A Multi-view Meta-learning Approach for Multi-modal Response Generation

Zhiliang Tian
National University of Defense Technology
Changsha, China
tianzhiliang@nudt.edu.cn

Zheng Xie
National University of Defense Technology
Changsha, China
xiezheng81@nudt.edu.cn

Fuqiang Lin
National University of Defense Technology
Changsha, China
linfuqiang13@alumni.nudt.edu.cn

Yiping Song*
National University of Defense Technology
Changsha, China
songyiping@nudt.edu.cn

ABSTRACT

As massive conversation examples are easily accessible on the Internet, we are now able to organize large-scale conversation corpora to build chatbots in a data-driven manner. Multi-modal social chatbots produce conversational utterances according to both textual utterances and vision signals. Due to the difficulty of bridging different modalities, the dialogue generation model of chatbots falls into local minima that only capture the mapping between textual input and textual output, as a result, it almost ignores the non-textual signals. Further, similar to the dialogue model with plain text as input and output, the generated responses from multi-modal dialogue also lack diversity and informativeness. In this paper, to address the above issues, we propose a Multi-View Meta-Learning (MultiVML) algorithm that groups samples in multiple views and customizes generation models to different groups. We employ a multi-view clustering to group the training samples so as to attend more to the unique information in non-textual modality. Tailoring different sets of model parameters for each group boosts the genereation diversity via meta-learning. We evaluate MultiVML on two variants of the OpenViDial benchmark datasets. The experiments show that our model not only better explore the information from multiple modalities, but also excels baselines in both quality and diversity.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**.

KEYWORDS

Multi-modal, Response generation, Meta-learning

ACM Reference Format:
Zhiliang Tian, Zheng Xie, Fuqiang Lin, and Yiping Song. 2023. A Multi-view Meta-learning Approach for Multi-modal Response Generation. In *Proceedings of the ACM Web Conference 2023 (WWW '23), April 30–May 04, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3543507.3583548

*Corresponding Author

WWW '23, April 30–May 04, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583548

1 INTRODUCTION

Along with the maturity of Web 2.0, there has been an explosion in the number of people having conversations on websites with chatbots on social media (e.g., Facebook, Twitter). Most chatbots lead a text-to-text conversation between chatbots and users [52]. However, people prefer face-to-face communication due to the accessibility of the speaker's visual signals, e.g., facial expressions. Recently, building multi-modal chatbots is becoming a hot research topic, where the multi-modal generation model behind the chatbots [37, 65] can comprehend both textual and non-textual signals.

Neural generation models are widely-used in text-to-text response generation, but in multi-modal response generation, they suffer from some issues when additionally handling the corresponding non-textual signals. First, researchers find that most multi-modal text generation models rely highly on the textual input and attend less to the non-textual signals (i.e. vision signals) [1], which is called the "modality-bias" problem [16, 32, 45, 53, 66]. The reason is the models find the input-output mapping between the same modalities (i.e. text-to-text) is easier to capture; thus, the model falls into the local minima that are mainly based on the text-to-text mapping. Some researchers address this issue by enhancing the multi-modal information fusion using attention mechanism [5, 60, 66]. But those works are good at incorporating non-textual information which is related to the textual input, but still tend to ignore the non-overlapped information between different modalities. Second, multi-modal response generation also inherits the disadvantages of the plain text-to-text generation model: the generated outputs lack diversity and informativeness [27]. The reason is that the models tend to capture only the most salient input-output mapping since it is trained via maximizing likelihood estimation (MLE) [22, 27, 72]. Hence, some researchers enhance the generation diversity by refining the training objective [25, 56], incorporating randomness in model training [75], and importing additional information [18, 48, 73]. These methods need either a quality-diversity trade-off or additional resources.

Facing the above two issues, we argue that (1) multi-modal text generation should cater more to the non-overlapped information

[1] According to our experiment (see details in Sec.5.7), for a Transformer-based generation model that feeds the image feature as one time step in the encoder, 97.61% attention of the decoder locates on textual input and 2.39% locates on visual input.

between the non-textual modalities and textual modalities; (2) existing models are expected to achieve diverse response generation via a universal model, but customizing generation models for different samples should be a better way, where each model learns to respond with its specific data. That is a new perspective to encourage generation diversity.

In this paper, to cater to non-textual signals and enhance the diversity in multi-modal response generations, we propose a Multi-View Meta-Learning (MultiVML) algorithm that customizes generation models for different groups of training samples. To group the multi-modal samples in a suitable way, we cluster samples in the perspective of multiple views, where each view corresponds to the unique information of each modality. By doing so, samples similar in all modalities (considering non-overlapped information) are likely to be assigned in a same mate-task and be customized with one model via mate learning.

Particularly, our model consists of three modules: (1) modal-specific encoders to represent the input data in each modality, (2) a disentanglement module to extract the unique information from image representations, and (3) a decoder to generate responses. To train the model, we first divide all samples into multiple groups by multi-view clustering, where the clustering considers the information in all modalities and each group has a few examples. Then, we regard each group as a task, and adopt a meta-learning algorithm, Reptile [40], to learn to adapt the generation model to each group. The experiments on two datasets show that our method surpasses strong baselines on multi-modal response generation tasks and enhances the generation diversity. Our contributions are threefold:

- We propose a multi-view meta-learning method catering to the multi-modal information, which conducts the meta task division according to multi-modal common information and unique information for each modality.
- We are the first one to enhance the output diversity with a meta-learning paradigm in multi-modal response generation.
- Our model achieves the state-of-the-art performances in image-grounded response generation, and it encourages the model to attend more to the non-textual input signals.

2 RELATED WORK

2.1 Multi-modal Text Generation

Multi-modal text generation has many practical applications, include image-grounded response generation [37, 62, 65], multi-modal summarization [43, 72], video caption generation [33], and multi-modal machine translation [5, 71].

Most work aims at fusing the multi-modal input into a same semantic space. Bahdanau et al. [3] employ a seq2seq [60] model with hierarchical attention to integrate information from different modalities. Yao and Wan [69] propose a Transformer model augmented by an additional memory. Pasi et al. [45] use both self-attention and cross-attention for multi-modal feature fusion. [5] estimate a joint distribution over texts and images to map the image objects and words into the same space. Lei et al. [23] introduce the idea of the memory cell from the Gated Recurrent Unit (GRU) to the Transformer to fuse textual and visual signals. To learn the mutual information (MI) [20] between visual contexts and text features, Wang et al. [66] incorporate the backward probability of generating features of visual contexts given text utterances. These methods are skilled in multi-modal feature alignment and representation fusion, and thus the model can make better use of multi-modal overlapping information. However, they are not good at alleviating the neglect of visual information.

Some work notice that the multi-modal text generation model tends to only optimize the text summary generation process, while the visual input is ignored during training, which is called the "modality-bias" problem [66, 77]. Meanwhile, it is observed that the unique and non-overlapped information from different modalities is important for the output generation [24, 71, 77]. Yin et al. [71] claim that visual context helps to resolve ambiguous multi-sense words. They build a heterogeneous graph to align the objects detected from images and the words in the sentence. Li et al. [24] prove that object detecting provides additional contextual information to raw text. Li et al. [28] represent each image or sentence as a graph, where each node represents an event or entity and each edge represents an argument role. The node and edge embeddings are represented in a multimedia-shared semantic space. These graph-based methods can explicitly capture unique and non-overlapped information from different modalities. However, they usually require image object detection results, which require additional effort and may cause the cascading error.

2.2 Diverse Response Generation

Neural network models can achieve a good performance on response generation tasks, but sometimes their generated response lack diversity and informativeness [25, 27]. The methods to address this issue can be divided into three categories.

Firstly, some researchers believe that the deep learning models trained via maximal likelihood estimation (MLE) tend to only pick the most possible and safe words instead of exploring the rare words. Li et al. [25] propose another training objective maximum mutual information (MMI) to balance MLE. Huang et al. [15] uses multi-head attention to selectively align salient visual objects with textual phrases, and further encourage diversity over multiple attention heads using a margin-based diversity loss. Song et al. [56] employ determinantal point processes (DPPs) [21] to select words with both good quality and diversity in response generation. Wang and Chan [64] apply the same idea to image captioning. Some researchers treat a soft target as the training label (ground-truth) [6, 67] so that the model accepts the various generated results during training.

The second groups of methods introduce randomness in the processing of model training and inference. Li et al. [29] observed that top-k sampling [9] helps to generate informative results rather than beam search. Conditional variational autoencoder (CVAE) [17, 54] model the input sentence with a latent variable and the latent variable follows Gaussian distribution. Zhao et al. [75] apply conditional variational autoencoder CVAE to response generation. Following this paper, Du et al. [7] extends the CVAE with a single latent variable to a sequence of latent variables to model a sequence of words in the input sentence. Gao et al. [11] equip the CVAE with interpretable latent variables. Lin et al. [31] incorporate the idea of CVAE with Transformer model [63].

The third branch imports additional knowledge or background information to response generation models. [48] and [50] incorporate the background document to assist response generations. Jung et al. [18] propose a bi-directional graph exploration model

to retrieve the facts from OpenDialKG, a commonsense knowledge graph [38], for response generation models. Zhang et al. [73] propose to explore concept-level conversation flows from a commonsense graph about the concepts, ConceptNet [57]. Majumder et al. [36] have built a persona-based conversational model by using a commonsense knowledge graph, COMET [4], to expand the persona sentences. Different from the above three types of methods, this paper boosts diversity by customizing models for different samples under the meta-learning framework.

2.3 Meta-learning for Text Generation

Meta-learning is widely used in few-shot tasks [35, 47, 55, 76] due to its strong ability of fast adaptation to a new task [59]. The meta-learning methods can be roughly divided into three categories: metric-based [8, 41, 70], model-based [42, 51], and optimization-based methods [10, 49, 68]. The first two methods are generally designed for classification tasks. The optimization-based methods are model-agnostic, so they can be widely used in text generation tasks. They learn a set of initial parameters, and then differentiate the initial parameters into multiple sets with respect to different tasks. The most widely used optimization-based algorithm is the Model-agnostic Meta-learning (MAML) [10]. In each iteration step, there will be an initial parameter for K tasks that carry out gradient updates using a support set to obtain new parameters, and then use the corresponding query set to update the global initial parameters. While FOMAML [10] algorithm is a variant of MAML. When calculating the gradient of K tasks, it is no longer divided into K different parameters but uses the gradient calculated by the previous task when calculating the gradient of the next task. Reptile [40] algorithm is also a first-order algorithm, but it does not require support and query set division so the training is much simpler than other methods. This is the base algorithm we used in this paper.

The optimization-based methods have been deeply explored recently in many text generation tasks, including personalized dialogue systems [26, 35, 61, 74], minority languages machine translation [34], multi-domain visual question answering (VQA) [2, 12], and multi-topic document summarization [1]. All the above works employ meta-learning methods to address the lack of training data.

3 GENERATION MODEL

3.1 Task Definition

In the multi-modal response generation, the input is a sequence of dialogue turns containing both textual utterances and visual signals, the output is a textual response. Each training sample is denoted as $\langle q_{t1}, q_{v1}, q_{t2}, q_{v2}, \cdots, q_{tn}, q_{vn}, r\rangle$. Here, each dialog turns q_{ti} and the corresponding image q_{vi} are paired, and r stands for the textual response. Notice that the original dataset is not in the meta-learning setting, so we do not divide the corpus into several groups so far.

3.2 Model Architecture

Our generation model is based on Transformer [63], which can take advantage of large pre-train models. As shown in Fig. 1, our model consists of three modules:

- **Multi-modal Encoding Module** represents the multi-modal query.
- **Disentanglement Module** employs a disentanglement strategy to align the overlapped information in all modalities and emphasize the unique information of each modality.
- **Textual Decoding Module** takes the representation in both the multi-modal encoding module and disentanglement module as inputs and generates the corresponding responses.

3.3 Multi-modal Encoding Module

This module mainly follows the structure of the Transformer's encoder [63]. For the textual input, we employ the prevailing Transformer to model the sentence. For the visual input, we represent each image as a vector via a pre-trained Resnet-50 [14]. Since there is a sequence of utterances and their corresponding images in the input, we feed the texts and images into the module alternatively, $\langle q_{t1}, q_{v1}, q_{t2}, q_{v2}, \cdots, q_{tn}, q_{vn}\rangle$. Either a word or an image takes one-time step in the Transformer. We apply a linear layer to the image representation, and the layer transfers the image representation into a vector with the same dimension as the word embedding.

3.4 Disentanglement Module

This module takes the input representation from the encoding module and disentangles each visual representation into two parts: the overlapping part with the corresponding textual input, and the unique part that does not exist in the textual input. The disentangled visual representation replaces the original representation in Transformer's encoder and participates in the self- and cross-attention in the whole model. The motivation for disentangling visual representation is that we observe that textual data usually dominates the generation model rather than non-textual modalities. Data from non-textual modalities can only play a subsidiary role, and the supplementary information is precious. Hence, we retain all the textual representation while explicitly extracting and highlighting the supplementary information in the visual representation.

Specifically, following the sentence classification tasks [58], the first token's hidden vector in the last layer of the Transformer serves as the textual representation $\boldsymbol{z_t}$. The image's hidden vector in the last layer of Transformer serves as the visual representation $\boldsymbol{z_v}$. We aim to disentangle the image representation z_v into two vectors: $z_{overlap}$ representing the information overlapped with text and z_{unique} carrying the supplementary image information not accessible in text. We first feed z_v into a fully-connected layer and a non-linear activation function ReLU, then we split the vector into two vectors, $z_{overlap}$ and z_{unique}. Since $z_{overlap}$ carries the information overlapping with textual input, we encourage $z_{overlap}$ and z_t to be similar by minimizing square loss as follows,

$$\mathcal{L}_{squ} = \|z_t - z_{overlap}\|_2^2. \tag{1}$$

Meanwhile, we remain supplementary information from visual modality into z_{unique} by pushing z_{unique} away from $z_{overlap}$. We minimize the cosine similarity between these two vectors as Eq. 2,

$$\mathcal{L}_{cos} = \frac{z_{overlap} \cdot z_{unique}}{\|z_{overlap}\|_2 \|z_{unique}\|_2}. \tag{2}$$

Thus, we obtain the unique information of visual input z_{unique} in each dialogue turn. z_{unique} replaces the image's last layer's hidden state $\boldsymbol{z_v}$ to take part in the self-attention in the multi-modal encoding module, as well as the cross-attention in the textual decoding module.

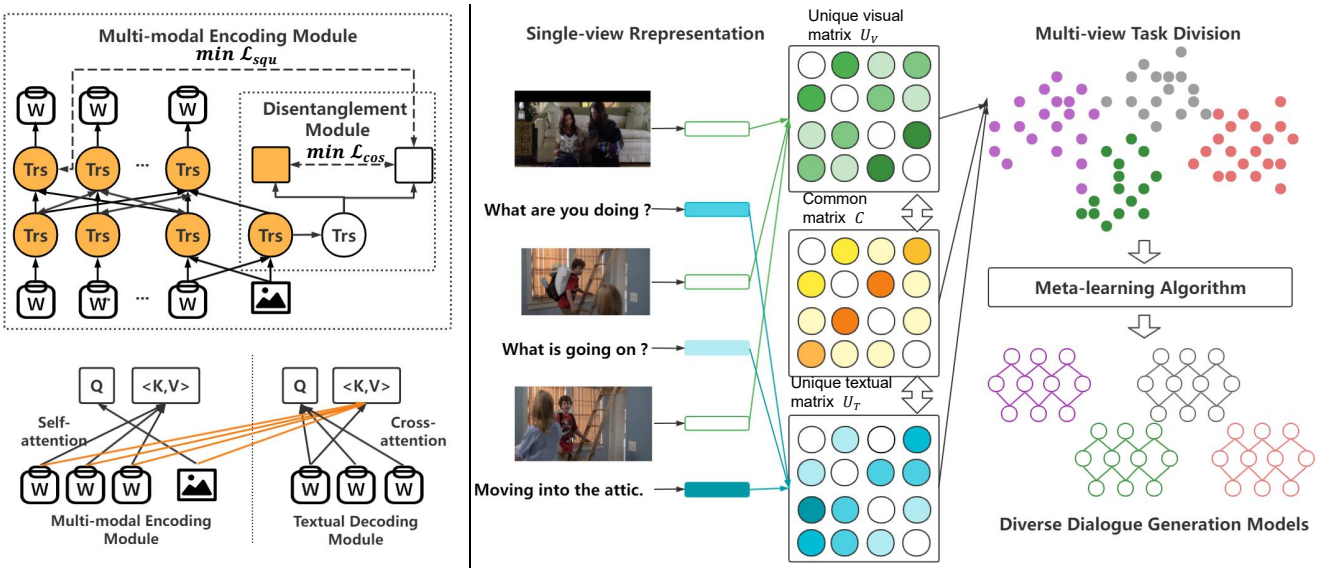

Figure 1: The sub-figure in the upper-left part shows the multi-modal dialogue generation model (omit the Textual Decoding Module). The sub-figure in the bottom-left part shows the self- and cross-attention in the generation model. The right side presents the procedure of the proposed Multi-view Meta-learning (MultiVML) algorithm.

3.5 Textual Decoding Module

The textual decoding module takes over the output of the multi-modal encoding module as well as the disentangled visual representation and generates the target response. We also use the Transformer as the decoder. The loss function of decoding is the negative log-likelihood of generating the response r given a sequence of input queries. So the total loss for the model is:

$$\mathcal{L} = -\log p(r|q_{t1}, q_{v1}, \cdots, q_{tn}, q_{vn}) + \lambda_1 \mathcal{L}_{squ} + \lambda_2 \mathcal{L}_{cos} \quad (3)$$

where λ_1 and λ_2 are balancing parameters.

4 MULTI-VIEW META LEARNING

We propose to imitate the meta-learning setting during training, so the similar training samples are clustered in the same group, while the dissimilar samples are distributed to different groups with different models. Here, we regard each group as a task in the meta-learning setting. We name this training algorithm as Multi-view Meta-Learning (MultiVML). In this section, we first introduce how to divide the training data into several meta-learning tasks, and then present the training and inference procedure of multi-modal response generation tasks. The whole training procedure is in Alg.1.

4.1 Meta-learning Task Division

To imitate the meta-learning setting, samples within a task (group) are similar from the perspective of multiple modalities. Multi-modal samples may have different characteristics in each modality, so we propose to view the samples from different views (e.g., textual view and image view) in the division of meta-learning tasks. Specifically, we first obtain the input's single-modal representation, and then draw out the commonness and individuality of each modality to get multi-view representations. Finally, we conduct a clustering of the training data from the perspective of multiple views and regard each cluster as a meta-learning task.

Algorithm 1: Multi-view Meta-learning

Input: Training set $\mathcal{S} = \{q_t, q_v, r\}_{i=0}^{n}$,
Cluster number n_k, Initial parameter Φ
Output: A set of generation models $\{\phi_i\}$.
//Meta-learning task division
//1. Single modal representation
for $\langle q_t, q_v, r\rangle$ *in* $\mathcal{S}$ **do**
  $x_t \leftarrow$ represent textual input q_t
  $x_v \leftarrow$ represent visual input q_v
$X_t \leftarrow \{x_t\}_i^n$;
$X_v \leftarrow \{x_v\}_i^n$;
//2. Multi-view representation
$C, U_t, U_v \leftarrow (X_t, X_v)$; // Use Eq. 4 to Eq. 7
//3. Multi-view clustering
$R_t, R_v \leftarrow (C, U_t, U_v)$; // Use Eq. 8 to Eq. 9
Obtain a set of tasks $\mathcal{T}$ from R_t and R_v
//Meta-Train
$\{\phi_i\} \leftarrow$ Reptile($\mathcal{T}, \Phi$);
return $\{\phi_i\}$

4.1.1 Single Modal Representation. The sample's input has two modalities: textual modality and visual modality. For the textual input, the averaged word embeddings of all the input sentences $q_t = \{q_{t1}, q_{t2}, \cdots, q_{tn}\}$ serve as the representations. Compared with the representations from the Multi-modal Encoding Module, word embedding can better reflect the differences among sentences. We use a matrix $X_t \in \mathbb{R}^{d \times n}$ to store all the textual features, where d is the dimension of the representation vector and n is the number of training samples. For the visual input, we use the average features of all the images $q_v = \{q_{v1}, q_{v2}, \cdots, q_{vn}\}$ from the pre-trained Resnet-50 model and denote the feature matrix as $X_v \in \mathbb{R}^{d \times n}$.

4.1.2 Multi-view Representation. Given the feature representation from two modalities X_t and X_v, we aim to cluster the training samples in the perspective of both textual and visual views. An intuitive solution for multi-view clustering is to directly apply the unsupervised clustering algorithm using the two feature matrices. This leads to two similar clustering results in two modalities because the two feature metrics contain almost the same information. While the unique information in each view is ignored. Hence, for each view, we need to not only emphasize the individuality of the data from this view but also consider its consistency with the other one.

We aim to learn a matrix $C \in \mathbb{R}^{n\times n}$ that describes the sample similarity regarding the common properties from all modalities, and we call it *common* view. We use two matrices $U_t \in \mathbb{R}^{n\times n}$ and $U_v \in \mathbb{R}^{n\times n}$ to describe the sample similarity regarding the unique properties from the textual and visual modality respectively, and we call them the *textual* and *visual* view. Given the feature matrix X_t and X_v, we aim to obtain the above matrices C, U_t and U_v. And the matrices C, U_t, and U_v ought to satisfy three properties as follows.

$$\mathcal{J}_{MV} = \mathcal{J}_R + \mathcal{J}_C + \mathcal{J}_H. \tag{4}$$

To explore the common and unique properties of the feature matrices, we borrow the idea from multi-view self-representation learning [53], which assumes that each sample can be reconstructed by a linear combination of similar samples from different perspectives. So a sample's feature representation x_i is close to a combination of the similar samples recorded in matrix C, U_t, and U_v. We use common matrix C and unique matrix U_t in textual view and U_v in visual view to reconstruct the original feature matrix X,

$$\mathcal{J}_R = \|X_t - X_t(C + U_t)\|_2^2 + \|X_v - X_v(C + U_v)\|_2^2. \tag{5}$$

The smaller $\mathcal{J}_R$ indicates that the original feature matrix and the reconstructed matrix are closer.

To ensure that the similarity calculation of C can reflect relationships of original samples, we calculate the L2 norm-based distance $\|C_i - C_j\|_2^2$ between every two samples from the *common* view C. The distance is weighted by the sample similarity based on *textual* view X_t and *visual* view X_v.

$$\mathcal{J}_C = \sum_{i,j=1,i\neq j}^{n} (W_{i,j}^t + W_{i,j}^v)\|C_i - C_j\|_2^2, \tag{6}$$

where $W_{i,j}^t$ and $W_{i,j}^v$ indicate the similarities of sample i and j. The similarity is calculated by the cosine similarity from X_t and X_v, and $W_{i,j} = \cos(x_i, x_j)$. The smaller $\mathcal{J}_C$ indicates two closer samples. That is, two samples similar in feature matrices (X_t and X_v) should also be similar in the perspective of the *common* view C.

Since the textual and visual inputs in the same sample have a large information overlap with each other, U_t and U_v will also be similar to each other. To alleviate this problem and enhance the uniqueness of each modality, we further use Hilbert-Schmidt Independence Criterion (HSIC) [13] to measure the dependency between two matrices U_t and U_v. HSIC calculates the squared norm of the cross-covariance over U_t and U_v in Hilbert Space, which is,

$$\begin{aligned} \mathcal{J}_H &= (n-1)^{-2}\mathrm{tr}(K_s H K_v H), \\ H &= I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \end{aligned} \tag{7}$$

where $\mathbf{K}$ is an $n \times n$ matrix, and $\mathbf{1}$ is an $n \times 1$ vector of ones. We adopt the Gaussian kernel to specify $k_{i,j} := \exp(-\sigma^{-2}\|U^i - U^j\|_2^2)$. The smaller $\mathcal{J}_H$ is, the more diverse U_t and U_v are.

4.1.3 Multi-view Clustering. We cluster the training samples using the multi-view representation obtained above. For the textual view, given the representation C and U_t, we conduct the clustering via a semi-nonnegative matrix factorization [46] as follows.

$$C + U_t = B_t(R_t)^\top, \tag{8}$$

where $B_t \in \mathbb{R}^{d\times n_k}$ is a matrix indicating the center of different clusters, and $R_t \in \mathbb{R}^{n\ times n_k}$ is a nonnegative matrix indicating the clustering result from the textual view. Note that n_k is the single-view cluster number. While for the visual view, to emphasize the unique information of visual input, we only use U_v to calculate the cluster indicator matrix $R_v \in \mathbb{R}^{n\times n_k}$.

Thus, the cluster results are not only dependent on the consistent representation C shared across multi-modal, but also the diverse representation U of different modalities. We apply L2-norm distance as the loss function to optimize and achieve Eq. 8 as Eq. 9. As a result, we obtain two different diverse clustering from the textual view and visual view.

$$\begin{aligned} J_{cluster} =& \|(C + U_t) - B_t(R_t)^\top\|_2^2 + \\ & \|U_v - B_v(R_v)^\top\|_2^2. \end{aligned} \tag{9}$$

To formulate the meta-learning setting and based on the cluster results from the textual and visual view, we distribute two samples belonging to the same task only if they are in the same cluster both in the textual and visual view. Otherwise, they are distributed into different tasks. We can adjust the cluster number by setting different values n_k, and there will be at most $n_k \times n_k$ clusters.

4.2 Meta-learning Training and Inference

4.2.1 Training. After the training sample division, we regard each cluster as a task and apply Reptile [40] to produce diverse generation models described in Sec. 3 for various tasks.

First, Reptile samples a task and k examples from this task. It updates the parameters ϕ by conducting the gradient descent for k steps to obtain the new task-specific parameter $\tilde{\phi}$ as,

$$\tilde{\phi} = \phi - g_1 - g_2 - \cdots - g_k \tag{10}$$

where g_i is the gradient descent at i-th step.

Second, to update the task-independent parameter ϕ, Reptile pushes ϕ toward the direction of $\tilde{\phi}$ at a slower pace than the updating pace of $\tilde{\phi}$ as

$$\phi = \phi + \epsilon\frac{1}{N}\sum_{i=1}^{N}(\tilde{\phi}_i - \phi) \tag{11}$$

where N is the task number and ϵ is the learning rate.

The two steps are alternatively performed. Finally, we obtain the initial parameters ϕ and the adapted task-specific parameter ϕ_i for each cluster.

4.2.2 Inference. For a sample from the test set, we calculate the similarity between it and each cluster's center via cosine, distribute it to the most similar cluster, and then generate responses via the corresponding generation model. Particularly, we use the same representation method introduced in Sec. 4.1.2.

5 EXPERIMENTS

5.1 Experimental Settings

We use two multi-modal open-domain dialogue datasets with millions of dialogue turns, OpenViDial 1.0 [37] and OpenViDial 2.0

Table 1: Dataset Statistics.

Statistics (OpenViDial)	V1.0	V2.0
Train	1M	4.6M
Dev	50K	0.5M
Test	50K	0.5M
Number of Turns	1.1M	5.6M
Vocab size before BPE	70K	278K
Vocab size after BPE	30K	30K
Average length of each episode	14	48
Average length of each turn	7.6	8.3

[65]. The dialogue turns and the corresponding images are extracted from movies and TV series.

Tab. 1 reports detailed statistics for OpenViDial 1.0 and OpenViDial 2.0. OpenViDial 1.0 [37] contains 1.1M dialogue turns, and the training/dev/test sets contain 1M/50K/50K turns. Each episode consists of 14 words on average, and each turn consists of 7.6 words on average. The vocabulary size after the BPE tokenizer is 30K. OpenViDial 2.0 [65] contains 5.6M turns, and the training/dev/test sets contain 4.6M/0.5M/0.5M turns. Each episode consists of 48 words on average, and each turn consists of 8.3 words on average. The dev set is only used for early stopping during the model training. The vocabulary size after the BPE tokenizer is 30K. The textual data of each dialogue turn comes from the subtitles of the movies or TV series. The image is the screenshot of the movies or TV series.

5.2 Implementation Details

The batch size is set to 64. We set the dropout rate to 0.3. The model is optimized by Adam [19] with the initial learning rate of 3e-5, and the default single-view cluster number n_k is 10. λ_1 and λ_2 in the loss function $\mathcal{L}$ are both set to 2, where we tried value $\{0.1, 0.25, 0.5, 1, 2, 5, 10\}$ to search the value. We train all the models on a single GeForce RTX 3090 GPU with 24GB memory. We implement our model based on a pre-trained multi-modal text generation model Oscar [30]. Following the setting of "Bert-base" [2], the hidden state dimension is 768, the dimension of image features is 1000, and we directly use the visual vector provided in the dataset (we use the coarse visual feature).

5.3 Evaluation Metrics

We perform both automatic and human evaluation. For the automatic metrics, **BLEU** [44] measures the overall quality by calculating the similarity between the ground truth and generated textual output. To measure diversity, **Dist-n** [27] calculates the count of unique n-grams (n=1,2,3,4) in all the generated responses, and **Entropy** [39] evaluates how evenly the empirical n-gram (n=3,4) distribution is.

We invited 5 voluntary annotators to conduct the human evaluation. Each annotator is asked to annotate 200 generated replies for each dataset. The generated replies are shuffled along the methods and the annotators cannot access the method name. The annotators are well-educated graduated students. The annotators are required to conduct the human evaluations in three aspects: (1) **H-quality** measures the quality of generated responses in terms of relevance, fluency, and grammaticality. The scoring range is from 1 to 5; (2) **H-Info** reflects the informativeness of generated responses. The scoring range is also from 1 to 5; (3) **H-Pct** is the percentage of universal responses among all generated responses. Following [56], annotators are asked to label whether a given response is a universal response or not. For example, the responses "I don't know" and "I'm OK" are universal responses.

[2]https://github.com/microsoft/Oscar

5.4 Competing Methods

There are four types of competing methods: **Conventional Generation** including *Seq2Seq* and *Trans*; **Diverse Generation** including *Trans+LS* and *Trans+CVAE*; **Multi-modal Generation** including *Trans+Mem*, *Trans+VDM-CV*, *Trans+VDM-FV*, *Trans+VDM-CV+Oscar* and *Trans+VDM-FV+Oscar*; **Meta-learning based Model** including the three variants of our proposed model. Except for **Conventional Generation**, all the single-modal models (text-to-text models) use the same way as our base model described in Sec.3 to process the visual signal (without the disentanglement strategy).

- *Seq2Seq* [52] is the widely used LSTM-based generation model. And it does not use image information.
- *Trans* (Transformer) [63] is proved to have strong ability for sequence modelling than LSTM and GRU-based model. Most of the following baselines use the Transformer as the base model, so it provides a basic comparison performance.
- *Trans+LS* [67] uses several soft labels instead of the ground truth sentence for each sample to supervise the model.
- *Trans+CVAE* [31] incorporates CVAE into the Transformer framework.
- *Trans+Mem* [23] use a memory to fuse multi-modal signals.
- *Trans+VDM-CV* and *Trans+VDM-FV* are from visual dialog model [66]. *CV* is the variant using image-level features; *FV* is the variant using objective-level features. We equip them with a multi-modal pre-trained model Oscar [30], denoting as *Trans+VDM-CV+Oscar* and *Trans+VDM-FV+Oscar*.
- *MultiVML* is our proposed model.
- *Trans+Reptile+Oscar* applies Reptile under the textual view task division to the conversation model.
- *MultiVML+Oscar* employs the pre-trained model Oscar as our model initialization.

5.5 Overall Performance

The overall performances are shown in Tab. 2. **Conventional Generation** methods provide a borderline for all the competing methods. They achieve relatively lower performance on all metrics since they only use the textual but no visual query for response generation. **Diverse Generation** methods work better than conventional ones. *Trans+LS* gains improvements by suppressing the probability of the target word and the probability of low-frequency words. While *Trans+CVAE* has higher diversity scores but lower quality scores, which indicates that CVAE helps to increase the uncertainty of the model but may lead to worse performance on quality. **Multi-modal Generation** methods generally work better than the above methods, since they fit well in the multi-modal setting. *Trans+Mem* builds a memory cell to store the multi-modal information, so the chatting history can be better memorized. *Trans+VDM-CV* and *Trans+VDM-FV* use the coarse and fine-grained image feature as the visual input, and the latter achieves better performance on almost all the metrics. However, when augmenting these two with

Table 2: The overall performance of all methods in quality and diversity. The bold numbers refer to the best performance among all the methods. The underlined numbers in automatic metrics indicate that our model's improvements over all baselines are statistically significant in the t-test of $p < 0.05$.

	Quality		Diversity							
	BLEU	**H-Quality**	**Dist-1**	**Dist-2**	**Dist-3**	**Dist-4**	**Ent-3**	**Ent-4**	**H-Info**	**H-Pct ↓**
OpenViDial V1.0										
Seq2Seq	0.45	1.91	0.001	0.001	0.002	0.003	2.69	2.81	1.41	41.3%
Trans	0.99	2.39	0.001	0.002	0.004	0.005	3.53	3.56	1.86	29.5%
Trans+LS	1.04	2.21	0.002	0.008	0.019	0.024	4.42	4.93	2.10	17.2%
Trans+CVAE	0.94	2.32	0.002	0.014	0.027	0.041	4.94	5.19	2.04	23.8%
Trans+Mem	1.15	2.41	0.002	0.007	0.019	0.031	4.16	4.80	1.95	32.1%
Trans+VDM-CV	1.16	2.56	0.002	0.009	0.018	0.028	5.01	5.18	2.33	26.4%
Trans+VDM-FV	1.19	2.41	0.003	0.012	0.026	0.043	4.82	4.99	2.09	19.3%
Trans+VDM-CV+Oscar	1.21	2.53	0.004	0.015	**0.036**	0.041	5.78	6.07	2.40	14.6%
Trans+VDM-FV+Oscar	1.23	2.40	0.003	0.012	0.024	0.035	5.27	5.98	2.25	17.7%
MultiVML (Ours)	1.22	**2.63**	0.004	**0.019**	0.032	0.048	5.66	5.84	2.51	**8.6%**
Trans+Reptile+Oscar	1.24	2.55	0.004	0.014	0.024	0.041	5.76	5.29	2.39	11.6%
MultiVML+Oscar (Ours)	**1.28**	2.61	**0.005**	0.018	0.033	**0.056**	**6.29**	**6.48**	**2.58**	9.7%
OpenViDial V2.0										
Seq2Seq	1.06	2.33	0.002	0.014	0.036	0.081	6.45	7.13	1.76	33.4%
Trans	1.96	2.56	0.004	0.031	0.095	0.163	7.84	8.49	2.31	24.0%
Trans+LS	1.71	2.45	0.004	0.043	0.113	0.189	8.12	8.38	2.39	13.5%
Trans+CVAE	1.64	2.51	0.004	0.034	0.118	0.230	8.41	8.75	2.45	19.8%
Trans+Mem	2.11	2.36	0.005	0.034	0.124	0.196	8.14	8.67	2.21	21.8%
Trans+VDM-CV	1.98	2.64	0.005	0.039	0.109	0.177	8.39	8.87	2.52	9.5%
Trans+VDM-FV	2.00	2.51	0.006	0.046	0.132	0.231	8.19	8.59	2.30	17.4%
Trans+VDM-CV+Oscar	1.95	2.71	0.007	0.044	0.151	0.253	8.14	8.77	2.58	**6.6%**
Trans+VDM-FV+Oscar	2.09	2.44	0.006	0.041	0.139	0.240	8.33	8.64	2.51	13.3%
MultiVML (Ours)	**2.18**	**2.79**	0.007	0.056	0.147	0.256	8.69	**9.13**	2.49	8.4%
Trans+Reptile+Oscar	2.02	2.49	0.006	0.043	0.135	0.248	8.26	8.69	2.47	8.6%
MultiVML+Oscar (Ours)	2.17	2.74	**0.008**	**0.057**	**0.167**	**0.268**	**8.75**	9.04	**2.89**	7.1%

the Oscar pre-training, the former surprisingly surpasses the latter. Hence, we choose the coarse image features in our full model.

As for **Meta-learning based Models**, they especially work well on diversity scores. This proves that meta-learning encourages the specific characteristics of each task. *MultiVML* works worse than *MultiVML+Oscar* on quality measurements, which indicates that multi-modal pre-training is helpful for response generation. *MultiVML+Oscar* method achieves the best results in most metrics, especially significant improvements on diversity criterion. We attribute it to the multi-view training algorithm encouraging unique information from multiple modalities, thereby making the generated sentence more informative and diverse.

Table 3: The ablation studies on OpenViDial V2.0. "−" indicates the variant without a specific component. TextVML and VisualVML conduct sample clustering on a single modality view, i.e., textual or visual view.

	BLEU	**Dist-3**	**Dist-4**	**Ent-4**
MultiVML+Oscar	2.17	**0.167**	**0.268**	**9.04**
− Disentangle	2.04	0.148	0.257	9.01
TextVML	2.11	0.149	0.251	8.76
VisualVML	1.96	0.154	0.258	8.67
− Meta-learning	**2.18**	0.137	0.243	8.59
− Reptile + MAML	2.13	0.159	0.262	8.84

5.6 Ablation Studies

The ablation study is shown in Tab. 3. The variant *−Disentangle* removes the Disentanglement module from the generation model. The worse performances on all metrics, especially the distinct scores, indicate that the disentanglement strategy is essential to enhance the unique information in the visual input. *TextVML* and *VisualVML* variants cluster the training data in a single view in a meta-learning setting. The obvious performance drops of both variants prove the effectiveness of multi-view clustering. As a comparison, *TextVML* is better than *VisualVML*, which caters to the intuition that textual data act as the dominant information resource in textual output generation. Instead of customizing diverse generation models via the meta-learning paradigm, the variant − *Meta-learning* trains a unified model using all the training data. This variant achieves a very slight improvement (only 0.1 compared to *MultiVML+Oscar*) in BLEU score but suffers from the lowest diversity scores among all variants. This proves that using meta-learning for model training can marginally improve diversity, and also confers a comparable quality. − *Reptile + MAML* replaces the meta-learning algorithm Reptile with MAML and achieves worse performance on all metrics. This indicates that the first-order method is more stable than the second-order one, especially in the text generation tasks.

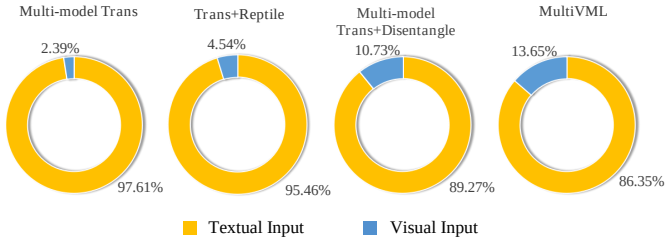

Figure 2: The average attention scores of textual input (in blue) and visual input (in orange), where the scores indicate the contribution of two modalities.

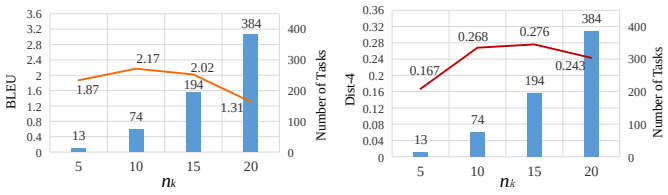

Figure 3: The histogram presents the number of multi-view tasks (bar chart) and performances (broken line) in different n_k (from 5 to 20) settings. The left broken line graph is the BLEU score, and the right is the corresponding Dist-4 score.

5.7 Analyses on Modality Bias

Our proposed model aims to relieve the "modality bias" [16, 32] in multi-modal text generation. To verify our model's effectiveness, we quantify the phenomenon of "modality bias" in baseline models and our proposed method, where we calculate the attention weight ratio over textual and visual input in the decoder. Fig. 2 shows the results in different methods. For *Multi-modal Trans*, which feeds the image feature as one time step into the encoder, attention is mostly located on textual data. *Trans+Reptile* uses Reptile to train the generation model under the textual view, and the attention on visual input increases, indicating that the meta-learning framework and data division during training helps the model to concentrate more information from various modalities. When adding the disentanglement module into the multi-modal generation model (*Multi-modal Trans+Disentangle*), more attention is shifted to visual input. This demonstrates the effectiveness of disentanglement operation in our model. Compared with *Multi-modal Trans+Disentangle* that makes task division only from the textual view, *MultiVML* uses the proposed multi-view clustering for task division. The result shows that both the multi-view cluster and disentanglement operation boost the contribution of visual input.

5.8 Contribution of Clustering

We verify the contribution of our proposed multi-view clustering. Ablation studies (Sec. 5.6) have shown the multi-view clustering performs better than single-view clustering. Here, we further investigate the contribution of clustering on samples compared to the randomly grouped samples. We randomly shuffle partial data in the task division and present the results in Tab 4. When 50% samples (Shuffle 50%) are not similar with other samples in one meta task, the performance drops in both quality and diversity. When all the samples are randomly distributed to compose the meta tasks (Shuffle 100%), the meta-learning based method achieves even worse performance than the non-meta-learning model (*Trans+Reptile+Oscar*). These phenomena suggest that a proper clustering is quite crucial for the model performance. The more similar the samples (from both modalities) within a task is, the better performance the proposed method achieves.

5.9 Analyses on Cluster Number

To analyze the influence of cluster number, we present the quality (BLEU) and diversity measurement (Dist-4) in different cluster number settings. n_k is the single-view cluster number that we can set as the hyper-parameter, and the multi-view cluster number is at least n_k and at most $n_k \times n_k$. As shown in Fig. 3, we set n_k to {5,10,15,20} for analyses on multi-view cluster numbers. We can see that when n_k is small (5), the BLEU score (1.87) is close to the highest one (2.17), while the Dist-4 (0.167) is much lower than the highest one (0.276). When n_k is large (20), the cluster number almost reaches the maximum value (400) and both quality and diversity are unsatisfactory. This illustrates that the samples are too dispersed, and similar samples may not be distributed into the same task, thereby harming the model performance. 10 and 15 are two good options for n_k, and we use 10 in our main experiment as it achieves a good balance in both quality and diversity.

Table 4: The task division analysis on OpenViDial V2.0. "Shuffle x%" indicates x% of samples are randomly shuffled and assigned to different task for model training.

	BLEU	Dist-3	Dist-4	Ent-4
MultiVML+Oscar	**2.17**	**0.167**	**0.268**	**9.04**
Shuffle 50%	2.02	0.141	0.237	8.71
Shuffle 100%	1.98	0.122	0.195	8.66
Trans+Reptile+Oscar	2.02	0.135	0.248	8.69

6 CONCLUSION

In this paper, we introduce a multi-view meta-learning (MultiVML) algorithm for building multi-modal chatbots. To alleviate the modality-bias problem, MultiVML first extracts the common information shared by all modalities and the modal-specific information belonging to each modality. MultiVML then divides the training data into several groups using the extracted modal-specific information and applies meta-learning to customize generation models for each group according to the data division. Generation diversity also benefits from customizing models for various scenarios. Our generation model is also equipped with a disentanglement strategy to emphasize non-textual information. Experiment results show that our method outperforms competitive baselines in both diversity and quality, and can also encourage the model to attend more to the visual information.

ACKNOWLEDGMENTS

This paper is supported by National Natural Science Foundation of China (NSFC Grant No. 62106275).

REFERENCES

[1] Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D Trippe, Juan B Gutierrez, and Krys Kochut. 2017. Text summarization techniques: a brief survey. *arXiv preprint arXiv:1707.02268* (2017).

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *2018 IEEE Conference on Computer Vision and Pattern Recognition*. 6077–6086.

[3] D. Bahdanau, K. Cho, and Y. Bengio. 2014. Neural machine translation by jointly learning to align and translate. *Computer Science* (2014).

[4] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 4762–4779.

[5] Iacer Calixto, Miguel Rios, and Wilker Aziz. 2019. Latent variable model for multimodal translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6392–6405.

[6] Yu Cao, Liang Ding, Zhiliang Tian, and Meng Fang. 2021. Towards Efficiently Diversifying Dialogue Generation Via Embedding Augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7443–7447.

[7] Jiachen Du, Wenjie Li, Yulan He, Ruifeng Xu, Lidong Bing, and Xuan Wang. 2018. Variational autoregressive decoder for neural response generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3154–3163.

[8] Yan Duan, John Schulman, Xi Chen, Peter L Bartlett, Ilya Sutskever, and Pieter Abbeel. 2016. Rl 2: Fast reinforcement learning via slow reinforcement learning. *arXiv preprint arXiv:1611.02779* (2016).

[9] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical Neural Story Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 889–898.

[10] Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (ICML)* (Sydney, NSW, Australia) *(ICML'17)*. JMLR.org, 1126–1135.

[11] Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, Guodong Zhou, and Shuming Shi. 2019. A Discrete CVAE for Response Generation on Short-Text Conversation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 1898–1908.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*. 6325–6334.

[13] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Sch¨olkopf. 2005. Measuring statistical dependence with hilbert-schmidt norms. In *16th International Conference of Algorithmic Learning Theory*. 63–77.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[15] Po-Yao Huang, Xiaojun Chang, and Alexander Hauptmann. 2019. Multi-head attention with diversity for learning grounded multilingual multimodal representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 1461–1467.

[16] Yan Huang, Qiang Wu, Jingsong Xu, Yi Zhong, Peng Zhang, and Zhaoxiang Zhang. 2022. Alleviating Modality Bias Training for Infrared-Visible Person Re-Identification. *IEEE Transactions on Multimedia* 24 (2022), 1570–1582. https://doi.org/10.1109/TMM.2021.3067760

[17] Unnat Jain, Ziyu Zhang, and Alexander G. Schwing. 2017. Creativity: Generating Diverse Questions Using Variational Autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Jaehun Jung, Bokyung Son, and Sungwon Lyu. 2020. Attnio: Knowledge graph exploration with in-and-out attention flow for knowledge-grounded dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 3484–3497.

[19] Diederik P Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations*.

[20] Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. 2019. Information Maximizing Visual Question Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *arXiv preprint arXiv:1207.6083* (2012).

[22] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020. MART: Memory-augmented recurrent transformer for coherent video paragraph captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, Online, 2603–2614. https://doi.org/10.18653/v1/2020.acl-main.233

[23] Jie Lei, Liwei Wang, Yelong Shen, Dong Yu, Tamara Berg, and Mohit Bansal. 2020. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2603–2614.

[24] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, and Daxin Jiang. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 11336–11344.

[25] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*. 110–119.

[26] Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 994–1003.

[27] Jiwei Li and Dan Jurafsky. 2016. Mutual information and diverse decoding improve neural machine translation. *arXiv preprint arXiv:1601.00372* (2016).

[28] Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. 2020. Cross-media structured common space for multimedia event extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2557–2568.

[29] Xin Li, Piji Li, Wei Bi, Xiaojiang Liu, and Wai Lam. 2020. Relevance-promoting language model for short-text conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8253–8260.

[30] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*. Springer, 121–137.

[31] Zhaojiang Lin, Genta Indra Winata, Peng Xu, Zihan Liu, and Pascale Fung. 2020. Variational transformers for diverse response generation. *arXiv preprint arXiv:2003.12738* (2020).

[32] Agnes Lisowska, Susan Armstrong, Mireille Betrancourt, and Martin Rajman. 2007. Minimizing Modality Bias When Exploring Input Preferences for Multimodal Systems in New Domains: The Archivus Case Study. In *CHI '07 Extended Abstracts on Human Factors in Computing Systems* (San Jose, CA, USA) *(CHI EA '07)*. Association for Computing Machinery, New York, NY, USA, 1805–1810. https://doi.org/10.1145/1240866.1240903

[33] Hui Liu and Xiaojun Wan. 2021. Video paragraph captioning as a text summarization task. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*. 55–60.

[34] Zequn Liu, Ruiyi Zhang, Yiping Song, and Ming Zhang. 2020. When does maml work the best? An empirical study on model-agnostic meta-learning in nlp applications. *arXiv preprint arXiv:2005.11700* (2020).

[35] Andrea Madotto, Zhaojiang Lin, Chien-Sheng Wu, and Pascale Fung. 2019. Personalizing dialogue agents via meta-learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5454–5459.

[36] Bodhisattwa Prasad Majumder, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Julian McAuley. 2020. Like Hiking? You Probably Enjoy Nature: Persona-grounded Dialog with Commonsense Expansions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 9194–9206.

[37] Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2021. OpenViDial: A Large-Scale, Open-Domain Dialogue Dataset with Visual Contexts. *arXiv preprint arXiv:2012.15015* (2021).

[38] Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 845–854.

[39] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to Backward and Forward Sequences: A Content-Introducing Approach to Generative Short-Text Conversation. In *26th International Conference on Computational Linguistics*. 3349–3358.

[40] Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999* (2018).

[41] Abiola Obamuyide and Andreas Vlachos. 2020. Model-agnostic meta-learning for relation classification with limited supervision. (2020).

[42] Abiola Obamuyide, Andreas Vlachos, et al. 2019. Meta-learning improves lifelong relation extraction. (2019).

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

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

[45] P Singh Pasi, S. Nemani, P. Jyothi, and G. Ramakrishnan. 2022. Investigating Modality Bias in Audio Visual Video Parsing. *arXiv e-prints* (2022).

[46] Chong Peng, Zhilu Zhang, Chenglizhao Chen, Zhao Kang, and Qiang Cheng. 2022. Two-dimensional Semi-nonnegative Matrix Factorization for Clustering. *Information Sciences* (2022), 106–141.

[47] Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2639–2649.

[48] Lianhui Qin, Michel Galley, Chris Brockett, Xiaodong Liu, Xiang Gao, Bill Dolan, Yejin Choi, and Jianfeng Gao. 2019. Conversing by reading: contentful neural conversation with on-demand machine reading. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 5427–5436.

[49] S. Ravi and H. Larochelle. 2017. Optimization as a model for few-shot learning. In *International Conference on Learning Representations*.
[50] Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2020. Thinking globally, acting locally: distantly supervised global-to-local knowledge selection for background based bonversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8697–8704.
[51] Adam Santoro, Sergey Bartunov, Matthew Botvinick, Daan Wierstra, and Timothy Lillicrap. 2016. Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning*. PMLR, 1842–1850.
[52] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53th Annual Meeting of the Association for Computational Linguistics*. 1577–1586.
[53] Yao Shixin, Yu Guoxian, Wang Jun, Carlotta Domeniconi, and Zhang Xiangliang. 2019. Multi-View Multiple Clustering. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. 4121–4127.
[54] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. 2015. Learning structured output representation using deep conditional generative models, In Proceedings of the 28th International Conference on Neural Information Processing Systems. *NeurIPS* 28, 3483–3491.
[55] Yiping Song, Zequn Liu, Wei Bi, Rui Yan, and Ming Zhang. 2020. Learning to customize model structures for few-shot dialogue generation tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5832–5841.
[56] Yiping Song, Rui Yan, Yansong Feng, Yaoyuan Zhang, Dongyan Zhao, and Ming Zhang. 2018. Towards a neural conversation model with diversity net using determinantal point processes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[57] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*.
[58] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to Fine-Tune BERT for Text Classification? *Springer, Cham* (2019).
[59] Qianru Sun, Yaoyao Liu, Tat-Seng Chua, and Bernt Schiele. 2019. Meta-transfer learning for few-shot learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 403–412.
[60] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems*. 3104–3112.
[61] Zhiliang Tian, Wei Bi, Zihan Zhang, Dongkyu Lee, Yiping Song, and Nevin L Zhang. 2021. Learning from my friends: few-shot personalized conversation systems via social networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 13907–13915.
[62] Zhiliang Tian, Zhihua Wen, Zhenghao Wu, Yiping Song, Jintao Tang, Dongsheng Li, and Nevin L Zhang. 2022. Emotion-Aware Multimodal Pre-training for Image-Grounded Emotional Response Generation. In *Database Systems for Advanced Applications: 27th International Conference, DASFAA 2022, Virtual Event, April 11–14, 2022, Proceedings, Part III*. Springer, 3–19.
[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 6000–6010.
[64] Qingzhong Wang and Antoni B Chan. 2019. Towards diverse and accurate image captions via reinforcing determinantal point process. *arXiv preprint arXiv:1908.04919* (2019).
[65] Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021. OpenViDial 2.0: A Larger-Scale, Open-Domain Dialogue Generation Dataset with Visual Contexts. *arXiv preprint arXiv:2109.12761* (2021).
[66] Shuhe Wang, Yuxian Meng, Xiaofei Sun, Fei Wu, Rongbin Ouyang, Rui Yan, Tianwei Zhang, and Jiwei Li. 2021. Modeling Text-visual Mutual Dependency for Multi-modal Dialog Generation. *arXiv preprint arXiv:2105.14445* (2021).
[67] Yida Wang, Yinhe Zheng, Yong Jiang, and Minlie Huang. 2021. Diversifying Dialog Generation via Adaptive Label Smoothing. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. 3507–3520.
[68] Huaxiu Yao, Ying Wei, Junzhou Huang, and Zhenhui Li. 2019. Hierarchically structured meta-learning. In *International Conference on Machine Learning*. PMLR, 7045–7054.
[69] Shaowei Yao and Xiaojun Wan. 2020. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4346–4350.
[70] Zhi-Xiu Ye and Zhen-Hua Ling. 2019. Multi-level matching and aggregation network for few-shot relation classification. *arXiv preprint arXiv:1906.06678* (2019).
[71] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3025–3035.
[72] Yongjian You, Weijia Jia, Tianyi Liu, and Wenmian Yang. 2019. Improving abstractive document summarization with salient information modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2132–2141.
[73] Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. 2020. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2031–2043.
[74] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2204–2213.
[75] Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. 654–664.
[76] Yingxiu Zhao, Zhiliang Tian, Huaxiu Yao, Yinhe Zheng, Dongkyu Lee, Yiping Song, Jian Sun, and Nevin L Zhang. 2022. Improving meta-learning for low-resource text classification and generation via memory imitation. *arXiv preprint arXiv:2203.11670* (2022).
[77] Junnan Zhu, Yu Zhou, Jiajun Zhang, Haoran Li, Chengqing Zong, and Changliang Li. 2020. Multimodal summarization with guidance of multimodal reference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 9749–9756.

A ETHICAL CONSIDERATIONS

Our work focuses on improving the diversity of generation models, so as to promote generating more informative and diverse responses. The proposed method can be applied to a wide range of real-world applications, such as chatbots and intelligent customer service. Nonetheless, our method, along with other text generation methods, may be potentially used with illegal intentions. For example, malicious users may deliberately manipulate the generated text for private interest. Besides, forgery issues, e.g., misinformative and fake response generation, are also a severe challenge to the use of our method. Fortunately, our method generates diverse responses is only designed for conversation tasks instead of the news generation tasks, and there is a gap between the two tasks. We couple our released code with strict misrepresentation, offensiveness, and bias checks, and restrict it only for academic use. With more discussion on the common ethical problem among all generation models, we encourage further works to explore scientific solutions for such cases.

Besides, our work verifies the effectiveness of the proposed method in human evaluation, we set the hourly pay to 20 US$, which exceeds the statutory minimum wage and also excels the average local salary level.