Multimodal Sentence Summarization via Multimodal Selective Encoding

Haoran Li[1], **Junnan Zhu**[2,3], **Jiajun Zhang**[2,3,4], **Chengqing Zong**[2,3,5], **Xiaodong He**[1]

[1] JD AI Research
[2]National Laboratory of Pattern Recognition, Institute of Automation, CAS
[3] University of Chinese Academy of Sciences
[4] Beijing Academy of Artificial Intelligence
[5] CAS Center for Excellence in Brain Science and Intelligence Technology
{lihaoran24, xiaodong.he}@jd.com
{junnan.zhu, jjzhang, cqzong}@nlpr.ia.ac.cn

Abstract

This paper studies the problem of generating a summary for a given sentence-image pair. Existing multimodal sequence-to-sequence approaches mainly focus on enhancing the decoder by visual signals, while ignoring that the image can improve the ability of the encoder to identify highlights of a news event or a document. Thus, we propose a multimodal selective gate network that considers reciprocal relationships between textual and multi-level visual features, including global image descriptor, activation grids, and object proposals, to select highlights of the event when encoding the source sentence. In addition, we introduce a modality regularization to encourage the summary to capture the highlights embedded in the image more accurately. To verify the generalization of our model, we adopt the multimodal selective gate to the text-based decoder and multimodal-based decoder. Experimental results on a public multimodal sentence summarization dataset demonstrate the advantage of our models over baselines. Further analysis suggests that our proposed multimodal selective gate network can effectively select important information in the input sentence.

1 Introduction

Text summarization is a task that condenses a long sentence to a short version. Existing researches (Rush et al., 2015; Chopra et al., 2016; Zeng et al., 2016; Li et al., 2017; Tan et al., 2017; Zhou et al., 2017; Zhang et al., 2018; Li et al., 2020b) produce summary only from the text. However, it has been proved that human understands the text relying on multimodal information (Waltz, 1980; Srihari, 1994; He and Deng, 2017), such as linguistic and visual signals. In this paper, we focus on the multimodal summarization task (Li et al., 2018a) that generates summary simultaneously drawing knowledge from coupled text and image, which can facilitate other applications such as image captioning (Xu et al., 2015; Vinyals et al., 2015), multimodal news summarization (Narayan et al., 2017; Zhu et al., 2018; Chen and Zhuge, 2018), and electronic commerce (e-commerce) product description generation (Chen et al., 2019; Elad et al., 2019; Zhang et al., 2019; Li et al., 2020a), etc.

Intuitively, it is easier for a reader to grasp the highlight of a news event by viewing an image than by reading a long text. Hence we believe that the image will benefit text summarization system. Figure 1 illustrates this phenomenon. For a given source sentence, a paired image visualizes a set of event highlight words, which highly correlates with the reference summary.

Multimodal sequence-to-sequence (seq2seq) learning has been widely explored in machine translation (MT) (Calixto et al., 2017; Caglayan et al., 2017; Helcl et al., 2018; Grönroos et al., 2018) in recent years, and the performances of their models surpass text-only models (Barrault et al., 2018). The main difference between multimodal MT and multimodal text summarization is: the model for MT is required to convert the same semantics from the input sentence and the paired image to the output, while for summarization, the model is expected to select the important information from the input. Li et al. (2018a) propose a hierarchical attention model for the multimodal sentence summarization task, while the image is not involved in the process of text encoding. Obviously, it will be easier for the decoder to generate

Proceedings of the 28th International Conference on Computational Linguistics, pages 5655–5667
Barcelona, Spain (Online), December 8-13, 2020

Source sentence: british prime minister tony blair has *rescued* a danish *swimmer* from the shark infested *sea* during his holiday in seychelles in africa , a government spokesman said friday .
Reference summary: uk pm rescues swimmer from sea .

Figure 1: The image visualizes the *highlight words* which are highly related to the summary. Our model distinguishes highlights via multimodal selective encoding.

an accurate summary if the encoder can filter out trivial information when encoding the input sentence. Based on this idea, we propose a multimodal selective mechanism which aims to select the highlights from the input text using visual signals, and then the decoder generates the summary using the filtered encoding information. Concretely, an encoder reads the input text and generates the hidden representations. Then, multimodal selective gates measure the relevance between the input words and the image to construct the selected hidden representation. Finally, a decoder generates the summary using the selected hidden representation. We explore the capacity of the image to select salient content from the input text at different levels including global image descriptor, activation grids, and object proposals in the image. Accordingly, as shown in Figure 2, we design three visual selective gates: global-level, grid-level, and object-level gates. Furthermore, some abstract concepts, such as "guilty" and "freedom", can hardly be well represented by the image, and thus we combine textual and visual selective gates to construct multimodal selective gates. In addition, we argue that a good summarizer should adequately capture the highlight words. Thus, we impose a modality regularization on our model that encourages the semantic similarity between (image, generated summary) higher than that between (image, source text).

Our main contributions are as follows:

- We propose a novel multimodal selective mechanism that can use both the textual and visual signals to select the important information from the source text.
- We propose a visual-guided modality regularization module to encourage the model focus on the key information in the source.
- The experimental results on a multimodal sentence summarization dataset demonstrate that our proposed system can take advantage of multimodal information and outperform baseline methods.

2 Related Work

2.1 Abstractive Sentence Summarization

Rush et al. (2015) first propose a seq2seq model to generate the summary for a sentence. Chopra et al. (2016) and Nallapati et al. (2016) further develop the seq2seq for this task. Gu et al. (2016), Zeng et al. (2016), and Gulcehre et al. (2016) incorporate a copying mechanism into the seq2seq. See et al. (2017) incorporate the pointer-generator model with the coverage mechanism. Chen et al. (2016) propose a distraction model to focus on the different parts of the input. Ma et al. (2017) focus on improving the semantic relevance between source and summary by encouraging high similarity of their representation. Zhou et al. (2017) employ a selective encoding mechanism to filter secondary information. Li et al. (2017) apply a deep recurrent generative decoder to the seq2seq framework. Cao et al. (2018) and Li et al. (2018b) solve the problem of fake facts in a summary using fact descriptions of the input. Zhou et al. (2018b) extend the copying mechanism from word to sequence level. Song et al. (2018) propose structure-infused copy mechanisms to copy important words and relations from the input sentence to the summary. Cohan et al. (2018) propose a discourse-aware hierarchical attention model for abstractive summarization. Duan et al. (2019) propose a contrastive attention mechanism that attends to irrelevant parts of the input. Wang et al. (2019) present a bi-directional selective encoding model with template to softly select key information from source text.

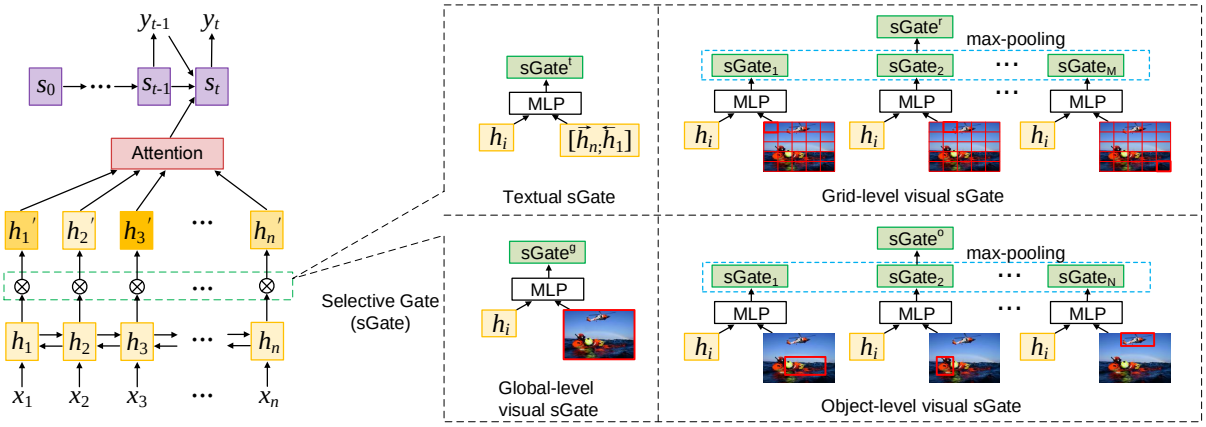


Figure 2: The framework of our model. We design visual selective gates including global-level, grid-level and object-level visual gates to select salient encoding information. We also integrate the textual and visual selective gates to construct multimodal selective gates. In this figure, the summary is generated with the text-based decoder, and we also apply multimodal selective gates to the multimodal-based decoder (Li et al., 2018a) in our work.

2.2 Multimodal Seq2seq Models

Libovický and Helcl (2017) propose multi-source seq2seq learning with hierarchical attention. Calixto and Liu (2017) use images as source words to improve translation quality. Delbrouck and Dupont (2017) adjust various attention for the visual modality. Calixto et al. (2017) propose attention mechanisms for textual and visual modalities and combine them to decode target words. Narayan et al. (2017) develop extractive summarization with side information including images and captions. Zhu et al. (2018), Chen and Zhuge (2018) and Zhu et al. (2020) propose to generate multimodal summary for multimodal news document. Li et al. (2018a) first introduce the multimodal sentence summarization task, and they propose a hierarchical attention model, which can pay different attention to image patches, words, and different modalities when decoding target words. Li et al. (2020a) propose a aspect-aware multimodal summarization model for e-commerce products. We use visual signals to enhance the encoder for the multimodal sentence summarization task, aiming to discriminate the important source information when encoding the input sentence.

3 Proposed Model

3.1 Overview

The input of the multimodal text summarization task is a pair of text and image, and the output is a textual summary. We propose visual selective gates to encourage important source information encoded into the second-level hidden sequence. We argue that the text can be pertinent to visual information at the level of the whole image, the image parts, and the object proposals in the image. Thus, as shown in Figure 2, we design three visual selective gates: global-level, grid-level, and object-level gates. Next, the summary decoder produces the summary based on the second-level hidden states. We apply two types of decoders including the text-based decoder and multimodal-based decoder.

3.2 Text Encoder

Given a source text $\mathbf{x} = (x_1, \cdots, x_n)$, an encoder builds the first-level hidden state sequence $\mathbf{h} = (h_1, \cdots, h_n)$. We apply a Bidirectional GRU to encode the source sentence forwardly and backwardly into two sequences of the hidden states: $(\overrightarrow{h}_1, \cdots, \overrightarrow{h}_n)$, $(\overleftarrow{h}_1, \cdots, \overleftarrow{h}_n)$, and $h_i = [\overrightarrow{h}_i; \overleftarrow{h}_i]$.

3.3 Image Encoder

Given an image, we apply a pre-trained VGG (VGG19) and a Faster R-CNN (Ren et al., 2017) model to extract visual features at multiple levels including the whole image, the image parts, and the object

proposals in the image. We extract the 4096-dimensional fully-connected layer of VGG19 (`fc7` layer) as the image global feature v^g. We extract a $7 \times 7 \times 512$ feature map of the last pooling layer (`pool5` layer) of VGG19 as the image grid-level feature v^r. For object proposal features v^o of an image, we use Faster R-CNN (Ren et al., 2017) initialized with ResNet-101 (He et al., 2016) pretrained for classification on ImageNet (Deng et al., 2009), and then we retrain it on Visual Genome Dataset (Krishna et al., 2017). $v_j^o \in \mathbb{R}^{2048}$ is obtained from the ROI pooling layer in the Region Proposal Network. We choose the top 36 object proposals after non-maximum suppression (Neubeck and Gool, 2006).

3.4 Textual Selective Gates

In this section, we briefly describe the selective mechanism (Zhou et al., 2017).

The selective encoding (Zhou et al., 2017) extends the seq2seq model by constructing a second-level hidden state $h_i^{'}$ to control the information flow from the encoder to the decoder as follows:

$$sGate_i^t = \sigma(\mathbf{W}_t h_i + \mathbf{U}_t r) \tag{1}$$

where σ denotes the sigmoid function, $\mathbf{W}$ and $\mathbf{U}$ are parameter matrices. r is the overall representation for the text $\mathbf{x}$.

Then, $h_i^{'}$ is computed as follows:

$$h_i^{'} = h_i \odot sGate_i^t \tag{2}$$

Finally, the decoder generates the summary depending on the second-level hidden sequence $\mathbf{h}^{'}$ as a standard seq2seq model does.

3.5 Visual Selective Gates

We design three kinds of visual selective gates to select salient encoding information: global-level, grid-level, and object-level gates. We further combine textual and visual selective gates to construct multi-modal selective gates.

Global-Level Visual Selective Gate

To conduct a global-level visual selective gate $sGate^g$, we explore the association between text hidden state h_i and global visual feature v^g:

$$sGate_i^g = \sigma(\mathbf{W}_g h_i + \mathbf{U}_g v^g) \tag{3}$$

Grid-Level Visual Selective Gate

Grid-level image features retain the spatial information of an image, which can be used to bridge the gap between text and image patch-by-patch as follows:

$$sGate_i^r = \mathbf{max}_j\{\sigma(\mathbf{W}_r h_i + \mathbf{U}_r v_j^r)\} \tag{4}$$

where $v_j^r \in \mathbb{R}^{512} (j \in [1, 49])$ is a grid-level image feature.

Specifically, for each h_i, we compute grid-level visual selective gate with each grid-level image patch v_j^r, and then we apply a max-pooling over the gates to take the maximum value of each dimension as the final selective gate for h_i. The idea behind this is that different dimensions of h_i represent different semantics, which correspond to different image patches, and the **max** operation can capture the most related image regions for each h_i.

Object-Level Visual Selective Gate

Grid-level image features correspond to a uniform grid of equally sized and shaped neural receptive fields for the images. In fact, salient image region, such as object proposal, is a much more natural basis for human attention (Egly et al., 1994). Thus, we compute the selective gate based on object-level image features as follows:

$$sGate_i^o = \mathbf{max}_j\{\sigma(\mathbf{W}_o h_i + \mathbf{U}_o v_j^o)\} \tag{5}$$

where $v_j^o \in \mathbb{R}^{2048} (j \in [1, 36])$ is an object-level image feature.

Multimodal Selective Gates

To avoid missing necessary information which can hardly be well represented by the image, such as abstract concepts, we combine textual and visual selective gates to construct multimodal selective gates. To jointly select important encoding information by linguistics and different levels of visual signals, we propose three kinds of multimodal selective gates as follows:

$$sGate_i^{tg} = \sigma(\mathbf{W}_a h_i + \mathbf{V}_a r + \mathbf{U}_a v^g) \tag{6}$$

$$sGate_i^{tr} = \mathbf{max}_j\{\sigma(\mathbf{W}_b h_i + \mathbf{V}_b r + \mathbf{U}_b v_j^r)\} \tag{7}$$

$$sGate_i^{to} = \mathbf{max}_j\{\sigma(\mathbf{W}_c h_i + \mathbf{V}_c r + \mathbf{U}_c v_j^o)\} \tag{8}$$

where $r = [\overrightarrow{h}_n; \overleftarrow{h}_1]$ is the overall representation for the source text.

Then, the second-level encoding states $h_i^{'}$ is computed by one of the $sGate \in \{sGate^g, sGate^r, sGate^o, sGate^{tg}, sGate^{tr}, sGate^{to}\}$ as follows:

$$h_i^{'} = h_i \odot sGate_i \tag{9}$$

Next, the decoder generates summaries based on $\mathbf{h}^{'}$.

3.6 Summary Decoder

To verify the generalization of our model, we apply multimodal selective gates to both the text-based decoder and multimodal-based decoder.

Text-Based Summary Decoder

The decoder for a standard text-only seq2seq model calculates the decoding state s_t as follows:

$$s_t = \mathrm{GRU}(s_{t-1}, y_{t-1}, c_t) \tag{10}$$

where $s_0 = \tanh(\mathbf{W}_h[\overrightarrow{h_n^{'}}; \overleftarrow{h_1^{'}}])$.

Context vector c_t is computed as a weighted sum of the source annotations as follows:

$$c_t = \sum_{i=1}^{N} \alpha_{t,i}^{txt} h_i^{'} \tag{11}$$

$$\alpha_{t,i}^{txt} = \mathrm{softmax}(\tanh(\mathbf{U}_d s_{t-1} + \mathbf{W}_d h_i^{'})) \tag{12}$$

The probability for the next target word y_t is computed using the decoder state s_t and the previous emitted word y_{t-1} as follows:

$$p(y_t) \propto \exp(\mathbf{W}_l \tanh(\mathbf{W}_s s_t + \mathbf{W}_y E[y_{t-1}]) \tag{13}$$

The loss function $\mathcal{L}_t$ for each time t is as follows:

$$\mathcal{L}_t = -\log p(y_t) \tag{14}$$

Multimodal-Based Summary Decoder

Li et al. (2018a) propose a multimodal-based summarization method. In this work, we adopt the summary decoder with weighted image local features initialization, and the hierarchical attention mechanism which can pay different attention to the input words and image patches, producing textual and visual context vectors. The textual context vector c_{txt} is calculated with Equation 11. The image context vector c_t^{img} is as follows:

$$c_t^{img} = \sum_{i=1}^{L} \alpha_{t,i}^{img} v_i^r \tag{15}$$

$$\alpha_{t,i}^{img} = \mathrm{softmax}(\tanh(\mathbf{U}_e s_{t-1} + \mathbf{W}_e v_i^r)) \tag{16}$$

The second-level modality attention is:

$$c_t = \beta_t^{txt} \mathbf{V}_T c_t^{txt} + \beta_t^{img} \mathbf{V}_I c_t^{img} \tag{17}$$

$$\beta_t^{txt} = \sigma(\mathbf{U}_f s_{t-1} + \mathbf{W}_f c_t^{txt}) \tag{18}$$

$$\beta_t^{img} = \sigma(\mathbf{U}_k s_{t-1} + \mathbf{W}_k c_t^{img}) \tag{19}$$

where β_t^{txt} is attention weight for textual context and β_t^{img} is attention weight for visual. The probability for the next target word y_t is computed based on Equation 10 and 13.

3.7 Modality Regularization

Given a source text, we argue that the paired image, especially for the foreground of an image, contains more centralized information that should be presented in the summary. To guarantee the summarizer can better capture the highlights expressed by the image, we attempt to constrain the image more related to the generated summary than to the source. To achieve this goal, we design a Modality Regularization (MR) module that scores (image, target) higher than (image, source).

We first learn representations for the source and summary by a cross-modal attention, which emphasize the words related to the image:

$$c^{src} = \sum_{i=1}^{N} \alpha_i^{src} h_i^{'} \tag{20}$$

$$\alpha_i^{src} = \mathrm{softmax}(\tanh(\mathbf{U}_v v^g + \mathbf{W}_h h_i^{'})) \tag{21}$$

$$c^{tgt} = \sum_{t=1}^{T} \alpha_t^{tgt} s_t \tag{22}$$

$$\alpha_t^{tgt} = \mathrm{softmax}(\tanh(\mathbf{U}_v v^g + \mathbf{W}_s s_t)) \tag{23}$$

Then we project the source, the target, and the image into the shared space: $e^{src} = \tanh(\mathbf{W}_h c^{src})$, $e^{tgt} = \tanh(\mathbf{W}_s c^{tgt})$, $e^v = \tanh(\mathbf{U}_v v^g)$. A pairwise ranking loss function to encourage similarity score between (image, summary) higher than (image, source):

$$\mathcal{L}^r = \max\{0, \delta - \cos(e^{tgt}, e^v) + \cos(e^{src}, e^v)\} \tag{24}$$

where $\delta = 2$, and $\cos$ is a cosine similarity function. The final loss function is:

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}_t + \mathcal{L}^r \tag{25}$$

In addition, inspired by Zhou et al. (2018a), we calculate s_0 with e^{src} as follows:

$$s_0 = \tanh(\mathbf{W}_h([\overrightarrow{h_n^{'}}; \overleftarrow{h_1^{'}}] + e^{src})) \tag{26}$$

4 Experiments

4.1 Dataset

We evaluated our methods on the public multimodal sentence summarization dataset (Li et al., 2018a). Each sample in this dataset is a triplet (sentence, image, summary). The dataset consists of 62,000 training samples, 2,000 test samples and 2,000 validation samples.

4.2 Experimental Settings

We set word embedding size to 300 and GRU hidden state size to 512. We use the full source and target vocabularies collected from the training data, which have 36,916 source words and 26,168 target words, respectively. The mini-batch size is 64, and beam search size is 10. Adam optimizer is applied with the learning rate of 0.0005, momentum parameters $\beta_1 = 0.9$ and $\beta_1 = 0.999$, and $\epsilon = 10^{-8}$. We use dropout (Srivastava et al., 2014) with probability of 0.2 and gradient clipping (Pascanu et al., 2013) with range $[-1, 1]$. During training, we test the ROUGE-2 (Lin, 2004) F1-score on the validation set for every 5,000 batches, and we halve the learning rate if the score drops for 5 consecutive testings.

Methods	RG-1	RG-2	RG-L
Non-Selective	44.53 (± 0.11)	22.67 (± 0.12)	41.91 (± 0.09)
T-Selective	44.81 (± 0.14)	23.13 (± 0.13)	41.88 (± 0.11)
Global V-Selective	45.31 (± 0.15)	23.39 (± 0.13)	42.48 (± 0.11)
Object V-Selective	44.83 (± 0.12)	23.28 (± 0.13)	42.03 (± 0.12)
Grid V-Selective	45.11 (± 0.16)	23.33 (± 0.15)	42.21 (± 0.12)
T + Global V-Selective	45.51 (± 0.13)	23.47 (± 0.17)	42.78 (± 0.10)
T + Object V-Selective	45.33 (± 0.14)	23.41 (± 0.15)	42.51 (± 0.12)
T + Grid V-Selective	45.58 (± 0.19)	23.58 (± 0.16)	42.80 (± 0.10)
T + Grid V-Selective + MR	**45.63** (± 0.12)	**23.68** (± 0.13)	**42.97** (± 0.09)

Table 1: Experimental results of our models with the text-based decoder. "T-Selective", "V-Selective" and "MR" denote Textual Selective, Visual Selective, and Modality Regularization, respectively. Our "T + Grid V-Selective + MR" model performs significantly better than the baselines by the 95% confidence interval in the official ROUGE script. Best result among different "V-Selective" is underlined.

4.3 Comparative Methods

Lead. A baseline uses the first eight words as the summary according to the reference length.

Compress. Clarke (2008) propose to compress sentence based on syntactic structure.

ABS. Rush et al. (2015) use an attentive CNN encoder and a neural network language model decoder to summarize the sentence.

SEASS. Zhou et al. (2017) propose a summarization model with textual selective encoding.

Multi-Source. Libovický and Helcl (2017) propose a hierarchical attention model.

Doubly-Attentive. Calixto et al. (2017) propose a doubly-attentive mechanism for multimodal machine translation.

Seq2seq. Bahdanau et al. (2015) propose a standard seq2seq model for machine translation.

Pointer-Generator Network (PGNet). See et al. (2017) present a text-based seq2seq network model with the copying mechanism.

MAtt. Li et al. (2018a) propose a hierarchical **M**odality-**Att**ention model, which is the state-of-the-art method for the multimodal sentence summarization.

4.4 Experimental Results

We report ROUGE-1 (RG-1), ROUGE-2 (RG-2), and ROUGE-L (RG-L) F1-scores[1]. We conduct the experiments with different initializations for eight times, and we report the mean and standard deviation. Table 1 shows the results for text-based decoders with different selective encoding mechanisms and modality regularization (MR). Generally, our models with visual selective gates perform better than the model with textual selective gate, and global-level feature constructs the most effective visual selective gate, while the object-level feature performs worst. We argue that some background regions, such as "sea" in Figure 1 and "highway" in Figure 3, are also necessary for our task, while object proposals may ignore these background. Recognition errors caused by Region Proposal Network may also influence the summarization performance. Multimodal Selective gates lead to further improvements, and the model with "Textual + Grid-level Visual Selective" achieves the highest ROUGE score. We hypothesize the reason may be that similar to the global-level visual selective gate, the textual gate is also a global-level selective gate. Thus, compared with the global-level visual selective gate, the local-level grid-level visual gate contribute more to the textual gate. In addition, we test with combining all of three different levels of visual information and textual information to compute the selective gate, while the result is a little worse than that of "Textual + Grid-level Visual Selective".

Table 2 show comparisons with other work. We can conclude that Multimodal Selective (Textual + Grid-Level Visual Selective, TGS) and Modality Regularization (MR) strategies exhibit good generalization ability for both the text-based decoder (Seq2seq) and multimodal-based decoder (MAtt), and the model "MAtt + PGNet + TGSMR" outperforms other baselines.

[1]The ROUGE evaluation option is -m -n 2 -w 1.2

Methods	RG-1	RG-2	RG-L
Lead*	33.64	13.40	31.84
Compress*	31.56	11.02	28.87
ABS*	35.95	18.21	31.89
SEASS*	44.86	23.03	41.92
Multi-Source*	39.67	19.11	38.03
Doubly-Attentive*	41.11	21.75	39.92
Seq2seq*	44.58	22.68	41.91
PGNet	46.05 ($\pm$ 0.14)	24.18 ($\pm$ 0.16)	44.16 ($\pm$ 0.14)
MAtt*	45.78	23.45	43.16
MAtt+Coverage*	47.28	24.85	44.48
Seq2seq + TGSMR	45.63 ($\pm$ 0.12)	23.68 ($\pm$ 0.13)	42.97 ($\pm$ 0.09)
PGNet + TGSMR	47.79 ($\pm$ 0.15)	25.36 ($\pm$ 0.13)	44.87 ($\pm$ 0.12)
MAtt + TGSMR	46.32 ($\pm$ 0.13)	24.25 ($\pm$ 0.14)	43.50 ($\pm$ 0.12)
MAtt + PGNet	47.17 ($\pm$ 0.14)	24.78 ($\pm$ 0.13)	44.85 ($\pm$ 0.11)
MAtt + PGNet + TGSMR	**48.19** ($\pm$ 0.14)	**25.64** ($\pm$ 0.12)	**45.27** ($\pm$ 0.13)

Table 2: Comparisons with other work. "TGSMR" denotes "Textual + Grid-Level Visual Selective + Modality Regularization". The model of "MAtt + PGNet + TGSMR" performs significantly better than other baselines by the 95% confidence interval in the official ROUGE script. "*" marks the results from Li et al. (2018a).

Methods	Precision	Recall	F1-score
TextRank	27.36	58.09	35.52
Seq2seq + T-Selective	33.83	63.90	42.83
Seq2seq + TGSMR	35.72	65.11	44.69
MAtt + TGSMR	36.24	66.52	45.29
BiLSTM-CRF	51.70	72.62	56.60

Table 3: Results (%, averaged precision, recall and F1-score) for keyword extraction.

4.5 How Reliable is Our Selective Gates?

In this section, we explore whether the selective gates can correctly discriminate the important words from the input. To this end, we evaluate how many activated words by the selective gates are ground-truth keywords (we take the overlapping words, except for stop-words, between the input sentence and the reference summary as the ground-truth keywords). Following Zhou et al. (2017), we use the method of Li et al. (2016) to calculate the contribution of the selective gates to the final summary. Considering that the average word count of the summaries in the validation set is eight, we take the words with top-8 contribution values as the activated keywords by the selective mechanisms.

Table 3 shows the results for keyword extraction. Our models with selective encoding perform better than unsupervised TextRank algorithm (Mihalcea and Tarau, 2004) (we also take the top-8 scored words as the keywords), and Multimodal Selective show advantages over Textual Selective. We further train a BiLSTM-CRF model (Huang et al., 2015) using sentence-keyword samples of the multimodal sentence summarization dataset, and the keyword extraction result for BiLSTM-CRF is better than our model, indicating a promising prospect for further development of the selective mechanism. In the future, we will dedicate our efforts to explore whether the selective gate can benefit from supervision signals of a special keyword extractor. A feasible research direction may be to explore whether the selective gate can benefit from supervision signals of a special keyword extractor.

4.6 Human Evaluation

We perform human evaluation with "Seq2seq + TGSMR" to avoid the influence of other modules such as multimodal decoder (MAtt) and copying (PGNet). We randomly choose 100 samples from the test set, and three postgraduates are involved to annotate whether "Seq2seq + TGSMR" outperforms "Seq2seq" model with respect to the informativeness and readability. The results are shown in Table 4. For example, "win" denotes that the results of "Seq2seq + TGSMR" is better than the "Seq2seq" model. We can see

Informativeness				Readability			
win	tie	lose	kappa	win	tie	lose	kappa
45.33	41.33	13.33	0.413	45.00	41.33	13.67	0.425

Table 4: Human evaluation results (%). "Win" denotes that the generated summaries of "Seq2seq + TGSMR" is better than "Seq2seq" model.

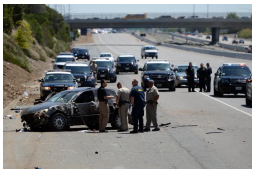
that the percentage of "win" is larger than "lose" (45.33% vs. 13.33% for informativeness, and 45% vs. 13.67% for readability. Kappa (Fleiss, 1971) is used to verified the consistency for different annotators.), demonstrating that the improvements of our multimodal summarizer are significant (p-value < 0.01, paired t-test).

4.7 Case Study

We show a sample from the test set, with comparisons of the reference summary and the summaries generated by the Seq2seq model, the text-based decoder with "Textual Selective", and "Textual + Grid-level Visual Selective". In Figure 3, the image shows the *"highway"* where the event occurs, and "Textual + Grid-level visual selective" model successfully spells out the place while other models fail.

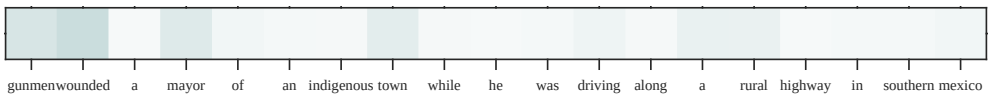
We visualize the gate values in a salience heat map for each source word. We observe that our multimodal selective gates accurately determine the salience of the source words, while textual selective gates fail, especially for the word *"highway"*. We also show visual selective gate values for "Textual + Grid-level Visual Selective" model in Figure 3 (b) for the target word "highway", which demonstrates that our visual grounded selective mechanism successfully captures the relationship between the image and the word that has corresponding visual semantics.

Source Sentence: gunmen wounded a mayor of an indigenous town while he was driving along a rural *highway* in southern mexico, authorities said tuesday .
Reference Summary: gunmen wound mexican mayor in attack on *highway*
Seq2seq Model: gunmen kill # people in southern mexico
Textual Selective Model: gunmen shoot mayor of town
Textual + Grid-Level Visual Selective Model: mayor of indigenous town is shot on rural *highway*

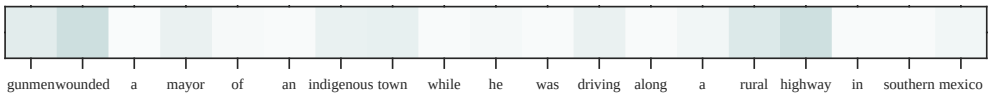


(a) Summaries generated by different models.

(b) Selective gate for *highway* (lighter denotes larger value).



(c) Salience heat map for "Textual selective gate" (deeper color denotes larger selective value).



(d) Salience heat map for "Textual + Grid-level visual selective gate".

Figure 3: A case study. "Textual + Grid-level Visual Selective" model accurately selects the important word *"highway"* while other models fail (see c and d). Visual selective gate values clearly show the reason why *"highway"* is selected (see b). Note that we simplify the input sentence to show the heat map.

5 Conclusion

This paper addresses the multimodal text summarization task, namely, transforming a pair of text and image into a summary. We propose multimodal selective encoding gates that can effectively control

the information flow from the encoder to the decoder. We adopt our multimodal selective encoding strategies to the text-based and multimodal-based decoders. The experimental results on the public multimodal sentence summarization corpus prove that our proposed model significantly outperforms other comparative methods. Note that our method is not specifically designed for RNN-based models, and it can be easily applied to Transformer-based models, which is left for our future work.

6 Acknowledgments

This work is partially supported by National Key R&D Program of China (2020AAA0105200), National Key R&D Program of China (2018YFB2100802), and Beijing Academy of Artificial Intelligence (BAAI).

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels, October. Association for Computational Linguistics.

Ozan Caglayan, Walid Aransa, Adrien Bardet, Mercedes García-Martínez, Fethi Bougares, Loïc Barrault, Marc Masana, Luis Herranz, and Joost van de Weijer. 2017. LIUM-CVC submissions for WMT17 multimodal translation task. In *Proceedings of the Second Conference on Machine Translation*, pages 432–439, Copenhagen, Denmark, September. Association for Computational Linguistics.

Iacer Calixto and Qun Liu. 2017. Incorporating global visual features into attention-based neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 992–1003, Copenhagen, Denmark, September. Association for Computational Linguistics.

Iacer Calixto, Qun Liu, and Nick Campbell. 2017. Doubly-attentive decoder for multi-modal neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1913–1924, Vancouver, Canada, July. Association for Computational Linguistics.

Ziqiang Cao, Furu Wei, Wenjie Li, and Sujian Li. 2018. Faithful to the original: Fact aware neural abstractive summarization. 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 4784–4791.

Jingqiang Chen and Hai Zhuge. 2018. Abstractive text-image summarization using multi-modal attentional hierarchical RNN. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4046–4056, Brussels, Belgium, October-November. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. 2016. Distraction-based neural networks for modeling documents. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2754–2760.

Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards knowledge-based personalized product description generation in e-commerce. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 3040–3050.

Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 93–98, San Diego, California, June. Association for Computational Linguistics.

James Clarke. 2008. *Global inference for sentence compression : an integer linear programming approach*. Ph.D. thesis, University of Edinburgh, UK.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana, June. Association for Computational Linguistics.

Jean-Benoit Delbrouck and Stéphane Dupont. 2017. An empirical study on the effectiveness of images in multimodal neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 910–919, Copenhagen, Denmark, September. Association for Computational Linguistics.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255.

Xiangyu Duan, Hongfei Yu, Mingming Yin, Min Zhang, Weihua Luo, and Yue Zhang. 2019. Contrastive attention mechanism for abstractive sentence summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3044–3053, Hong Kong, China, November. Association for Computational Linguistics.

R Egly, J Driver, and R. D. Rafal. 1994. Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. *Journal of Experimental Psychology General*, 123(2):161–77.

Guy Elad, Ido Guy, Slava Novgorodov, Benny Kimelfeld, and Kira Radinsky. 2019. Learning to generate personalized product descriptions. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, pages 389–398.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Stig-Arne Grönroos, Benoit Huet, Mikko Kurimo, Jorma Laaksonen, Bernard Merialdo, Phu Pham, Mats Sjöberg, Umut Sulubacak, Jörg Tiedemann, Raphael Troncy, and Raúl Vázquez. 2018. The MeMAD submission to the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 603–611, Belgium, Brussels, October. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany, August. Association for Computational Linguistics.

Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 140–149, Berlin, Germany, August. Association for Computational Linguistics.

Xiaodong He and Li Deng. 2017. Deep learning for image-to-text generation: A technical overview. *IEEE Signal Processing Magazine*, 34(6):109–116.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778.

Jindřich Helcl, Jindřich Libovický, and Dušan Variš. 2018. CUNI system for the WMT18 multimodal translation task. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 616–623, Belgium, Brussels, October. Association for Computational Linguistics.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.

Jiwei Li, Xinlei Chen, Eduard Hovy, and Dan Jurafsky. 2016. Visualizing and understanding neural models in NLP. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 681–691, San Diego, California, June. Association for Computational Linguistics.

Piji Li, Wai Lam, Lidong Bing, and Zihao Wang. 2017. Deep recurrent generative decoder for abstractive text summarization. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2091–2100, Copenhagen, Denmark, September. Association for Computational Linguistics.

Haoran Li, Junnan Zhu, Tianshang Liu, Jiajun Zhang, and Chengqing Zong. 2018a. Multi-modal sentence summarization with modality attention and image filtering. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 4152–4158.

Haoran Li, Junnan Zhu, Jiajun Zhang, and Chengqing Zong. 2018b. Ensure the correctness of the summary: Incorporate entailment knowledge into abstractive sentence summarization. In *Proceedings of COLING*, pages 1430–1441.

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*, pages 8188–8195.

Haoran Li, Junnan Zhu, Jiajun Zhang, Chengqing Zong, and Xiaodong He. 2020b. Keywords-guided abstractive sentence summarization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8196–8203.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada, July. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*.

Shuming Ma, Xu Sun, Jingjing Xu, Houfeng Wang, Wenjie Li, and Qi Su. 2017. Improving semantic relevance for sequence-to-sequence learning of Chinese social media text summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 635–640, Vancouver, Canada, July. Association for Computational Linguistics.

Rada Mihalcea and Paul Tarau. 2004. TextRank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, Barcelona, Spain, July. Association for Computational Linguistics.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Çağlar Gùlçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 280–290, Berlin, Germany, August. Association for Computational Linguistics.

Shashi Narayan, Nikos Papasarantopoulos, Mirella Lapata, and Shay B. Cohen. 2017. Neural extractive summarization with side information. *CoRR*, abs/1704.04530.

Alexander Neubeck and Luc Van Gool. 2006. Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China*, pages 850–855.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. 2013. On the difficulty of training recurrent neural networks. In *International conference on machine learning (ICML)*, pages 1310–1318.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2017. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(6):1137–1149.

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, September. 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, July. Association for Computational Linguistics.

Kaiqiang Song, Lin Zhao, and Fei Liu. 2018. Structure-infused copy mechanisms for abstractive summarization. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1717–1729, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Rohini K. Srihari. 1994. Computational models for integrating linguistic and visual information: A survey. *Artif. Intell. Rev.*, 8(5-6):349–369.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research (JMLR)*.

Jiwei Tan, Xiaojun Wan, and Jianguo Xiao. 2017. Abstractive document summarization with a graph-based attentional neural model. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1171–1181, Vancouver, Canada, July. Association for Computational Linguistics.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 3156–3164.

David L Waltz. 1980. Generating and understanding scene descriptions. *Elements of Discourse Understanding*, pages 266–282.

Kai Wang, Xiaojun Quan, and Rui Wang. 2019. BiSET: Bi-directional selective encoding with template for abstractive summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2153–2162, Florence, Italy, July. Association for Computational Linguistics.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 2048–2057.

Wenyuan Zeng, Wenjie Luo, Sanja Fidler, and Raquel Urtasun. 2016. Efficient summarization with read-again and copy mechanism. *CoRR*, abs/1611.03382.

Jiajun Zhang, Yang Zhao, Haoran Li, and Chengqing Zong. 2018. Attention with sparsity regularization for neural machine translation and summarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(3):507–518.

Tao Zhang, Jin Zhang, Chengfu Huo, and Weijun Ren. 2019. Automatic generation of pattern-controlled product description in e-commerce. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2355–2365.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2017. Selective encoding for abstractive sentence summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1095–1104, Vancouver, Canada, July. Association for Computational Linguistics.

Mingyang Zhou, Runxiang Cheng, Yong Jae Lee, and Zhou Yu. 2018a. A visual attention grounding neural model for multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3643–3653, Brussels, Belgium, October-November. Association for Computational Linguistics.

Qingyu Zhou, Nan Yang, Furu Wei, and Ming Zhou. 2018b. Sequential copying networks. 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 4987–4995.

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, October-November. Association for Computational Linguistics.

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