中国科学院大学
University of Chinese Academy of Sciences

基于场景理解的沙盒主题多标签识别

第25组 陈若愚 杨嘉瑞 唐源民 涂皓钦 刘康威 颜广

2022年5月24日

中国科学院大学
University of Chinese Academy of Sciences

目录

CONTENTS

问题分析

- 本任务是一个基于场景理解的有监督多标签分类问题
  - 要求模型把沙盘游戏图像分到八种类别中的一个或者多个类别
  - 模型需要理解不同沙盘中的object排列的场景语义信息

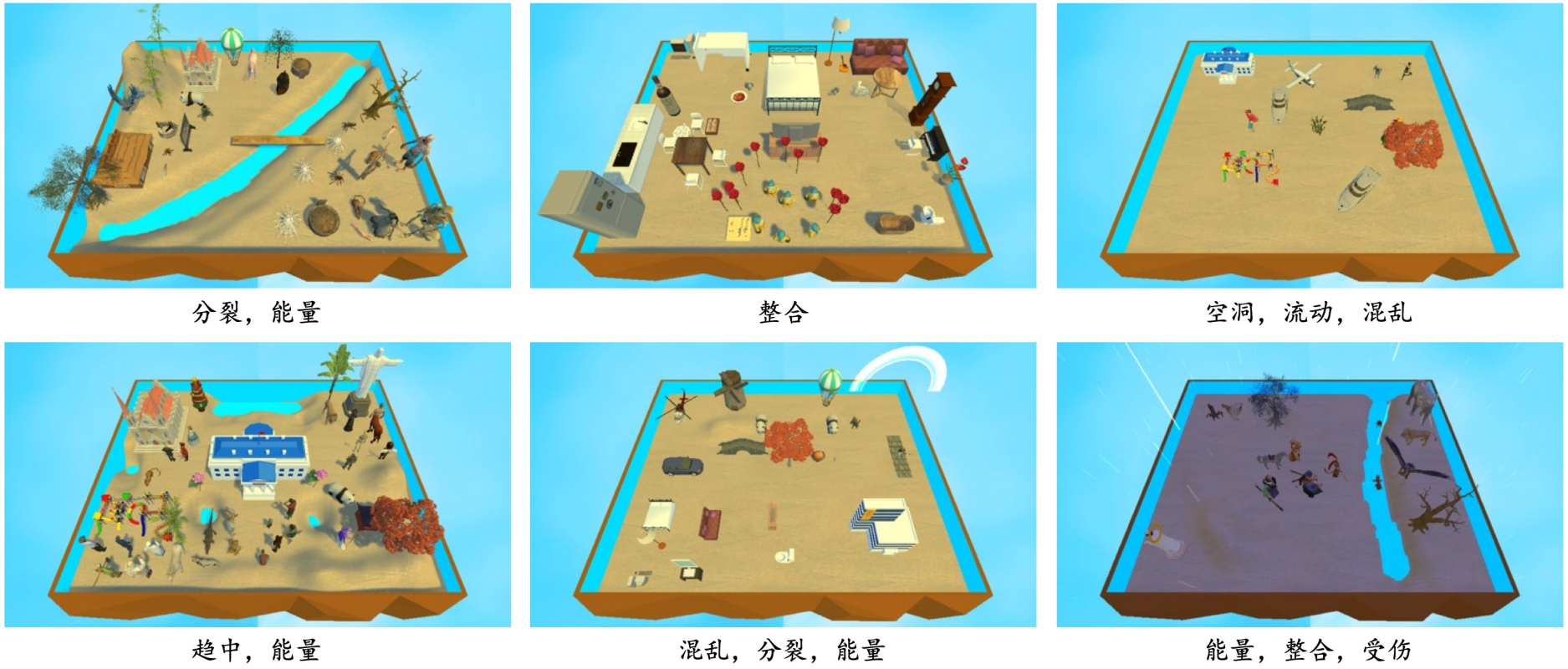

图1. 沙盒游戏图像以及其对应的主题

模型设计与验证方案

设计的模型需要什么？

- 需要解码复杂的沙盘场景语义信息 – 解码场景
- 需要依据解码出的场景信息正确进行多标签分类 – 理解场景
- 需要有效的学习方式 – 少量监督信息

模型设计方案

- 采用预训练模型提取沙盘场景特征+分类的End-to-End架构 - 解码并理解场景
- 采取

  （1）模型+从头训练+对比学习增强样本

  （2）预训练模型+微调

两种模型设计方案，并对各个方案进行实验验证

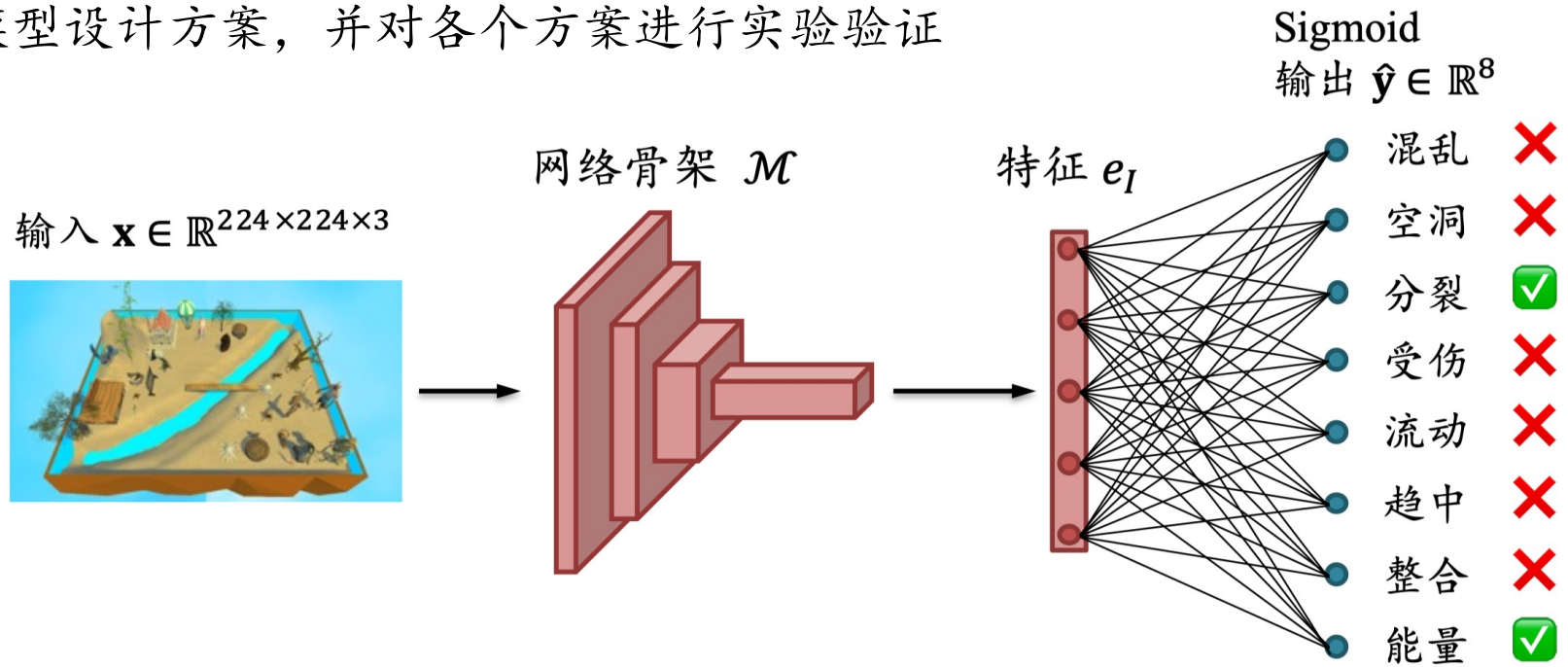

图2. 模型基本架构设计

模型设计与验证方案

为了验证模型

- 139张图像的数据集中抽取了123张图像作为训练集，其余图像作为测试集
- 同时，如图3所示，我们需要考虑训练集样本分布不均衡问题

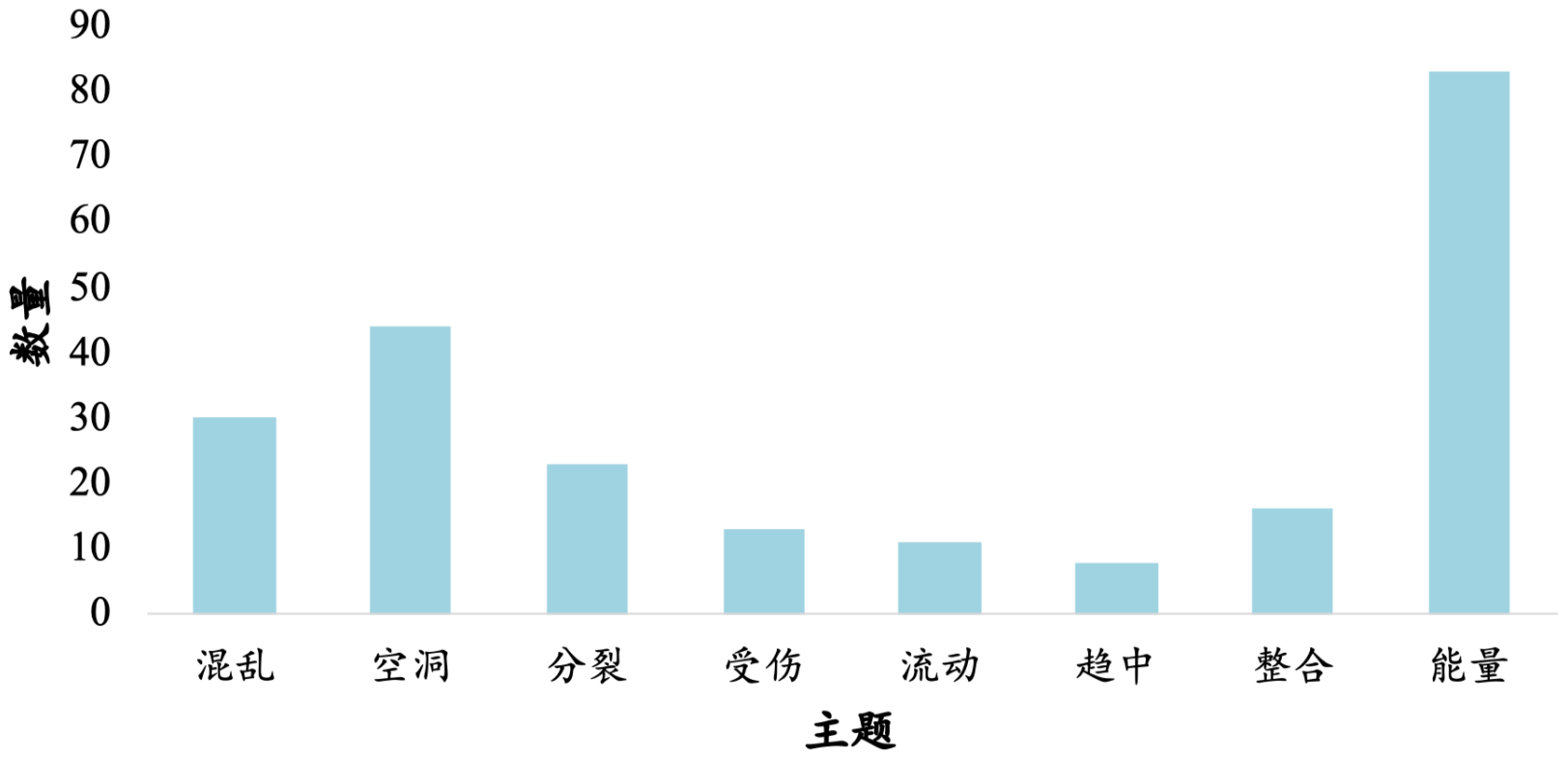

图3. 训练集数据分析

模型损失函数设计

考虑正负样本不均衡，设计如公式(1)所示的加权的二分类交叉熵损失函数

在二分类交叉熵损失函数增加权重系数 p

- 用于平衡正负样本不均衡的情况
- 每个类别的p为表1中**正例/反例**

$$\mathcal{L} = -[\mathbf{p} \cdot \mathbf{y} \cdot \log \hat{\mathbf{y}} + (1 - \mathbf{y}) \cdot \log (1 - \hat{\mathbf{y}})] \quad (1)$$

标签	0	1	2	3	4	5	6	7
主题	混乱	空洞	分裂	受伤	流动	趋中	整合	能量
数量	30	44	23	13	11	8	16	83
正例/反例	3.10	1.80	4.35	8.46	10.18	14.38	6.69	0.48

表1. 不同类别的权值

模型损失函数设计

- 考虑到有监督数据量过少的问题
- 设计如公式 (2) 所示的基于对比学习的损失函数
  - 集合P, A – 正，负样本集
  - z_p, z_a – 正负样本经过解码器的特征
  - τ – 模型训练的超参数

$$\mathcal{L}_{\text{out}}^{sup} = \sum_{i \in I} \mathcal{L}_{\text{out},i}^{sup} = \sum_{i \in I} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau\right)}{\sum_{a \in A(i)} \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_a / \tau\right)} \tag{2}$$

- 最终基于对比学习模型损失函数如公式 (3) 所示

$$\mathcal{L}_{all} = \mathcal{L} + \beta \mathcal{L}_{\text{out}}^{sup} . \tag{3}$$

模型解码器优化 - 提升模型对复杂场景语义的理解能力

- 考虑到沙盒图像具有复杂场景语义信息
  - 采取CLIP，一种视觉-语言预训练模型 - 更好地提取场景语义信息

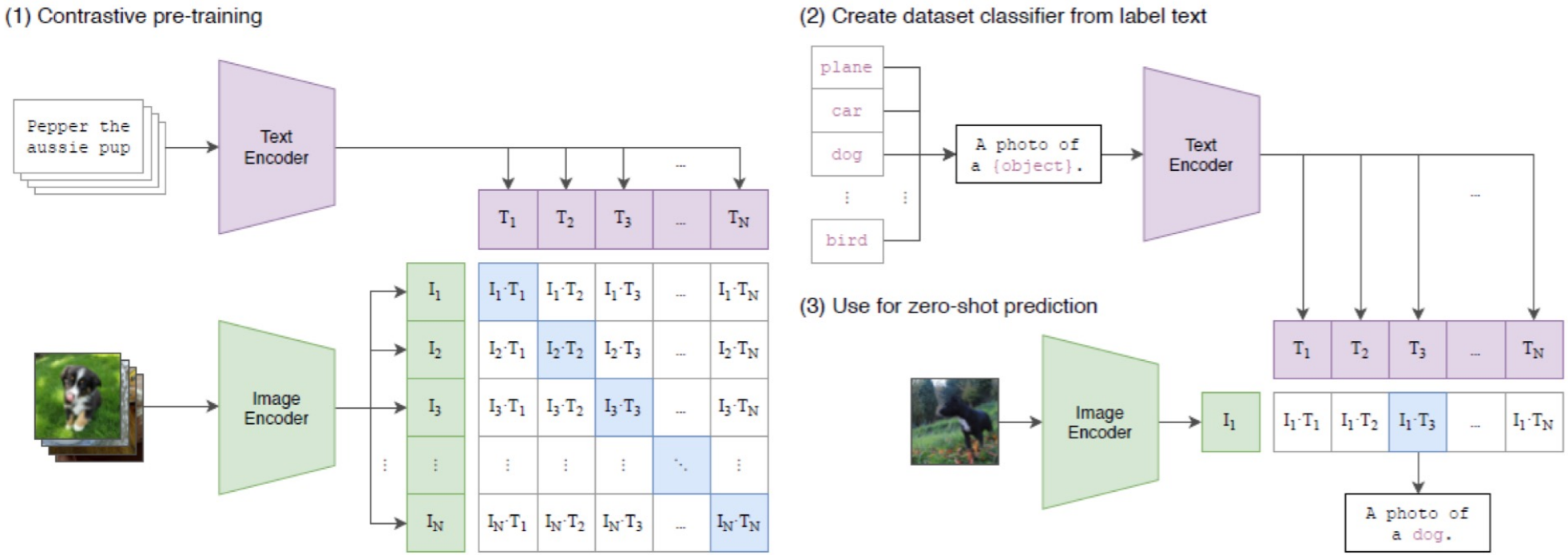

图4. CLIP模型架构图

实验评价指标

- 将多标签分类看作一个二分类问题
- 结果数值加入阈值 $T = 0.5$，当输出值大于阈值判断存在该主题，否则不存在。
- 我们采纳的评价指标包括 $TP, TN, FP, FN, Precise, Recall, Accuracy$。

$$\text{Precise} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

1.1 模型+从头训练 - ResNet50最优

ResNet18

Index	主题	TP	FN	FP	TN	Precision	Recall	ACC
0	混乱	4	0	12	0	0.25	1	0.25
1	空洞	0	6	0	10	0	0	0.625
2	分裂	0	4	0	12	0	0	0.75
3	受伤	0	1	0	15	0	0	0.9375
4	流动	0	1	0	15	0	0	0.9375
5	趋中	0	1	0	15	0	0	0.9375
6	整合	1	0	15	0	0.0625	1	0.0625
7	能量	0	11	0	5	0	0	0.3125

ResNet50

Index	主题	TP	FN	FP	TN	Precision	Recall	ACC
0	混乱	4	0	10	2	0.2857	1	0.375
1	空洞	6	0	2	8	0.75	1	0.875
2	分裂	4	0	10	2	0.2857	1	0.375
3	受伤	0	1	5	10	0	0	0.625
4	流动	1	0	5	10	0.1667	1	0.6875
5	趋中	1	0	0	15	1	1	1
6	整合	0	1	0	15	0	0	0.9375
7	能量	3	8	1	4	0.75	0.2727	0.4375

ResNet101

Index	主题	TP	FN	FP	TN	Precision	Recall	ACC
0	混乱	4	0	11	1	0.2667	1	0.3125
1	空洞	5	1	1	9	0.8333	0.8333	0.875
2	分裂	4	0	11	1	0.2667	1	0.3125
3	受伤	1	0	14	1	0.0667	1	0.125
4	流动	0	1	0	15	0	0	0.9375
5	趋中	0	1	0	15	0	0	0.9375
6	整合	0	1	0	15	0	0	0.9375
7	能量	1	10	0	5	1	0.0909	0.375

VGG16

Index	主题	TP	FN	FP	TN	Precision	Recall	ACC
0	混乱	0	4	0	12	0	0	0.75
1	空洞	0	6	0	10	0	0	0.625
2	分裂	0	4	0	12	0	0	0.75
3	受伤	0	1	0	15	0	0	0.9375
4	流动	0	1	0	15	0	0	0.9375
5	趋中	1	0	15	0	0.0625	1	0.0625
6	整合	0	1	0	15	0	0	0.9375
7	能量	11	0	5	0	0.6875	1	0.6875

表2. 小参数模型+从头训练学习方案实验结果

1.2 模型+从头训练 - 加权系数p消融(无系数p训练不出有效模型)

Index	主题	TP	FN	FP	TN	Precision	Recall	ACC
0	混乱	0	4	0	12	0	0	0.75
1	空洞	0	6	0	10	0	0	0.625
2	分裂	0	4	0	12	0	0	0.75
3	受伤	0	1	0	15	0	0	0.9375
4	流动	0	1	0	15	0	0	0.9375
5	趋中	0	1	0	15	0	0	0.9375
6	整合	0	1	0	15	0	0	0.9375
7	能量	11	0	5	0	0.6875	1	0.6875

表3. ResNet50 损失函数没有加权系数p的实验结果

1.3 小参数模型+从头训练+对比学习增强 -（能有效增强模型整体性能）

Avg. Acc. 结果为不同主题准确率平均，每个实验在5个不同随机种子设定下运行取最终平均值。

β	0.0	0.1	0.3	0.5	0.8	1.0
Avg. Acc.	0.5109	0.6828	**0.7359**	0.7172	0.7172	0.6828

表4. ResNet50 损失函数在不同对比损失加权系数β 下的实验结果。

2.1 大参数预训练模型+微调 - Resnet101 训练全部参数

- 计算数据集的图片三通道的均值和标准差，对图片进行预处理
- 采用十折交叉验证，对每一个fold在相同的epoch下取平均

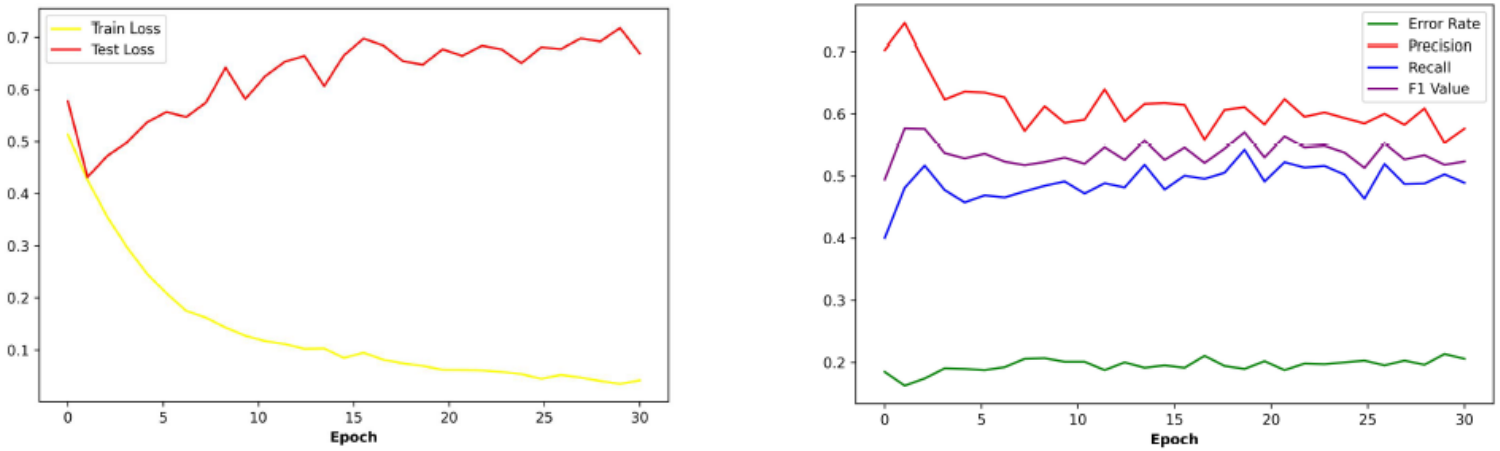

图5. 基于Resnet101的预训练模型+微调-训练全部参数的实验结果

2.2 大参数预训练模型+微调 - Resnet101 训练分类器参数

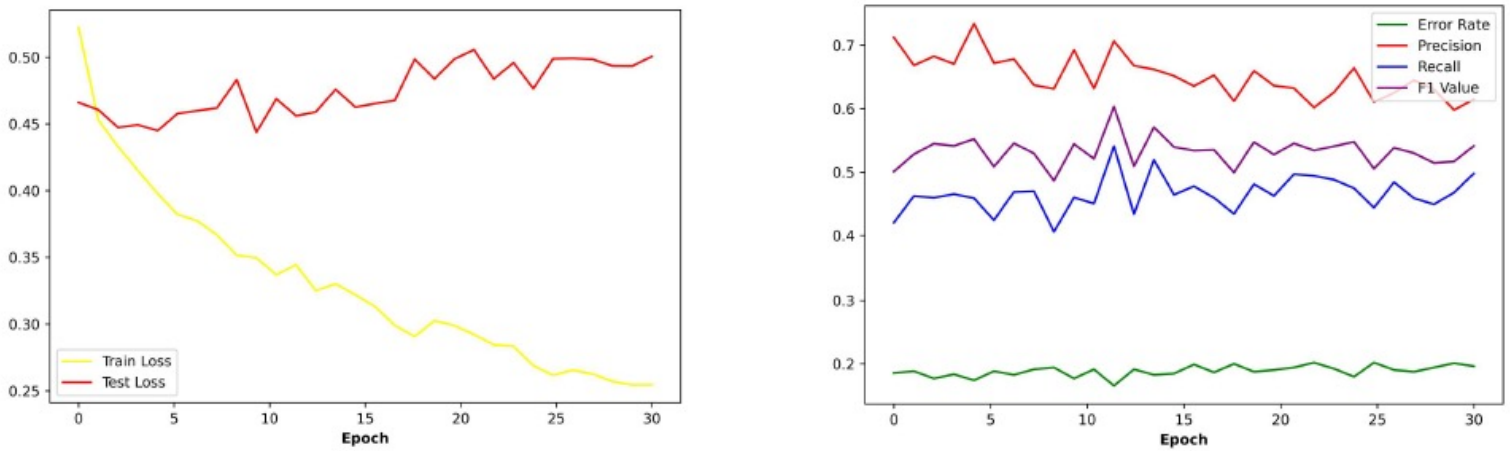

图5. 基于Resnet101的预训练模型+微调-训练分类器参数的实验结果

结论：由于resnet101 模型过于复杂，因此网络**过拟合**的现象十分严重；
固定了特征提取部分的参数可以从一定程度上减轻过拟合；

2.3 大参数预训练模型+微调 -Vision Transformer 训练分类器参数

Index	主题	TP	FN	FP	TN	Precision	Recall	ACC
0	混乱	4	0	5	7	0.4444	1	0.6875
1	空洞	6	0	0	10	1	1	1
2	分裂	3	1	4	8	0.4286	0.75	0.75
3	受伤	0	1	1	14	0	0	0.6875
4	流动	1	0	1	14	0.5000	1	0.8790
5	趋中	0	1	1	14	0	0	0.9375
6	整合	0	1	2	13	0	0	0.8125
7	能量	7	4	2	3	0.7778	0.6364	0.6250

表4. 基于CLIP的Vision Transformer+训练分类器参数的实验结果

结论：与纯视觉模型相比，多模态预训练模型能更好地解码具有复杂场景信息的沙盒图像；

结论

- 我们针对沙盒图像提出了一个基于场景理解的沙盒主题多标签分类模型。
- 针对有限监督的**正负样本分布不均衡**设计了有效的学习方法。
- 实验结果证明了模型的有效性。
- 我们的分析实验表明
  - 深度网络具有对复杂场景理解的能力。
  - 相对于纯视觉模型，多模态模型能够更好地理解复杂场景的语义信息。

小组分工

陈若愚：组长，组织安排
多标签模型与加权损失函数，论文初稿。

唐源民：PPT第一版制作，多模态预训练模型设想与实验。

杨嘉瑞：预训练模型微调实验。

刘康威：PPT微调与润色，实验结果与内容核对。

涂皓钦：对比学习消融实验。

颜　广：论文终稿撰写。

感谢各位聆听

请老师批评指正

第25组 陈若愚 杨嘉瑞 唐源民 涂皓钦 刘康威 颜广

2022年5月24日