# Emoji2movie 数据集说明书

## 摘要（说明本数据集做了什么）

本数据集构建了50条根据emoji表情猜电影名称的测试数据。例如，给出emoji表情🐰🦊🚔🏙，可以猜到这串emoji表情对应的电影是《疯狂动物城》。通过这个任务，可以评价语言模型对于emoji表情符号的理解能力、对电影知识的储备以及隐式推理的能力。

## 动机（为什么要构建本数据集）

表情符号，是1997年引入的用于信息化交流中的象形符号和笑脸符号的集合，已经深深地融入到我们的日常交流中。到2015年，超过10%的推文和超过35%的Instagram帖子包含一个或多个表情符号[1]，[2]。随着emoji表情符号使用越来越频繁，越来越多的研究正在研究它们在语言和文化方面的影响（例如[3]）；以及我们如何利用emoji表情符号来帮助解决自然语言处理（NLP）任务（例如[4]）。

随着预训练语言模型的发展，最近的大语言模型在理解和生成自然语言方面取得了巨大的成功。由于emoji符号在今天的数字世界中无处不在，评估这些语言模型理解emoji符号的能力是至关重要的。此外，研究发现，一般来说，不同文化和语言之间的emoji符号使用规范和习惯是相似的[3]。因此，衡量语言模型对emoji符号的理解对于AI技术发展具有关键的意义。

## 数据集构建方法（如何构建本数据集）

我们通过人工编写的方式，构造了50条数据集。首先，我们从imdb中收集了一些热门电影的名称，然后我们找了一些电影爱好者，根绝电影剧情编写其最合适的emoji表情。对于一些已经在网上有表情符号描述的电影，我们特意为编写了不同的表情符号表达。例如电影《断背山》在网上已有一些表情符号描述，如🧔🏻‍♂️❤️🧔🏻‍♂️🗻和🧔🏻‍♂️🧔🏻‍♂️⛺️🗻，在本数据集中我们将其重写为🤠❤️🤠。

Emoji符号的编写可以从多个角度思考，常见的是电影标题和电影剧情，在本数据集构造时，我们只关注于电影剧情而非电影标题。例如，电影《海底总动员(finding nemo)》可以根据其标题描述为🔍🐠。然而，按照其剧情，我们将其写为👧🐟🐠🐡。这种编码要求语言模型有电影剧情知识的储备，而不是将简单的将电影标题中的单词映射到emoji符号。

此外，最近的许多emoji符号会使用unicode修饰符来引入性别和肤色的变体。例如，unicode U+1F44B代表👋（挥手），加上肤色修饰符U+1F44B U+1F3FD，它变成了👋🏽（中等肤色的挥手）。为了使本数据集更具普遍性并且字符数量更少，我们优先选择没有修饰符的中性表情符号。例如，电影《超人总动员》也可以写成🦸🏻‍♀️🦸🏼‍♂️👦🏼👧🏻👶🏻（加了肤色和性别修饰），但我们会优先使用🦸🦸‍♂️👦👧👶（没有使用肤色和性别修饰）。

最后，由于emoji符号对应的电影名称可能并不唯一，为了可以使用精确匹配的方式进行评测，我们将数据集设计成单项选择题的形式，4个候选电影中只有一项是和emoji符号匹配的。

## 数据集评测方式（如何计算模型表现的指标）

判断模型预测结果是否和答案完全精确匹配，计算准确率。

## 数据集格式说明

注：数据集提供excel/json其中一种格式即可，下面分别给出两种格式的说明

Excel版本：

 本数据集为Excel格式，每条数据为一行，共包含7列，第一列是问题，第二列到第六列为五个候选项，第七列是答案。如下所示：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 问题 | A | B | C | D | E | 答案 |
| What movie does this emoji describe? 👧🐟🐠🐡 | Finding nemo | The wolf of wall street | se7en | the shining | mr. smith goes to washington | A |
| What movie does this emoji describe? 🐀👨‍🍳👩‍🍳 | zootopia | a clockwork orange | e.t. the extra-terrestrial | mad max: fury road | ratatouille | E |

Json版本：

本数据集为Json格式，每条数据包括以下字段：

* input: 一个关于emoji符号对应哪部电影的问题
* target\_scores:包括几个不同选项，每个选项是一部电影的名称，其中只有一个选项是和input中的emoji符号是正确的映射,错误的候选项后面填0，正确的选项后面填1
* target:emoji符号对应的电影名称

具体样例如下：

"examples": [

 {

 "input": "What movie does this emoji describe? 👧🐟🐠🐡",

 "target\_scores": {

 "finding nemo": 1,

 "the wolf of wall street": 0,

 "se7en": 0,

 "the shining": 0,1

 "mr. smith goes to washington": 0

 },

 "target": "finding nemo"

 },

 {

 "input": "What movie does this emoji describe? 🐀👨‍🍳👩‍🍳",

 "target\_scores": {

 "mad max: fury road ": 0,

 "e.t. the extra-terrestrial": 0,

 "zootopia": 0,

 "a clockwork orange": 0,

 "ratatouille": 1

 },

 "target": "ratatouille"

}

]

## 局限性和未来工作

电影剧情和它们的emoji描述之间并没有一一对应的映射：可以使用相同emoji表情的不同变体，或者用不同的emoji表情来描述同一部电影。未来可以适当对当前任务进行扩展，例如引入emoji表情的变体，以进一步评估语言模型对emoji的理解。

我们希望这个任务可以成为探索语言模型解读表情符号能力的第一步。未来的研究可以探索其他的探测方法；例如，可以为流行的游戏、音乐等设计类似的表情符号测试。

## 参考文献

[[1]](https://blog.twitter.com/en_us/a/2015/emoji-usage-in-tv-conversation.html) Cruse, J. (2015) Emoji usage in TV conversation. Twitter blog.

[[2]](https://instagram-engineering.com/emojineering-part-1-machine-learning-for-emoji-trendsmachine-learning-for-emoji-trends-7f5f9cb979ad) Instagram Engineering. (2015) Emojineering part 1: Machine learning for emoji trends. Medium.

[[3]](https://ojs.aaai.org/index.php/ICWSM/article/view/3224) Chandra Guntuku, S., et al. (2019) Studying cultural differences in emoji usage across the east and the west. Proceedings of the International AAAI Conference on Web and Social Media **13**(10):226–235.

[[4]](https://aclanthology.org/W16-6208/) Eisner, B., et al. (2016) Emoji2vec: Learning emoji representations from their description. In Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media, pp. 48–54. Stroudsburg, PA: Association for Computational Linguistics.

[[5]](https://andrewmayneblog.wordpress.com/2020/06/24/open-ai-alchemy-emoji-storytelling) Mayne, A. (2020) OpenAI API Alchemy: Emoji storytelling. Andrew Mayne blog.

[[6]](https://www.fluate.net/en/travaux/vectoglyph) Boillot, N. (2019) Vector forms as a foreign language.