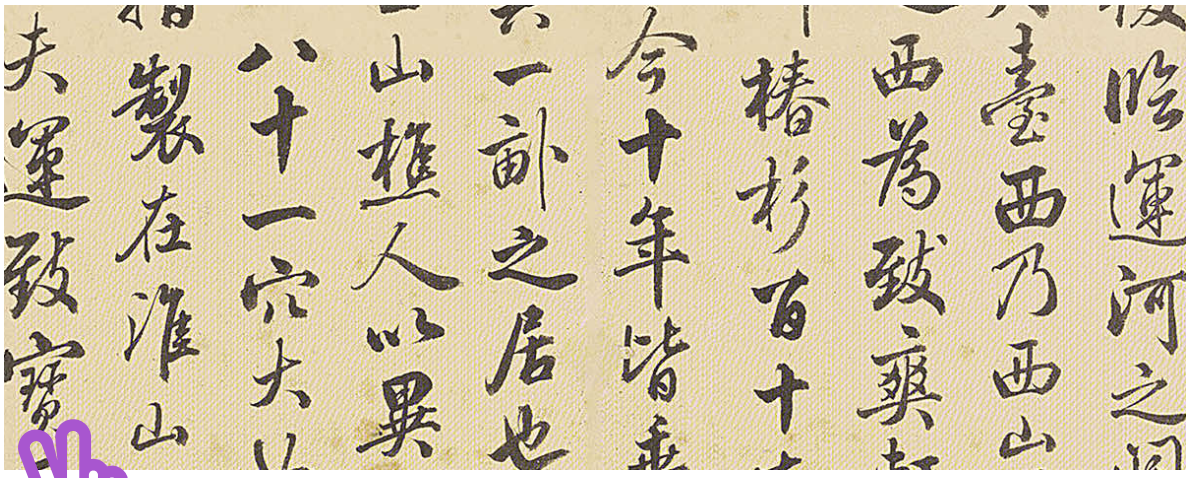

MMDialog

摘要

本文提出了 MMDialog 数据集，由 108 万个真实世界对话组成，包含 153 万个独特的图像，涉及 4184 个主题。

MMDialog 有两个独特的优势：

1. 按对话数量计算的最大的多模态对话数据集
2. 包含了大量的话题

贡献：

1. 提出了两个基于检索和生成的反应预测任务
2. 建立了两个基线模型
3. 评估多模态反应的 MM-relevance 指标

开源

https://github.com/victorsungo/MMDialog

Introducation

1. 多模态对话代理的需求很大

2. 数据集有一些
    a. IGC，imagechat
    b. openvidial
    c. photochat ，mmchat
3. 这些数据集有局限性
    a. visual dialog，igc，imagechat 都是谈论给定图像，话题由图像支持或者引发，跟日常交流不一样
    b. openvidial 等数据集不是来自于真实的对话场景
    c. photochat，mmchat 引入的真实对话，但是规模小，缺乏多样性
4. 为了解决上面的问题，本文贡献
    a. mmdialog 数据集
        i. 包含 100 万个对话场景，153 万张图片
        ii. 我们在数据收集阶段设计了一系列的数据过滤过程
            1. 平均每个对话 session 有 2.59 张图片
    b. 两个任务：检索式的和生成式的
        i. 实现了两个 baseline
    c. 多模态反应预测中，生成的模态顺序与真实 response 不一致
        i. MM-relevance

Dialogue Creation

来自全球性的社交媒体，数据集收集有三个阶段：

1. 广泛的手工收集用户常用的标签，覆盖尽可能多的领域
2. 从种子标签开始收集，收集所有带有上述标签的对话，并且只保留包含至少一张图片的对话（称之为 anchors，锚点）对于每个锚点，检索所有回复他的对话和他回复的对话
3. 设计一系列数据过滤和后处理脚本，提高对话质量

hashtag collection

抓取了最有影响力的在线社交平台，使用学术可用的 API，tags：#travel，#friends，#golf

手动筛选了 4184 个流行标签，每个标签至少有 1000 个对话

Figure 4: 200 most popular hashtags in MMDialog weighted by their frequencies.

multi-modal conversations construction

对于每个标签，抓取含有标签，且至少包含一个图像的对话（锚点）

递归的追踪包含锚点的对话

data filtering and post-processing

1. 删除包含明确攻击性词语和有毒声明的对话
2. 忽略并丢弃含有 gif 和视频的对话
3. 对话中删除不规则自负，比如 url 等
4. 将表情符号（emoji）和 hashtag 转换成相应的自然语言形式，保证对话的连贯性
5. 删除了自言自语（连续两次或者多次回复自己的情况）
6. 丢弃了不完整或者缺失的对话
7. 只保留不少于 3 个对话

corpus statistics

与 photochat 对比

Statistics	PhotoChat	MMDialog
#Language	English	English
#Open-domain	✘	✔
#Dialogues	12.29K	1.08M
#Images	10.92K	1.53M
#Turns	156.10K	4.92M
#Topics/Objects	89	4,184
Avg. #Turns per Dialogue	12.71	4.56
Avg. #Images per Dialogue	0.89	2.59
Avg. #Tokens per Turn	6.33	15.90

Table 1: Statistics of MMDialog and previous multi-modal dialogue dataset PhotoChat.

mmdialog 每个对话都比 photochat 包含更多的文本标记，可以传达更多的语义信息

Task definition

假设有一个多轮对话数据集 $\mathcal{D} = \{(U_i, R_i)\}_{i=1}^{n}$,ui 是对话上下文，Ri 是 Ui 的回复，U 和 R 可能包含文本元素（utterance）和视觉元素（image）

对于任意 U 和 R， $U_i = \{u_k^m\}_{k=1}^{K}$ 和 $R_i = \{r_l^m\}_{l=1}^{L}$ 作为多模态元素的序列。K 和 L 是元素的数量，m 是模态类型（t：文本，v：图像）

目标是学习多模态对话模型 g，从 U 中预测反应 R

task1：multi-modal response generation

模型 $P(R|U;\theta)$ ，给定一个上下文 U，可以直接生成一个多模态回应 R（文本，视觉，或者两者都有）

task2：multi-modal response retrieval

检索模型：每一个对话例子（U，R）都会额外一共一系列负面语料作为干扰，然后将 R 中的真实回复和负面例子构成一个候选集 C

response modal intent prediction

$$\forall j \in [1, J], \mathcal{I}\left(U, \tilde{R}_{<j}\right) \in \{0, 1, 2\}$$

多模态响应可以出现在任何一个地方，意图预测任务是一个分类任务，在给定的对话上下文中，不同模态的顺序；文本：0，图像：1，eof：2

Evaluation of multi-modal dialogue tasks

task1

文本部分，使用 bleu 和 rouge 来进行评估

task2

计算 recall 值

MM-Relevance

基于 clip 模型进行视觉语言匹配，以缓解模态错位的问题

Baseline

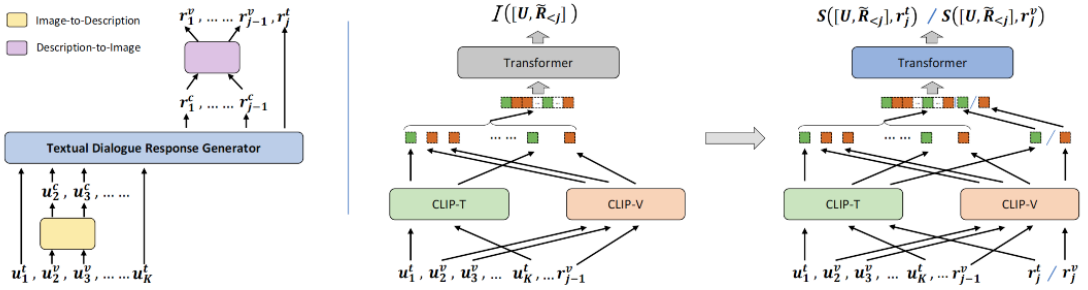

Figure 2: The overview of multi-modal response generation (a) and retrieval (b) baselines.

Multi-modal response generation model

使用最先进的多模态对话响应生成模型 Divter，该模型由两部分组成：文本对话响应生成器 G 和描述-图像翻译器 F

具体来说，G 将对话上下文 U 作为输入，然后生成一个可能包含文字回应 rt 或者文字描述 rc 的文本序列

U 中的图像会被替换成他们的描述 uc（借助描述-图像翻译器 F）

通过这种方式，可以将文本语句 ut 和图片描述串联起来作为 G 的输入。

此外，[UTT]作为文本的开头标识符，[DST]作为图像描述的开头。然后对于生成描述 rc（[DST]开头），F 将会生成一个真实的改分辨率的图像 rv

图片评价使用 IS 指标

Multi-modal response retrieval model

受 Parekh 等人的启发，建立了一个检索模型 R，名为 DE++；包含一个意图预测模块 $\mathcal{R}_\alpha$ 以及一个排名模块 $\mathcal{R}_\beta$

在每个 ranking action 开始之前 $\mathcal{R}_\alpha$ 首先将对话上下文 U 和先前的响应元素 R 作为输入并且预测：

1. 回复是否已经完成，是否需要停止检索一个新的元素；
2. 下一个元素的模态

如果是 1，那么 $\mathcal{R}_\alpha$ 将 U 和 R 作为输入来预测意图

如果是 2，那么 $\mathcal{R}_\beta$ 将会计算相关性分数 S，在相同情况下 $\mathcal{R}_\beta$ 将会计算 $\{r_z^m\}_{z=1}^Z$ 所有的候选答案并且选择相关性最高的作为最终的回复元素

具体来说， $\mathcal{R}_\alpha$ 和 $\mathcal{R}_\beta$ 有相似的结构，我们使用 clip 的文本编码器和图像编码器去分别表示文本和图像。

在 $\mathcal{R}_\alpha$ 中，我们链接所有的上下文使用一个特殊的可学习 token[CLS]放在句首，并且将序列送入 transformer 模块去预测意图；

在 $\mathcal{R}_\beta$ 中，我们将[CLS]嵌入到串联的语境嵌入或者候选嵌入中，然后将它们分别送入 transformer 模块中，之后我们可以得到上下文和候选前乳的保湿向量，并通过对上下文和候选向量的点乘相似度来计算相关性分数

Experiments

在 MMDialog 数据集上进行实验，评估两个任务和基线模型。我们评估了 response/intent 除了第一轮以外的所有轮次，并将之前的所有轮次视为上下文（context）

experimental setup

1. 对数据集抽样制作测试集和交叉验证集

2. 检索任务：随机抽取 999 个负面文本和负面图片（从同一个集合里），保持候选集合的总数为 1000.在训练阶段，负面内容是批量抽样的。
    a. 文本生成器：fine-tune 了 dialogpt（huggingface 的 dialogpt-medium）
    b. 描述-图片翻译器：实现了 DALL-E 使用 mega 版本，本文 finetune 了 DALL-E mega 了 1 个 epoch，学习率 1e-8，minibatch64
    c. 为了获得 MMDialog 的图片描述，我们使用 OFA-huge
    d. 所有版本的 clip model 都是用 clip-vit-base-patch32
3. 生成任务 Divter：
    a. 使用相同的实验结构

results of multi-modal baselines

生成模型

Models	Intent	Image Generation	Textual Response Generation			Multi-Modal Generation
	F1	IS↑	BLEU-1	BLEU-2	ROUGE-L	MM-Relevance↑
Divter (Sun et al., 2022)	71.77	20.53 ± 0.50	9.44	7.45	11.19	61.85

Table 3: Automatic evaluation results of the generative baseline on the test set of MMDialog. All numbers except "IS" and "MM-Relevance" are in percentage.

评估了文本响应生成，图像生成和意图预测任务

Divter 的文本生成性能比较差，但是图像生成任务比较好；意图预测任务的分数比较高，表明该模型由能力确定对话中该生成文本还是图像

检索模型

各个指标都不错