MMChat

Title

MMChat：Multi-Modal Chat Dataset on Social Media

Abstract

本文开发了 MMChat 数据集，包含 32.4M 原始对话和 120.84K 个过滤后的对话。对话是由区中提供或者从虚构电影中收集的。MMChat 包含社交媒体上收集来的以图像为基础的对话。特别的，以图像为起点的对话可能会偏移到与图像无关的话题。为了更好的解决这个问题，我们手工注释了 100K 条丢话并且过滤了语料，从而得到了 MMChat-HF。我们开发了一个 benchmark 模型来解决对话生成问题的稀疏性问题通过调整图像特征的注意力路由机制。

开源

https://github.com/silverriver/MMChat

Introduction

当前常识去建立像人类一样的开放域对话系统普遍有两个角度：

1. 用文本或者结构背景来丰富对话系统，比如知识，比如人格
2. 使对话系统理解多模态的上下文，比如视觉，语音甚至手势

当前存在的数据集的重要问题是，他们假设对话中的每一个话语都是基于给定的图像，但是正常的交流并不是这样。具体来说，通过图片触发的话题可能会在对话中偏离图像。我们把这种现象称之为稀疏性，表现出这种现象的对话叫做稀疏的以图像为基础的对话。

为了解决稀疏性问题，我们提出了 MMChat，不像之前的多模态对话数据集在每轮对话中只提供单独的原始图片，每一个过滤的对话场景在 MMChat 中都有一个或者多个图像应答，每张图片的语义信息通过一个预先训练好的 image-caption 模型来实现语义对齐。

具体来说，一系列被检测的对象标签和一个生成描述标题被对应到每一个 MMChat 的图片中。

为了改善 MMChat 的对话质量，我们抽样了 10W 个对话场景，并且手工检查图像与对话是否强关联。这就产生了一个人工过滤的 MMChat-HF 数据集包含 19.90K 个对话场景和 52.24K 个图像。

本文设计了一个比较强的 benchmark 模型去处理图像稀疏性问题，本文的贡献如下：

1. 我们建立了一个大型多模态对话数据集 MMChat，并且解决了稀疏性问题，提出了一个自动过滤过程来清理数据集。
2. 我们提供了一个人类过滤数据集 MMChat-HF
3. 我们建立了 benchmark 模型在 MMChat 中，结果表明视觉上下文对对话模型有积极作用，并且我们的模型在解决稀疏性问题上有帮助。

Dataset Construction：MMChat

来源：中文社交媒体

#(Dialogues)	120.84K
#(Total Images)	204.32K
#(Total Utterances)	314.13K
#(Dialogue Sessions) Longer than 4	17.32K
#(Image) per Dialogue	2.91
#(Utterance) per Dialogue	2.59
#(Character) per Utterance	8.52
#(Raw Dialogues)	32.4M

Table 1: Statistics of MMCHAT.

Data Collection

两阶段 pipeline 去收集对话：

1. 第一阶段的目标是收集在社交媒体活跃的种子用户；
2. 收集种子用户的对话，收集他们的图片和图片评论；

Data Filtering and Post-processing

为了改善 MMChat 的质量，一系列规则被谨慎的设计用来过滤低质量图像和对话。具体来说，低质量图像或者高长宽比的图像被丢弃，并且对话包含极长句子（>200）被过滤掉。此外，我们只保留超过 3 轮的对话。冒犯性的对话也被过滤掉（冒犯词表和冒犯评论过滤器）。

为了确保对话文本与图像相关（第一轮），我们只保留那些直接上传模式的图像。这个模式允许用户不需要提供文本就分享图片。我们认为，在这些仅有图像的帖子之后，对话的最初几个回合通常是由视觉信息引发的，因为除了上传的图像之外，之前没有任何文本背景。

采用这样的规则是希望利用有限的计算资源多关注多模态之间更加密切的语境。

Human Filtering

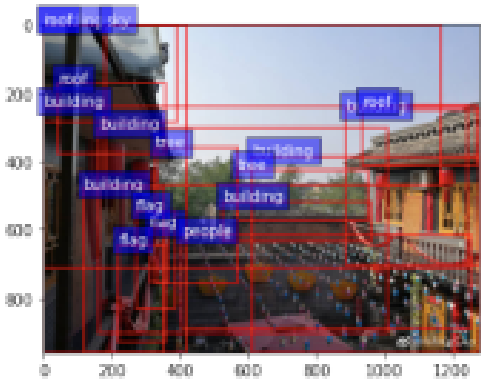
具体来说，我们随机抽样 100K 个对话场景，并且令标注员去标注每一个对话场景，主要标注以下方面：

1. 相关图像是否合格，不合格的标准：
    a. 重叠/长宽比过大，文本的截图
    b. 自拍，有攻击性的图片
    c. 对话内容是否与图像密切相关
2. 如果内容包含物体/人/背景，会被注释为 true

在 MMChat-HF 中，只保留被注释为 true 的对话场景。

Data Analysis

我们尝试去揭示 MMChat 中图像的语义信息。具体来说，我们使用 Faster R-CNN 模型来检测每张图片的污图，任何类别的置信度超过阈值（0.2）就会被选中。每张图片平均检测到 11.42 个物体。



Caption: Flags on the street.
Detected Objects: building, blue sky, red flags, roof, people, trees

Dialogue Generation on MMChat

Task Definition

MMDS 的任务是去学习函数 f 能够映射文本上下文 $\mathcal{C}$ 和多模态上下文 $\mathcal{I}$ 为对话回复 Y，$\{\mathcal{C},\mathcal{I}\} \rightarrow Y$

在这个研究中，我们关注图像模态， $\mathcal{I}$ 由一系列图像组成： $\{I_n\}_{n=1}^{N}$

Dialogue Generation Model

seq2seq 架构被用作我们的主干网络。如图，

双编码器被用作分别编码文本上下文和图像上下文为 E_c, E_I

一个注意力路由模型被用作去混合 E_c, E_I 到 decoder 中，response Y 是自回归编码的。

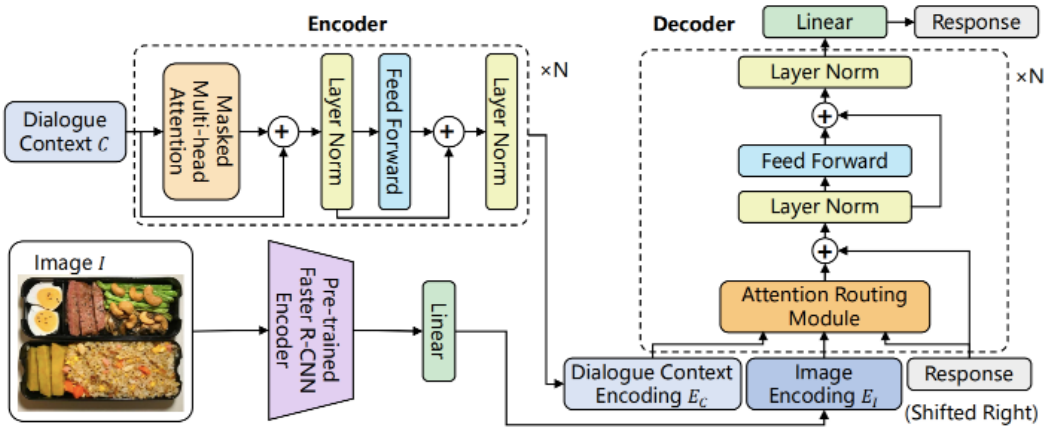


Figure 2: Overview of multi-modal dialogue generation model.

Encoder

text：

文本上下文使用 Transformer 结构来实现参数化（12 层，12 注意力头，768 隐藏层）

为了改善生成质量，我们使用预训练的 GPT 模型来初始化权重。

对话历史被[SEP]分割，将串联的文本序列送入文本编码器中，得到 EC

image：

图像编码器是带有 ResNet-101 主干的 Faster-RCNN 模型。具体来说，每个图片区域提取 2048 维度向量，top-50 置信度最高的区间被用作产生 EI 通过一个线性层来调整特征维度。EI 包含 50 个特征，特征长度是 768.

Decoder

我们使用 Transformer 结构实现对话解码器并且与我们的文本编码器共享权重。

为了解决稀疏性问题，我们部署对话解码器通过注意力路由机制去平衡每一个区域特征的贡献。具体来说，给定对话上下文的编码 EC，图片上下文 EI 以及先前的解码 token Epre，三个注意力被计算为：

$$
\begin{aligned}
O_{\mathcal{C}} &= \mathrm{MHA}\left(E_{\mathrm{pre}}, E_{\mathcal{C}}, E_{\mathcal{C}}\right) \\
O_{\mathcal{I}} &= \mathrm{MHA}\left(E_{\mathrm{pre}}, \gamma E_{\mathcal{I}}, \gamma E_{\mathcal{I}}\right) \\
O_{\mathrm{pre}} &= \mathrm{MMHA}\left(E_{\mathrm{pre}}, E_{\mathrm{pre}}, E_{\mathrm{pre}}\right)
\end{aligned}
$$

MHA 和 MMHA 表示有 mask 和无 mask 的多头注意力。每个注意力的结果操作在进行下一个模块时被平均：

$$
O_{\mathrm{merge}} = \frac{O_{\mathcal{C}} + O_{\mathcal{I}} + O_{\mathrm{pre}}}{3}
$$

Experiments

实验被用来评估我们模型和数据集的性能。具体来说，同时训练模型使用 MMChat 和 MMChat-HF。对于 MMCHAT，我们抽样 4.0K 和 2.0K 个对话场景作为测试集和交叉验证集。对于 MMChat-HF，我们分别抽样 1000 个座位测试集和交叉验证集。

实现细节

在我们提出的对话模型（seq2seq+IMG），编码器和解码器都是 12 层 transformer。我们的 encoder 和 decoder 共享权重，参数初始化是 GPT 模型。

baselines

我们同样实现了两个 baseline 去评估数据集：

1. Seq2Seq：只考虑文本上下文
2. Seq2Seq+PIMG：一个基于图像的对话模型，用单一的图像表示。具体来说，对 EI 进行最大池化操作得到 EI，注意力路径不被使用。

为了公平比较，所有的 baseline 使用相同的架构，超参数和初始化方式一样。

评价指标

BLEU，DIST，Ent

$$Ent = \frac{1}{\sum_w F(w)} \sum_{w \in V} F(w) \log \frac{F(w)}{\sum_w F(w)},$$

Results

Model	BLEU-2	BLEU-3	BLEU-4	Dist-1	Dist-2	Ent-1	Ent-2
Seq2Seq	2.830	1.376	0.805	2.63	33.92	6.00	9.47
Seq2Seq+PIMG	2.928 (+3.46%)	1.469 (+6.76%)	0.888 (+10.31%)	2.73 (+3.80%)	34.34 (+1.24%)	6.01 (+0.17%)	9.45 (-0.21%)
Seq2Seq+IMG	3.001 (+6.04%)	1.588 (+15.41%)	1.006 (+24.97%)	2.82 (+7.22%)	35.38 (+4.30%)	6.07 (+1.17%)	9.52 (+0.53%)
Human Reference	N/A	N/A	N/A	9.09	48.77	6.69	9.64

Table 4: Evaluation Results on MMCHAT. Relative improvements compared to the Seq2Seq baseline is shown in parentheses.

Model	BLEU-2	BLEU-3	BLEU-4	Dist-1	Dist-2	Ent-1	Ent-2
Seq2Seq	3.779	2.405	1.641	5.35	45.62	6.11	9.26
Seq2Seq+PIMG	4.576 (+21.09%)	3.094 (+28.65%)	2.230 (+35.89%)	5.04 (-5.79%)	42.61 (-6.60%)	6.01 (-1.64%)	9.14 (-0.21%)
Seq2Seq+IMG	4.818 (+27.49%)	3.381 (+40.58%)	2.541 (+54.84%)	5.75 (+7.48%)	45.35 (-0.59%)	6.05 (-0.98%)	9.15 (+1.19%)
Human Reference	N/A	N/A	N/A	6.17	47.43	5.98	9.00

Table 5: Evaluation Results on MMCHAT-HF. Relative improvements compared to the Seq2Seq baseline is shown in parentheses.