhuge, 2018), EC-product (Li et al., 2020a), MM-AVS (Fu et al., 2021), MM-Sum (Liang et al., 2022b), and M^{3}Sum (Liang et al., 2023). All these datasets, covering video summarization, movie summarization, meeting records summarization, sentence summarization, product summarization, and news summarization, aim to generate a summary based on multimodal inputs (text, vision, or audio). With the data resources extensively used, the MAS task has attracted much attention, where the existing work mainly focuses on how to effectively exploit the additional visual features, having achieved impressive performance in recent years (Li et al., 2018b; Li et al., 2020b; Zhu et al., 2020a; Zhu et al., 2021; Zhang et al., 2021b; Zhang et al., 2021a; Yu et al., 2021a). The difference from ours lies in the cross-lingual summarization where we hope to generate a summary in any target language.

2.3 Cross-Lingual Summarization (CLS)

Cross-lingual summarization aims to generate a summary in a cross-lingual language, which has achieved significant progress (Wang et al., 2022b; Wang et al., 2023a). Generally, besides some work of constructing datasets (Ladhak et al., 2020; Scialom et al., 2020; Yela-Bello et al., 2021; Zhu et al., 2019; Bhattacharjee et al., 2022; Perez-Beltrachini and Lapata, 2021; Varab and Schluter, 2021), existing methods mainly include: the pipeline methods (Leuski et al., 2003; Ouyang et al., 2019; Orăsan and Chiorean, 2008; Wan et al., 2010; Wan, 2011; Yao et al., 2015; Zhang et al., 2016), *i.e.*, translation and then summarization or summarization and then translation, mixed-lingual pre-training (Xu et al., 2020a), knowledge distillation (Nguyen and Tuan, 2021), contrastive learning (Wang et al., 2021a), zero-shot approaches (Ayana et al., 2018; Duan et al., 2019; Dou et al., 2020), and multi-task learning (Zhu et al., 2020b; Takase and Okazaki, 2020; Bai et al., 2021a; Cao et al., 2020b; Cao et al., 2020a; Bai et al., 2021b; Liang et al., 2022d). Wang et al. (2022a) concentrate on building a benchmark dataset for CLS on the dialogue field. We focus on offering additional visual features for multimodal cross-lingual summarization.

2.4 Multilingual Abstractive Summarization

It aims to train a model that can produce a summary in any language. Existing studies mainly pay attention to constructing the multilingual abstractive summarization dataset and there have been many datasets publicly available: MultiLing2015 (Giannakopoulos et al., 2015), GlobalVoices (Nguyen and Daumé III, 2019), MultiSumm (Cao et al., 2020b), MLSUM (Scialom et al., 2020), MultiHumES (Yela-Bello et al., 2021), MassiveSumm (Varab and Schluter, 2021), MLGSum (Wang et al., 2021a), and XL-Sum (Hasan et al., 2021). Most of these datasets are automatically constructed from online websites due to high human cost, which involves at least two languages. Essentially, this line of work is still monolingual while we aim to generate summaries in a cross-lingual manner.

2.5 Knowledge Distillation (KD)

Knowledge distillation (Hinton et al., 2015) is a method to train a model, called the student, by leveraging valuable information provided by soft targets output by another model, called the teacher. In particular, the framework initially trains a model on one designated task to extract useful features. Subsequently, given a dataset $D = \{(X_1, Y_1), (X_2, Y_2), \dots (X_{|D|}, Y_{|D|})\}$, where $|D|$ is the size of the dataset, the teacher model will generate the output $\mathbf{H}_i^T = \{\mathbf{h}_1^T, \mathbf{h}_2^T, \dots, \mathbf{h}_{L_T}^T\}$ for each input X_i. Dependent on the researchers' decision, the output might be hidden representations or final logits. As a consequence, to train the student model, the framework will use a KD loss that discriminates the output of the student model $\mathbf{H}_i^S = \{\mathbf{h}_1^S, \mathbf{h}_2^S, \dots, \mathbf{h}_{L_S}^S\}$ given input X_i from the teacher output $\mathbf{H}_i^T$. Eventually, the KD loss for input X_i will possess the form as follows

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

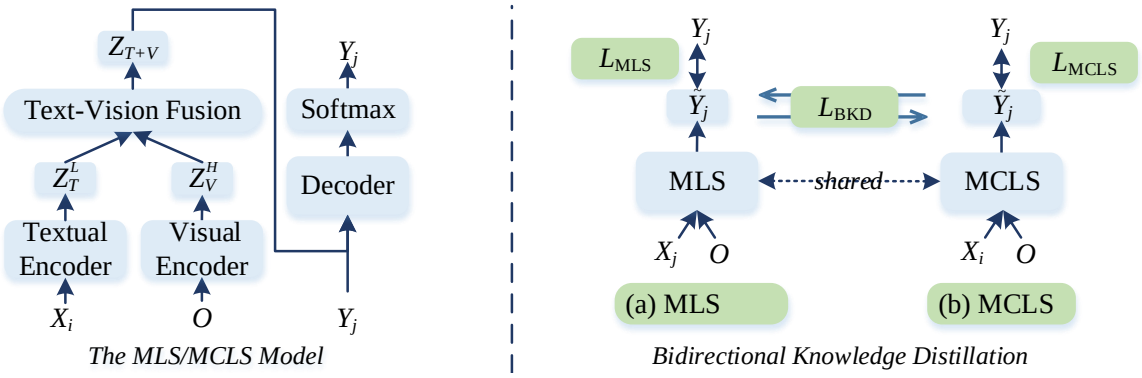

Figure 2: The overview of our model architecture.

$$\mathcal{L}_{\text{KD}} = dist(\mathbf{H}_i^T, \mathbf{H}_i^S), \tag{1}$$

where $dist$ is a distance function to estimate the discrepancy of teacher and student outputs.

The explicated knowledge distillation framework has shown its effectiveness in many NLP tasks, such as question answering (Hu et al., 2018; Arora et al., 2019; Yang et al., 2020) and neural machine translation (Tan et al., 2019; Wang et al., 2021b; Li and Li, 2021; Sun et al., 2020; Zhang et al., 2023). Nonetheless, its application for multimodal cross-lingual summarization has received little interest.

3 Method

3.1 Problem Formulation

Given an input article $\mathcal{X}_{L1}=\{x_k\}_{k=1}^{|\mathcal{X}_{L1}|}$ in the source language and the corresponding object sequence $\mathcal{O}=\{o_{ij}\}_{i=1,j=1}^{i\leq n,j\leq m}$, where x_k denotes the k-th token and o_{ij} represents the detected j-th object of the i-th image (n, m is the number of images and detected objects in each image, respectively), the MCLS task is defined as:

$$p(\mathcal{Y}_{L2}|\mathcal{X}_{L1}, \mathcal{O}) = \prod_{t=1}^{|\mathcal{Y}_{L2}|} p(y_t|\mathcal{X}_{L1}, \mathcal{O}, y_{<t}),$$

where $y_{<t}$ indicates the previous tokens before the t-th time step of the summary $\mathcal{Y}_{L2}=\{y_t\}_{t=1}^{|\mathcal{Y}_{L2}|}$ in target language and $L_1 \neq L_2$.

3.2 The MCLS Model

Yu et al. (2021a) design a text-vision fusion method to inject the visual features into the generative pre-trained language models (*e.g.*, BART), which achieves state-of-the-art performance on MAS (Liang et al., 2022b). As shown in the left part of Figure 2, the backbone of the MAS model is a variant of transformer (Vaswani et al., 2017) with four modules: textual encoder, visual encoder, text-vision fusion, and decoder.

Textual Encoder. The input text $\mathcal{X}_{L1}$ is firstly tokenized and mapped to a sequence of token embeddings $\mathbf{X}$. Then, the positional encodings $\mathbf{E}_{pe}$ are pointwisely added to $\mathbf{X}$ to keep the positional information (Vaswani et al., 2017):

$$\mathbf{Z}_T^0 = \mathbf{X} + \mathbf{E}_{pe},\ \{\mathbf{Z}_T^0, \mathbf{X}, \mathbf{E}_{pe}\} \in \mathbb{R}^{|\mathcal{X}_{L1}|\times d},$$

where d is the feature dimension. It forms the input features $\mathbf{Z}_T^0$ to the encoder, which consists of L stacked layers and each layer includes two sub-layers: 1) Multi-Head Attention (MHA) and 2) a position-wise Feed-Forward Network (FFN):

$$\mathbf{S}_T^l = \text{MHA}(\mathbf{Z}_T^{l-1}) + \mathbf{Z}_T^{l-1},\ \mathbf{S}_T^l \in \mathbb{R}^{|\mathcal{X}_{L1}|\times d},$$
$$\mathbf{Z}_T^l = \text{FFN}(\mathbf{S}_T^l) + \mathbf{S}_T^l,\ \mathbf{Z}_T^l \in \mathbb{R}^{|\mathcal{X}_{L1}|\times d},$$

CCL 2023

where $\mathbf{Z}_T^l$ is the state of the l-th encoder layer.

Visual Encoder. Following previous work (Liang et al., 2021; Liang et al., 2022a; Liang et al., 2022c), the object sequence $\mathcal{O}$ is typically extracted from the image by the Faster R-CNNs (Ren et al., 2015) (actually, we have several images instead of only one image. Then the visual features are fed into the visual encoder with H layers. Finally, we obtain the output visual features $\mathbf{Z}_V^H$:

$$\begin{aligned}\mathbf{S}_V^h &= \mathrm{MHA}(\mathbf{Z}_V^{h-1}) + \mathbf{Z}_V^{h-1},\ \mathbf{S}_V^h \in \mathbb{R}^{|\mathcal{O}|\times d_v},\\ \mathbf{Z}_V^h &= \mathrm{FFN}(\mathbf{S}_V^h) + \mathbf{S}_V^h,\ \mathbf{Z}_V^h \in \mathbb{R}^{|\mathcal{O}|\times d_v},\end{aligned}$$

where $\mathbf{Z}_V^0$ is the extracted visual features $\mathbf{O}$.

Text-Vision Fusion. The fusion method is vision-guided multi-head attention (Yu et al., 2021a). Firstly, the query $\mathbf{Q}$ is linearly projected from the textual features $\mathbf{Z}_T^L$, and the key $\mathbf{K}$ and value $\mathbf{V}$ are linearly projected from the visual features $\mathbf{Z}_V^H$. Secondly, a Cross-modal Multi-Head Attention (CMHA) is applied to get the text queried visual features $\mathbf{M}$. Then, a forget gate $\mathbf{G}$ is used to filter redundant and noisy information from the visual features. Finally, we obtain the vision-guided output $\mathbf{Z}_{T+V}$ by concatenating the textual features $\mathbf{Z}_T^L$ and the result of a point-wise multiplication $\mathbf{G}\otimes\mathbf{M}$, and then linearly project it to the original dimension d. Formally, the text-vision fusion process is:

$$\begin{aligned}\mathbf{Q} &= \mathbf{Z}_T^L\mathbf{W}_q,\ \mathbf{Q}\in\mathbb{R}^{|\mathcal{X}_{L1}|\times d_c},\\ \mathbf{K} &= \mathbf{Z}_V^H\mathbf{W}_k,\ \mathbf{V}=\mathbf{Z}_V^H\mathbf{W}_v,\ \mathbf{K},\mathbf{V}\in\mathbb{R}^{|\mathcal{O}|\times d_c},\\ \mathbf{M} &= \mathrm{CMHA}(\mathbf{Q},\mathbf{K},\mathbf{V}),\ \mathbf{M}\in\mathbb{R}^{|\mathcal{X}_{L1}|\times d_c},\\ \mathbf{G} &= \mathrm{Sigmoid}(\mathrm{Concat}(\mathbf{Z}_T^L,\mathbf{M})\mathbf{W}_g+\mathbf{b}_g),\\ \mathbf{Z}_{T+V} &= \mathrm{Concat}(\mathbf{Z}_T^L,\mathbf{G}\otimes\mathbf{M})\mathbf{W}_z+\mathbf{b}_z,\end{aligned}$$

where Concat is the concatenation operation and $\mathbf{W}_*$ and $\mathbf{b}_*$ are trainable weights.

Decoder. Similar to the encoder, but each of L decoder layers includes an additional Multi-Head Cross-Attention sub-layer (MHCA):

$$\begin{aligned}\mathbf{S}_{dec}^l &= \mathrm{MHA}(\mathbf{Z}_{dec}^{l-1}) + \mathbf{Z}_{dec}^{l-1},\ \mathbf{S}_{dec}^{l-1}\in\mathbb{R}^{|\mathcal{Y}_{L2}|\times d},\\ \mathbf{C}_{dec}^l &= \mathrm{MHCA}(\mathbf{S}_{dec}^l,\mathbf{Z}_{T+V}) + \mathbf{S}_{dec}^l,\\ \mathbf{Z}_{dec}^l &= \mathrm{FFN}(\mathbf{C}_{dec}^l) + \mathbf{C}_{dec}^l,\ \mathbf{C}_{dec}^l\in\mathbb{R}^{|\mathcal{Y}_{L2}|\times d},\end{aligned}\tag{2}$$

where $\mathbf{Z}_{dec}^l\in\mathbb{R}^{|\mathcal{Y}_{L2}|\times d}$ denotes the state of the l-th decoder layer. Then, at each decoding time step t, the top-layer (L-th) decoder hidden state $\mathbf{Z}_{dec,t}^L$ is fed into the softmax layer to produce the probability distribution of the next target token as:

$$p(y_t|\mathcal{X}_{L1},\mathcal{O},y_{<t}) = \mathrm{Softmax}(\mathbf{W}_o\mathbf{Z}_{dec,t}^L+\mathbf{b}_o),$$

where $\mathbf{W}_o$ and $\mathbf{b}_o$ are trainable weights.

3.3 Bidirectional Knowledge Distillation

Our framework is shown in the right part of Figure 2, where we initiate the process by training the teacher model on the multimodal monolingual summarization task. In detail, given an input $X^{L_1}=\{x_1,x_2,\ldots,x_N\}$ and corresponding image features, the teacher model will aim to generate its monolingual summary $Y^{L_1}=\{y_1^{L_1},y_2^{L_1},\ldots,y_{M_1}^{L_1}\}$. Similar to previous multimodal monolingual summarization schemes, our model is trained with the cross-entropy loss:

$$\mathcal{L}_{\mathrm{MLS}} = -\sum_{t=1}^{|\mathcal{Y}_{L1}|}\log(p(y_t^{L_1}|y_{<t}^{L_1},\mathcal{X}^{L_1},\mathcal{O})).\tag{3}$$

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

After finetuning the teacher model, we progress to train the student model, which also uses the Transformer architecture. Contrary to the teacher, the student model's task is to generate the cross-lingual output $Y^{L_2} = \{y_1^{L_2}, y_2^{L_2}, \ldots, y_{M_2}^{L_2}\}$ in language L_2, given the input document X^{L_1} in language L_1 and corresponding image features. We update the parameters of the student model by another cross-entropy loss:

$$\mathcal{L}_{\text{MCLS}} = -\sum_{t=1}^{|\mathcal{Y}_{L2}|} \log(p(y_t^{L_2}|y_{<t}^{L_2}, \mathcal{X}^{L_1}, \mathcal{O}). \tag{4}$$

To pull the cross-lingual and monolingual representations nearer, we implement a KD loss to penalize the large distance of two vector spaces. Specifically, let $\mathbf{H}^T = \{\mathbf{h}_1^T, \mathbf{h}_2^T, \ldots, \mathbf{h}_{L_T}^T\}$ denote the contextualized representations produced by the decoder of the teacher model, and $\mathbf{H}^S = \{\mathbf{h}_1^S, \mathbf{h}_2^S, \ldots, \mathbf{h}_{L_S}^S\}$ denote the representations from the decoder of the student model, our KD loss are defined as:

$$\mathcal{L}_{\text{KD}} = dist(\mathbf{H}^T, \mathbf{H}^S), \tag{5}$$

where $dist$ is the distance function to evaluate the difference between two representations (*e.g.*, KL, and cosine similarity). Conversely, when the student model achieves better performance, we also distill its knowledge into the teacher model. Therefore, the knowledge between the teacher and student models can be transferred to each other and thus enhance both of them. The bidirectional knowledge distillation loss function can be defined as:

$$\mathcal{L}_{\text{BKD}} = dist(\mathbf{H}^T, \mathbf{H}^S) + dist(\mathbf{H}^S, \mathbf{H}^T). \tag{6}$$

3.4 Training and Inference

For training, the model can deal with inputs in multiple languages and predict the summary in the corresponding language. Specifically, for each language L_k in the set of K languages $Lang = \{L_1, l_2, ..., L_K\}$, the training objective is:

$$\mathcal{J} = \sum_{k=1}^{K} (\mathcal{L}_{\text{MLS}}^{L_k} + \mathcal{L}_{\text{MCLS}}^{L_k} + \alpha * \mathcal{L}_{\text{BKD}}). \tag{7}$$

During inference, the BKD is not involved and only the MLS or MCLS model is used to conduct summarization.

4 Experiments

4.1 MM-CLS Dataset

There is no large-scale multimodal cross-lingual benchmark dataset until now. We construct one as follows.

Data Source and Data Construction. Based on the CrossSum dataset (Bhattacharjee et al., 2022), we construct our **M**ulti**M**odal **C**ross-**L**ingual **S**marization (MM-CLS) dataset. The original CrossSum dataset is automatically crawled from the BBC website[1]. However, the lacking of the associated image sequence in CrossSum, makes it impossible to directly conduct research on multimodal cross-lingual summarization. Therefore, we strictly follow the procedure of Bhattacharjee et al. (2022) to crawl the images for the corresponding textual summarization dataset given the article URL, where we maintain the article-summary pair if it contains images and keep the image order that appeared in the article.

Dataset Statistics and Splits. Table 4 of Appendix A shows that our MM-CLS covers 44 languages and totally includes 1,073,301 article-summary pairs with 3,381,456 images, where each article-summary pair contains about 3.15 images on average. According to the language directions, we select six languages and conduct experiments in the many-to-many setting.

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

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Src \ Trg	Models	English	French	Hindi	Chinese	Japanese	Russian
English	mT5	35.80 / 13.45 / 27.99	31.29 / 11.17 / 22.28	33.22 / 11.72 / 26.20	29.49 / 15.24 / 23.85	30.62 / 15.02 / 23.94	24.47 / 8.22 / 19.88
	VG-mT5	36.08 / 13.84 / 28.23	31.67 / 11.56 / 22.77	33.47 / 11.98 / 26.58	29.88 / 15.76 / 24.34	30.99 / 15.54 / 24.61	24.85 / 8.77 / 20.44
	VG-mT5+BKD (Ours)	**36.85 / 14.51 / 29.44**	**32.55 / 12.45 / 23.67**	**34.67 / 13.48 / 27.89**	**30.49 / 17.13 / 25.67**	**31.86 / 16.74 / 25.87**	**25.88 / 9.88 / 21.58**
French	mT5	23.29 / 8.75 / 18.66	38.31 / 19.19 / 29.21	22.11 / 7.44 / 18.41	25.45 / 11.21 / 18.55	26.78 / 12.44 / 20.01	23.44 / 7.47 / 18.42
	VG-mT5	23.80 / 8.99 / 18.99	38.53 / 19.59 / 29.67	22.45 / 7.93 / 18.85	25.78 / 11.56 / 18.93	26.99 / 12.78 / 20.56	23.83 / 7.82 / 18.90
	VG-mT5+BKD (Ours)	**24.72 / 9.45 / 19.78**	**39.79 / 20.24 / 30.66**	**23.62 / 8.95 / 19.77**	**26.91 / 13.04 / 19.89**	**28.18 / 14.21 / 22.05**	**24.91 / 9.05 / 20.31**
Hindi	mT5	27.05 / 11.67 / 21.72	22.11 / 7.16 / 17.28	36.41 / 14.82 / 27.34	26.12 / 11.59 / 19.89	21.32 / 9.21 / 16.78	22.11 / 7.41 / 16.11
	VG-mT5	27.62 / 11.99 / 22.07	22.34 / 7.45 / 17.61	36.84 / 15.25 / 27.76	26.54 / 11.87 / 20.21	21.67 / 9.56 / 17.15	22.60 / 7.88 / 16.70
	VG-mT5+BKD (Ours)	**28.34 / 13.07 / 23.24**	**23.52 / 8.41 / 18.78**	**37.49 / 16.56 / 29.04**	**27.54 / 13.11 / 20.99**	**22.87 / 10.56 / 18.86**	**23.83 / 8.41 / 17.40**
Chinese	mT5	29.10 / 13.08 / 27.37	26.29 / 11.17 / 21.28	27.70 / 12.12 / 22.22	33.47 / 15.24 / 28.81	28.60 / 13.06 / 21.95	22.81 / 7.49 / 16.42
	VG-mT5	29.49 / 13.52 / 27.78	26.56 / 11.57 / 21.71	27.92 / 12.71 / 22.55	33.91 / 15.60 / 29.23	28.87 / 13.55 / 22.19	23.11 / 7.90 / 16.82
	VG-mT5+BKD (Ours)	**30.54 / 14.51 / 28.29**	**27.45 / 13.07 / 23.16**	**28.83 / 13.79 / 23.71**	**35.38 / 16.82 / 30.84**	**30.68 / 15.01 / 23.88**	**23.99 / 8.89 / 17.58**
Japanese	mT5	29.97 / 14.18 / 24.44	24.22 / 9.15 / 18.25	25.21 / 10.72 / 21.20	24.49 / 11.21 / 18.80	39.60 / 18.08 / 33.91	25.04 / 8.44 / 20.44
	VG-mT5	30.31 / 14.54 / 24.93	24.62 / 9.56 / 18.70	25.63 / 10.95 / 21.57	24.81 / 11.62 / 19.09	39.97 / 18.50 / 34.33	25.60 / 8.92 / 20.87
	VG-mT5+BKD (Ours)	**31.57 / 15.78 / 25.77**	**25.86 / 10.59 / 19.77**	**26.78 / 12.17 / 22.45**	**25.66 / 12.33 / 19.98**	**40.97 / 19.41 / 35.16**	**26.77 / 9.49 / 21.89**
Russian	mT5	29.47 / 9.86 / 22.82	25.28 / 10.17 / 20.26	28.01 / 11.28 / 26.51	27.49 / 13.24 / 20.85	27.62 / 12.02 / 20.94	29.32/11.32/23.72
	VG-mT5	29.89 / 10.05 / 23.18	25.67 / 10.51 / 20.60	28.60 / 11.57 / 26.97	27.91 / 13.65 / 21.28	27.98 / 12.55 / 21.46	29.66/11.70/24.12
	VG-mT5+BKD (Ours)	**30.56 / 11.18 / 24.13**	**26.76 / 11.45 / 21.85**	**29.45 / 12.88 / 27.59**	**28.88 / 14.41 / 22.87**	**28.88 / 14.01 / 22.91**	**30.93/12.88/24.87**

Table 1: Results on MM-CLS (ROUGE-1 / ROUGE-2 / ROUGE-L).

4.2 Implementation Details and Metrics

Data Pre-Processing. Following Bhattacharjee et al. (2022), we pre-process the textual data by truncating or padding them into sequences of 512 tokens for $\mathcal{X}$ and the outputs $\mathcal{Y}$ to 84 tokens after using the 250k wordpiece (Xue et al., 2021) vocabulary provided with the mT5 checkpoint. For the image sequence, we also truncate or pad the sequence length to 180 (*i.e.*, five images: 5 * 36; n=5, m=36).

Hyper-Parameters. Following Bhattacharjee et al. (2022), we use the $base$[2] model of mT5 (Xue et al., 2021), in which $L = 12$ for both encoder and decoder. For the vision-related hyper-parameters mentioned in subsection 3.2, we follow Yu et al. (2021a) for a fair comparison. Specifically, we use a 4-layer encoder (*i.e.*, $H = 4$) with 8 attention heads and a 2048 feed-forward dimension. For all models, the dropout is set to 0.1 and the label smoothing is set to 0.1. The d, d_c, and d_v are 768, 256, and 2048, respectively. During the training, following a similar training strategy (Conneau and Lample, 2019; Bhattacharjee et al., 2022), we sample each batch from a single language containing 256 samples and use a smoothing factor (0.5) so that batches of low-resource languages would be sampled at a higher rate, increasing their frequency during training. We set the training step to 35,000 steps on a distributed cluster of 8 NVIDIA Tesla V100 GPUs and trained for about 5 days. We use the Adafactor optimizer (Shazeer and Stern, 2018) with a linear warm-up of 5,000 steps and the "inverse square root" learning rate schedule.

For inference, we use beam search with beam size 4 and length penalty of $\gamma = 0.6$. When calculating the ROUGE scores, we use the multi-lingual rouge[3] toolkit following Hasan et al. (2021). All experimental results reported in this paper are the average of three runs with different random seeds.

Metrics. Following Bhattacharjee et al. (2022), we use the standard ROUGE scores (R-1, R-2, and R-L) (Lin, 2004) with the statistical significance test (Koehn, 2004) for a fair comparison.

4.3 Comparison Models

Text-Only MAS Systems.

mT5: We choose the mT5 (Xue et al., 2021), a multilingual language model pre-trained on a large dataset of 101 languages, as the text-only baseline which is fine-tuned on our dataset.

Vision-Guided MAS Systems.

VG-mT5: We implement the fusion method described in subsection 3.2 to inject visual features into the mT5 model, which is a strong baseline.

VG-mT5+BKD (Ours): It is the proposed model where we design two summary-oriented vision modeling tasks to enhance the VG-mT5 model.

[2] https://huggingface.co/google/mt5-base/tree/main

[3] https://github.com/csebuetnlp/xl-sum/tree/master/multilingual_rouge_scoring

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

	Models	English→*	French→*	Hindi→*	Japanese→*	Russian→*	Chinese→*
0	Baseline (VG-mT5)	31.15/12.90/24.49	26.89/11.44/20.93	26.26/10.66/20.25	28.31/12.47/23.38	28.49/12.34/23.24	28.28/11.67/22.93
1	*w/* $\mathcal{L}_{\text{MLS}}^{L_k}$	31.62/13.41/24.92	27.45/11.86/21.45	26.69/11.06/20.77	28.87/12.88/23.81	28.66/12.58/23.66	28.51/11.99/23.35
2	*w/* $\mathcal{L}_{\text{BKD}}$	31.75/13.77/25.04	27.80/11.99/21.80	26.89/11.35/21.02	28.99/13.37/24.13	28.96/12.82/23.92	28.65/12.27/23.59
3	*w/* $\mathcal{L}_{\text{MLS}}^{L_k}$&$\mathcal{L}_{\text{BKD}}$	32.05/14.03/25.68	28.02/12.49/22.07	27.26/11.68/21.38	29.47/13.68/24.57	29.60/13.29/24.17	29.24/12.80/24.04

Table 2: Ablation results under different language directions (Avg. R-1/R-2/R-L results), where each loss is separately added on the baseline.

Models	Chinese→English			English→Chinese		
	Fluency	Conciseness	Informativeness	Fluency	Conciseness	Informativeness
mT5	4.21	3.54	3.04	3.56	3.14	3.04
VG-mT5	4.44	3.68	3.26	3.82	3.36	3.22
VG-mT5+BKD (Ours)	**4.26**	**4.38**	**3.76**	**4.32**	**3.88**	**3.68**

Table 3: Human evaluation results.

4.4 Main Results

Table 1 present the main results on many-to-many scenarios. Overall, our model obtains notably better results than the text-only "mT5" model and the vision-guided "VG-mT5" model no matter if it is the MLS or MCLS setting. Compared with the text-only model, the VG-mT5 model can substantially surpass it, showing that the vision plays a vital role and suggesting the value of our MM-Sum dataset. After adding the BKD approach, the model performance obtains further significant improvement, up to **1.35/0.92/1.42** ROUGE scores on average, showing the effectiveness of our proposed approach.

5 Analysis

5.1 Ablation Study

We conduct ablation studies to investigate how well the two auxiliary tasks work. The results are shown in Table 2. We have the following findings:

- The MLS task shows a positive impact on the model performance (row 1 vs. row 0), demonstrating that the knowledge of MLS can be transferred to MCLS, which is beneficial to the summary generation;
- The BKD substantially improves the MCLS model in terms of ROUGE scores (row 2 vs. row 0), suggesting that transferring knowledge into each other is helpful for summarization;
- The two loss functions exhibit notable cumulative benefits (row 3 vs. rows 0∼2), showing that transferring the knowledge of MLS to the MCLS is effective;

5.2 Human Evaluation

To further evaluate the performances of mT5, VG-mT5 and our VG-mT5+BKD, we conduct human studies on 50 samples randomly selected from English and Chinese test sets. We invite three Chinese postgraduate students who highly proficient in English comprehension to compare the generated summaries under the multilingual training setting, and assess each summary from three independent perspectives: **fluency**, **conciseness** and **informativeness**. We ask them to assess each aspect with a score ranging from 1 (worst) to 5 (best). The average results are presented in Table 3.

Table 3 show the human results on Chinese→English and English→Chinese. We find that our model outperforms all comparison models from all criteria in both languages, which further demonstrates the effectiveness and superiority of our model. The Fleiss' Kappa scores (Fleiss and Cohen, 1973) of Flu., Conci and Info. are 0.72, 0.68 and 0.59, respectively, which indicates a substantial agreement among three evaluators.

CCL 2023

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

6 Conclusion and Future Work

In this paper, we propose to benchmark the MCLS task and provide a large-scale MM-CLS dataset. We also propose a bidirectional knowledge distillation approach, which can explicitly enhance the knowledge transferring between VG-mT5 and MCLS, and thus improve the summary quality. Extensive experiments on multiple settings, show that our model significantly outperforms related baselines in terms of ROUGE scores. In the future, due to the difficulty of simultaneously learning cross-lingual alignment and cross-modal alignment, future work should focus on these directions.

Acknowledgements

The authors would like to thank the anonymous reviewers for their insightful comments and suggestions to improve this paper.

References

Siddhartha Arora, Mitesh M Khapra, and Harish G Ramaswamy. 2019. On knowledge distillation from complex networks for response prediction. In *NAACL*, pages 3813–3822.

Ayana, shi-qi Shen, Yun Chen, Cheng Yang, Zhi-yuan Liu, and Maosong Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM TASLP*, 26(12):2319–2327.

Yu Bai, Yang Gao, and Heyan Huang. 2021a. Cross-lingual abstractive summarization with limited parallel resources. In *ACL-IJCNLP*, pages 6910–6924.

Yu Bai, Heyan Huang, Kai Fan, Yang Gao, Zewen Chi, and Boxing Chen. 2021b. Bridging the gap: Cross-lingual summarization with compression rate. *CoRR*, abs/2110.07936.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2022. Crosssum: Beyond english-centric cross-lingual abstractive text summarization for 1500+ language pairs.

Yue Cao, Hui Liu, and Xiaojun Wan. 2020a. Jointly learning to align and summarize for neural cross-lingual summarization. In *ACL*, pages 6220–6231.

Yue Cao, Xiaojun Wan, Jinge Yao, and Dian Yu. 2020b. Multisumm: Towards a unified model for multi-lingual abstractive summarization. In *AAAI*, volume 34, pages 11–18, Apr.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *EMNLP*, pages 4046–4056.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *NIPS*.

Zi-Yi Dou, Sachin Kumar, and Yulia Tsvetkov. 2020. A deep reinforced model for zero-shot cross-lingual summarization with bilingual semantic similarity rewards. In *NGT*, pages 60–68.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *ACL*, pages 3162–3172.

B. Erol, D.-S. Lee, and J. Hull. 2003. Multimodal summarization of meeting recordings. In *ICME*, pages III–25.

Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, 15(7):1553–1568.

Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and Psychological Measurement*, pages 613–619.

Xiyan Fu, Jun Wang, and Zhenglu Yang. 2021. MM-AVS: A full-scale dataset for multi-modal summarization. In *NAACL*, pages 5922–5926.

George Giannakopoulos, Jeff Kubina, John Conroy, Josef Steinberger, Benoit Favre, Mijail Kabadjov, Udo Kruschwitz, and Massimo Poesio. 2015. MultiLing 2015: Multilingual summarization of single and multi-documents, on-line fora, and call-center conversations. In *SIGDIAL*, pages 270–274.

CCL 2023

Travis Goodwin, Max Savery, and Dina Demner-Fushman. 2020. Flight of the PEGASUS? comparing transformers on few-shot and zero-shot multi-document abstractive summarization. In *COLING*, pages 5640–5646.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of ACL-IJCNLP*, pages 4693–4703.

Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. In *NIPS*, page 1693–1701.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Minghao Hu, Yuxing Peng, Furu Wei, Zhen Huang, Dongsheng Li, Nan Yang, and Ming Zhou. 2018. Attention-guided answer distillation for machine reading comprehension. *arXiv preprint arXiv:1808.07644*.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *EMNLP*, pages 388–395.

Faisal Ladhak, Esin Durmus, Claire Cardie, and Kathleen McKeown. 2020. WikiLingua: A new benchmark dataset for cross-lingual abstractive summarization. In *Findings of EMNLP*, pages 4034–4048.

Anton Leuski, Chin-Yew Lin, Liang Zhou, Ulrich Germann, Franz Josef Och, and Eduard Hovy. 2003. Cross-lingual c*st*rd: English access to hindi information. *ACM TALIP*, 2(3):245–269.

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 *ACL*, pages 7871–7880.

Yongqi Li and Wenjie Li. 2021. Data distillation for text classification. *arXiv preprint arXiv:2104.08448*.

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 *EMNLP*, pages 1092–1102.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, Chengqing Zong, et al. 2018a. Multi-modal sentence summarization with modality attention and image filtering. In *IJCAI*, pages 4152–4158.

Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhang, and Chengqing Zong. 2018b. Read, watch, listen, and summarize: Multi-modal summarization for asynchronous text, image, audio and video. *IEEE TKDE*, 31(5):996–1009.

Haoran Li, Peng Yuan, Song Xu, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020a. Aspect-aware multimodal summarization for chinese e-commerce products. In *AAAI*, volume 34, pages 8188–8195.

Mingzhe Li, Xiuying Chen, Shen Gao, Zhangming Chan, Dongyan Zhao, and Rui Yan. 2020b. VMSMO: Learning to generate multimodal summary for video-based news articles. In *EMNLP*, pages 9360–9369.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. *AAAI*, pages 13343–13352, May.

Yunlong Liang, Fandong Meng, Jinan Xu, Yufeng Chen, and Jie Zhou. 2022a. MSCTD: A multimodal sentiment chat translation dataset. In *ACL*, pages 2601–2613.

Yunlong Liang, Fandong Meng, Jinan Xu, Jiaan Wang, Yufeng Chen, and Jie Zhou. 2022b. Summary-oriented vision modeling for multimodal abstractive summarization. *arXiv preprint arXiv:2212.07672*.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2022c. Emotional conversation generation with heterogeneous graph neural network. *Artificial Intelligence*, 308:103714.

Yunlong Liang, Fandong Meng, Chulun Zhou, Jinan Xu, Yufeng Chen, Jinsong Su, and Jie Zhou. 2022d. A variational hierarchical model for neural cross-lingual summarization. In *ACL*, pages 2088–2099.

Yunlong Liang, Fandong Meng, Jiaan Wang, Jinan Xu, Yufeng Chen, and Jie Zhou. 2023. D2tv: Dual knowledge distillation and target-oriented vision modeling for many-to-many multimodal summarization. *arXiv preprint arXiv:2305.12767*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *TSBO*, pages 74–81.

CCL 2023

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. In *EMNLP-IJCNLP*, pages 3730–3740.

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 *EMNLP*, pages 1834–1845.

Nayu Liu, Kaiwen Wei, Xian Sun, Hongfeng Yu, Fanglong Yao, Li Jin, Guo Zhi, and Guangluan Xu. 2022. Assist non-native viewers: Multimodal cross-lingual summarization for how2 videos. In *EMNLP*, pages 6959–6969.

Khanh Nguyen and Hal Daumé III. 2019. Global Voices: Crossing borders in automatic news summarization. In *NFS*, pages 90–97.

Thong Nguyen and Luu Anh Tuan. 2021. Improving neural cross-lingual summarization via employing optimal transport distance for knowledge distillation. *CoRR*, abs/2112.03473.

Constantin Orăsan and Oana Andreea Chiorean. 2008. Evaluation of a cross-lingual Romanian-English multi-document summariser. In *LREC*.

Jessica Ouyang, Boya Song, and Kathy McKeown. 2019. A robust abstractive system for cross-lingual summarization. In *NAACL*, pages 2025–2031.

Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *ACL*, pages 6587–6596.

Romain Paulus, Caiming Xiong, and Richard Socher. 2018. A deep reinforced model for abstractive summarization. In *ICLR*.

Laura Perez-Beltrachini and Mirella Lapata. 2021. Models and datasets for cross-lingual summarisation. In *EMNLP*, pages 9408–9423.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *NIPS*, volume 28.

Sascha Rothe, Joshua Maynez, and Shashi Narayan. 2021. A thorough evaluation of task-specific pretraining for summarization. In *EMNLP*, pages 140–145.

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 *ViGIL*.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *EMNLP*, pages 8051–8067.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost. In Jennifer Dy and Andreas Krause, editors, *ICML*, volume 80, pages 4596–4604.

Haipeng Sun, Rui Wang, Kehai Chen, Masao Utiyama, Eiichiro Sumita, and Tiejun Zhao. 2020. Knowledge distillation for multilingual unsupervised neural machine translation. *arXiv preprint arXiv:2004.10171*.

Sho Takase and Naoaki Okazaki. 2020. Multi-task learning for cross-lingual abstractive summarization.

Xu Tan, Yi Ren, Di He, Tao Qin, Zhou Zhao, and Tie-Yan Liu. 2019. Multilingual neural machine translation with knowledge distillation. *arXiv preprint arXiv:1902.10461*.

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 *IEEE WACV*, pages 471–478.

Daniel Varab and Natalie Schluter. 2021. MassiveSumm: a very large-scale, very multilingual, news summarisation dataset. In *EMNLP*, pages 10150–10161.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, pages 5998–6008.

Xiaojun Wan, Huiying Li, and Jianguo Xiao. 2010. Cross-language document summarization based on machine translation quality prediction. In *ACL*, pages 917–926, Uppsala, Sweden.

Xiaojun Wan. 2011. Using bilingual information for cross-language document summarization. In *ACL*, pages 1546–1555.

CCL 2023

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Danqing Wang, Jiaze Chen, Hao Zhou, Xipeng Qiu, and Lei Li. 2021a. Contrastive aligned joint learning for multilingual summarization. In *Findings of ACL-IJCNLP*, pages 2739–2750.

Fusheng Wang, Jianhao Yan, Fandong Meng, and Jie Zhou. 2021b. Selective knowledge distillation for neural machine translation. *arXiv preprint arXiv:2105.12967*.

Jiaan Wang, Fandong Meng, Ziyao Lu, Duo Zheng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022a. Clidsum: A benchmark dataset for cross-lingual dialogue summarization. *arXiv preprint arXiv:2202.05599*.

Jiaan Wang, Fandong Meng, Tingyi Zhang, Yunlong Liang, Jiarong Xu, Zhixu Li, and Jie Zhou. 2022b. Understanding translationese in cross-lingual summarization. *arXiv preprint arXiv:2212.07220*.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2022c. A survey on cross-lingual summarization. *Transactions of the Association for Computational Linguistics*, 10:1304–1323.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023a. Cross-lingual summarization via chatgpt. *arXiv preprint arXiv:2302.14229*.

Jiaan Wang, Fandong Meng, Duo Zheng, Yunlong Liang, Zhixu Li, Jianfeng Qu, and Jie Zhou. 2023b. Towards unifying multi-lingual and cross-lingual summarization. *arXiv preprint arXiv:2305.09220*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *ACL*, pages 5245–5263.

Ruochen Xu, Chenguang Zhu, Yu Shi, Michael Zeng, and Xuedong Huang. 2020a. Mixed-lingual pre-training for cross-lingual summarization. In *AACL*, pages 536–541, Suzhou, China.

Song Xu, Haoran Li, Peng Yuan, Youzheng Wu, Xiaodong He, and Bowen Zhou. 2020b. Self-attention guided copy mechanism for abstractive summarization. In *ACL*, pages 1355–1362.

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 *NAACL*, pages 483–498.

Ze Yang, Linjun Shou, Ming Gong, Wutao Lin, and Daxin Jiang. 2020. Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In *WSDM*, pages 690–698.

Jin-ge Yao, Xiaojun Wan, and Jianguo Xiao. 2015. Phrase-based compressive cross-language summarization. In *EMNLP*, pages 118–127.

Jenny Paola Yela-Bello, Ewan Oglethorpe, and Navid Rekabsaz. 2021. MultiHumES: Multilingual humanitarian dataset for extractive summarization. In *EACL*, pages 1713–1717.

Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. 2021a. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *EMNLP*, pages 3995–4007.

Tiezheng Yu, Zihan Liu, and Pascale Fung. 2021b. AdaptSum: Towards low-resource domain adaptation for abstractive summarization. In *NAACL*, pages 5892–5904.

Jiajun Zhang, Yu Zhou, and Chengqing Zong. 2016. Abstractive cross-language summarization via translation model enhanced predicate argument structure fusing. *IEEE/ACM TASLP*, 24(10):1842–1853.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. PEGASUS: Pre-training with extracted gap-sentences for abstractive summarization. In *ICML*, volume 119, pages 11328–11339.

Litian Zhang, Xiaoming Zhang, Junshu Pan, and Feiran Huang. 2021a. Hierarchical cross-modality semantic correlation learning model for multimodal summarization. *arXiv preprint arXiv:2112.12072*.

Zhengkun Zhang, Xiaojun Meng, Yasheng Wang, Xin Jiang, Qun Liu, and Zhenglu Yang. 2021b. Unims: A unified framework for multimodal summarization with knowledge distillation. *arXiv preprint arXiv:2109.05812*.

Songming Zhang, Yunlong Liang, Shuaibo Wang, Wenjuan Han, Jian Liu, Jinan Xu, and Yufeng Chen. 2023. Towards understanding and improving knowledge distillation for neural machine translation. *arXiv preprint arXiv:2305.08096*.

Junnan Zhu, Haoran Li, Tianshang Liu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2018. MSMO: Multimodal summarization with multimodal output. In *EMNLP*, pages 4154–4164.

CCL 2023

Proceedings of the 22nd China National Conference on Computational Linguistics, pages 862-874, Harbin, China, August 3 - 5, 2023.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. NCLS: Neural cross-lingual summarization. In *EMNLP-IJCNLP*, pages 3054–3064, Hong Kong, China.

Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020a. Multimodal summarization with guidance of multimodal reference. In *AAAI*, volume 34, pages 9749–9756.

Junnan Zhu, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2020b. Attend, translate and summarize: An efficient method for neural cross-lingual summarization. In *ACL*, pages 1309–1321.

Junnan Zhu, Lu Xiang, Yu Zhou, Jiajun Zhang, and Chengqing Zong. 2021. Graph-based multimodal ranking models for multimodal summarization. *TALLIP*, 20(4):1–21.

A Dataset Statistics.

Due to space limit, here we show 6 * 5 language pairs in Table 4. In fact, we construct the MM-CLS dataset based on CrossSum (Bhattacharjee et al., 2022) where 62% data of CrossSum are maintained. Therefore, our MM-CLS covers 44 * 43 language pairs and totally includes 1,073,301 article-summary pairs with 3,381,456 images, where each article-summary pair contains about 3.15 images on average. The average article and summary length for all languages is about 520 and 84, respectively.

Languages	English	French	Hindi	Chinese	Japanese	Russian
English	-	1,881	4,256	4,561	2,447	7,854
French	1,881	-	546	288	256	656
Hindi	4,256	546	-	1,234	5,23	4,256
Chinese	4,561	288	1,234	-	956	2,432
Japanese	2,447	256	523	956	-	1,253
Russian	7,854	656	4,256	2,432	1,253	-

Table 4: An example of 6 * 5 Language pairs covered by our MM-CLS dataset, and the number of images with the corresponding article-summary pair is 3 4. Here, we do not list them for simplicity.

CCL 2023