基于视觉理解的沙盒主题多标签识别

作者：陈若愚*，唐源民，杨嘉瑞，涂皓钦，刘康威，颜广
导师：黄凯奇，陈晓棠，赵鑫，操晓春，于静，王川，赵险峰，侯锐

2022 年 6 月 24 日

Abstract

随着人工智能与深度学习技术的发展，计算机视觉在识别领域取得了突破性的进展，在 ImageNet[1] 等数据集上的识别准确率大幅度提升，甚至超越了人类的识别能力。然而，在视觉系统对图像的理解方面距离人类对图像的理解还有非常大的探索空间。沙盒游戏，是由一个或多个地图区域构成，往往包含多种游戏要素，包括角色扮演，动作、射击、驾驶等等。能够改变或影响甚至创造世界是沙盒游戏的特点。沙盒游戏的各种主题与要素往往反应着多种视觉感知上的情感，例如受伤，分裂等等。沙盒游戏能够非常好的帮助人们研究视觉理解问题 [2]。针对视觉理解的问题，本文以沙盒游戏的主题识别为研究对象，探索现有卷积神经网络方法对沙盒主题识别的性能。我们的数据集包括 139 张沙盒图像，每个图像对应有 8 个主题标签——混乱、空洞、分裂、受伤、流动、趋中、整合和能量。我们基于卷积神经网络 VGGNet[3] 与 ResNet[4]，构建了一个多标签分类模型，以实现对沙盒图像的主题识别。由于沙盒图像与自然图像存在较大的域差异，因此本文也讨论了基于沙盒图像内容与自然图像理解内容的差异。针对沙盒图像，我们从零训练了识别模型。我们也使用了由自然图像训练的预训练模型，对特征提取器冻结，然后在训练沙盒图像中微调网络，从而探讨沙盒图像与自然图像语义理解的差异。实验结果表明，尽管在很少的图像用于训练，我们的模型依然能较好的识别出沙盒图像的主题。

1 数据集描述与分析

1.1 数据主观分析

我们的实验数据基于 139 张沙盒图像，每一张图像被标注归属于单个或者多个主题。图1展示了部分数据集中的图像以及其对应的主题标签。主题总共有 8 种，包括**混乱**、**空洞**、**分裂**、**受伤**、**流动**、**趋中**、**整合**和**能量**。

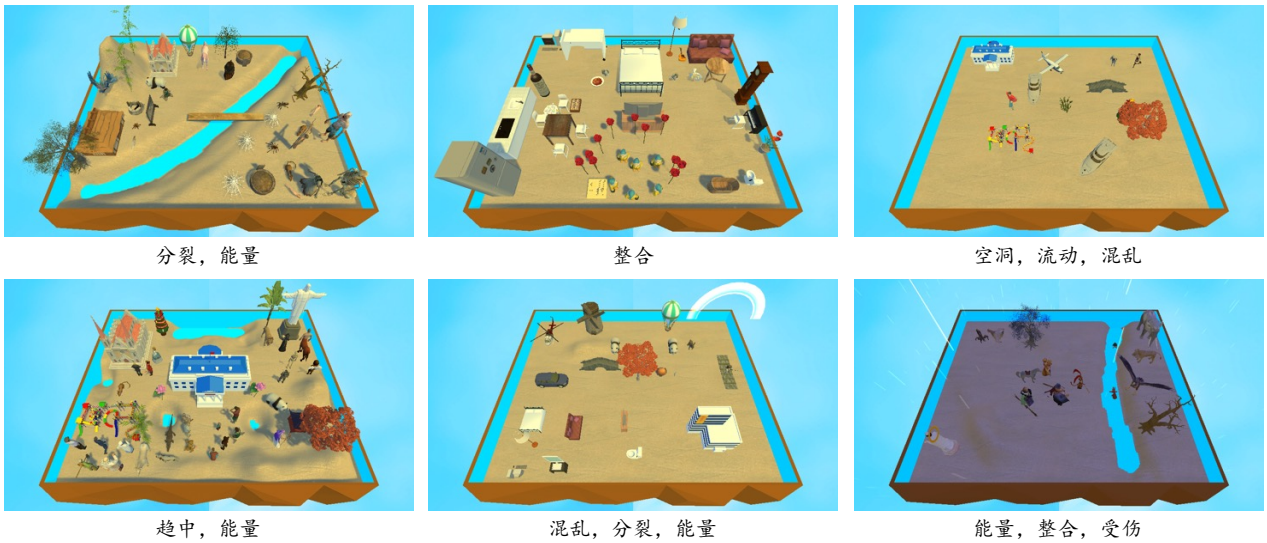


Figure 1: 沙盒游戏图像以及其对应的主题。

*组长

我们观察到，同一张图像可能会属于多种主题，并且多数主题之间并不存在互斥现象。例如图1第 1 行最后一张图像同时包括了空洞与混乱两种看上去关联性很低的主题。因此我们认为该任务为一个多标签分类问题。即给定一张图像，判断图像存在哪些主题。在这些图像中我们发现，沙盒在图像中的位置总是固定不变的，沙盒中存在各种各样的物体，因此我们不建议进行一些图像旋转或者图像平移的图像增强操作，因为经过旋转或平移操作的图像将会是一种 Out of Distribution 图像。此外，绝大多数沙盒的地表是金黄色的沙滩，但是也存在较为阴暗的表面，如图1第 2 行最后一张沙河图像。因此，将沙盒的背景去除也是不合理的，依然需要考虑沙盘的表面情况。

1.2 训练数据标签分布分析

为了能够更好的评价模型，在 139 张图像的数据集中，我们抽取了 123 张图像作为训练集 $\mathcal{D}_{\text{train}}$，另外的 16 张图像作为测试集 $\mathcal{D}_{\text{test}}$。表1展示了训练集中包含的标签以及每个标签所包含的图像的数量。如图2所示，训练集中每个主题正例样本数量的分布是严重不均衡的，例如包括受伤，流动，趋中的主题的图像较多，而不包括能量主题的图像较小。这些数据分布情况表明我们在训练模型时必须要考虑到模型数据的分布情况。

Table 1: 训练集 $\mathcal{D}_{\text{train}}$ 的标签，以及每个标签包含的图像的数量。

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

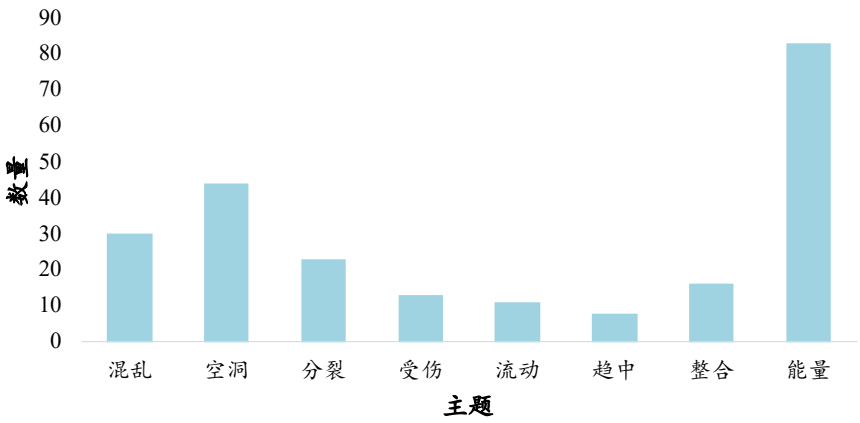


Figure 2: 训练集中每个主题正例样本数量的分布。

2 多标签分类模型

在深度学习中，常常需要一个前向的神经网络，例如 VGGNet[3] 或者 ResNet[4]。同时，通过反向传播 [5] 方法，反向传播的主要前提便是损失函数的设计。因此本章节分为三个小节，在小节2.1中我们描述我们的网络结构，在小节2.2中我们描述我们所设计的损失函数。在小节2.3中我们引入了对比学习分支来优化多标签分类模型。我们的模型都是从零训练，探索模型对沙盒图像的理解。

2.1 多标签分类网络模型

给定训练集图像，我们将图像缩放到 224 尺度 $\mathbf{x} \in \mathbb{R}^{224\times224\times3}$，其对应的标签 $\mathbf{y} \in \mathbb{R}^{8}$，标签在各个维度 t 取值 $\mathbf{y}^t = \{0, 1\}$，代表是否属于第 t 个主题。图3展示了我们的模型的基本结构。对于给定的输入图像 $\mathbf{x}$，我们将其缩放到 224×224，并对每一个通道归一化均值与方差到 0.5。首先我们将预处理后的图像输入到一个骨架网络 $\mathcal{M}$ 以提取特征 $e_I = \mathcal{M}(\mathbf{x})$，这个骨架网络 $\mathcal{M}$ 可

以是 VGGNet 或 ResNet 去掉全连接层后的模型。之后，我们通过一层全连接层 $f(\cdot)$ 将特征 e_I 映射到 8 个输出的维度，并用 σ（Sigmoid）函数激活得到网络的输出 $\hat{\mathbf{y}} = \sigma(f(e_I))$。

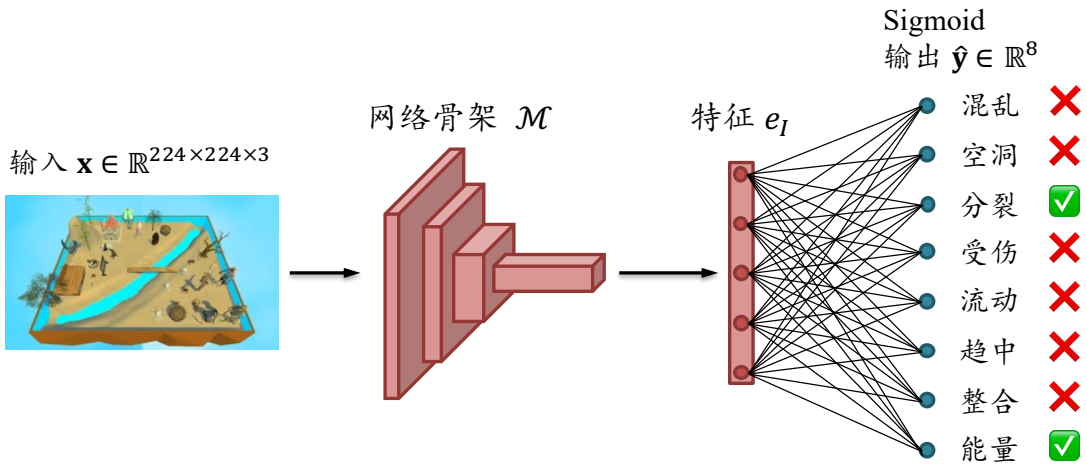


Figure 3: 本文提出的多标签分类模型架构。

2.2 损失函数设计

给定图像 $\mathbf{x}$ 经过模型的输出结果 $\hat{\mathbf{y}}$ 与对应的标签 $\mathbf{y}$，假设第 t 个主题的输出和标签分别为 $\hat{\mathbf{y}}^t$ 和 $\mathbf{y}^t$。我们考虑到二分类的样本存在严重的不均衡性，因此在二分类交叉熵损失函数的基础上增加权重系数 $\mathbf{p}$。如表1的**正例/反例**的数值，我们将此作为权重系数，加入到正样本激励的二分类交叉熵损失函数中，即 $\mathbf{p} = [3.10, 1.80, 4.35, 8.46, 10.18, 14.38, 6.69, 0.48]$。这里，我们定义一个加权的二分类损失函数

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

这里我们要声明，$\mathbf{p}$ 对于网络的训练至关重要，我们将在实验讨论中描述。

2.3 模型增强：对比学习的引入

在上述模型的基础上，我们在训练阶段针对输入图像的标签加入了有监督的对比损失以增强模型的分类性能。我们将首先简单介绍对比学习技术的主要思想，接着具体讲解在上述模型的基础上使用对比学习的步骤和细节。

无监督对比学习的思想与聚类和自编码器等无监督技术的思想类似，即对于给定的样本 A，其目标是找到与之相似的正例 B 以及与之相反的负例 C，通过距离学习最小化 $d(A,B)$ 同时最大化 $d(A,C)$，其中函数 $d(\cdot)$ 为自定义的距离函数。然而一方面在无监督的条件下，如何取得正例以及负例是一项有待探究的话题。另一方面，无监督对比学习是为了解决在没有标签的情况下学习事物的表征，但是其得到的表征依然是为了做一些有监督的任务。那么在我们拥有数据标签时，根据数据标签来分别精确地识别正例和负例进行对比损失的计算能够更大发挥数据和模型的性能 [6]。具体来说，有监督的对比学习可以用以下公式进行计算：

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

其中对于每个样本 $i \in I$，$P(i), |P(i)|$ 分别代表其正样本的集合以及正样本集合的个数。相应地，$A(i), |A(i)|$ 分别代表其负样本的集合以及负样本集合的个数。z_i, z_p, z_a 分别代表样本 i, p, a 的模型表征，这里我们选取 ResNet50 经过 sigmoid 函数前的输出作为对比表征空间。参数 τ 为训练超参数。由于我们知道图片标签，因此正，负样本的构造由给定的标签决定，即相同标签的样本属于相同正例，否则为负例。最终模型的损失为：

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

其中 β 为控制对比学习损失的权重。

3 预训练模型微调

对于数据集较小且存在类别不均衡的分类任务，从头开始训练整个网络可能并不合适。一方面，相较于那些优异的模型，往往它们训练的数据集足够大 (例如 ImageNet 数据集有 1.2M 图像)，所以泛化性能通常较好，而我们的数据集数据相对较少，网络学习可能不够充分；另一方面，在我们的数据集中还存在着类别不均衡的问题，这也会对网络的性能带来一些负面影响。为了解决这类问题，除了 Section 2中介绍的方案之外，另外一种有效的解决方案是使用风格迁移的方式进行域自适应。同时我们可以探索自然图像理解与沙盒图像理解的差异。在本章节，我们选择了两种预训练模型在我们的数据集上进行微调，分别是基于 ConvNet 架构的纯视觉模型和基于 Vision-Transformer[7] 架构的多模态视觉语言模型。

3.1 基于 ConvNet 架构的纯视觉模型

我们选择了在 ImageNet[1] 上预训练的 ResNet101 模型，并以两种方式进行微调，用以说明不同微调方式对模型性能的影响。

3.2 基于 Vision-Transformer 架构的多模态视觉语言模型

纯视觉模型虽然在一些物体识别任务上表现较好，但对于具有丰富且复杂的场景语义信息的沙盘图像，纯视觉模型由于缺乏对场景语义的理解而不能很好地解码沙盘图像。因而，为了优化我们的解码器，即能使我们的模型能更好地对解码复杂场景语义信息，我们采用一种基于 Vision-Transformer 架构的多模态视觉语言模型 CLIP[8] 作为我们的预训练模型。CLIP 利用了大量自然语言与图像的配对来实现泛化和迁移，因而能够更好地理解复杂场景语义信息，因而在沙盘主题识别任务上也有着不错的性能。

4 实验分析

4.1 实现细节

网络的骨架 $\mathcal{M}$ 我们采用了 VGG16，ResNet18，ResNet50，ResNet101。网络的优化器我们采用了 SGD 优化器，学习率设置为 0.01。并且采用了指数衰减学习率，衰减因子 $\gamma = 0.1$，每 40 次迭代衰减一次学习率。我们总共训练 100 轮 epoch 并保存最后的模型。我们的实验是在一块 RTX 3090 显卡上进行的。对于加入对比学习的增强模型，我们基于 ResNet50 进行实现，并在不同的对比学习权重下进行 5 折验证实验，最后的结果取每轮训练测试结果最佳准确率的平均值。

4.2 评价指标

由于是二分类问题，我们对输出结果 $\hat{\mathbf{y}}$ 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。其中

$$\text{Precise} = \frac{TP}{TP + FP} \tag{4}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

4.3 实验结果

ResNet 我们在 ResNet18，ResNet50 与 ResNet101 上训练模型并输出结果。ResNet18 的结果如表2所示，ResNet50 的结果如表3所示，ResNet101 的结果如表4所示。我们发现 ResNet50 取得了最好的结果。

Table 2: ResNet18 实验结果。

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

Table 3: ResNet50 实验结果。

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

Table 4: ResNet101 实验结果。

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

VGGNet 我们在 VGG16 上训练了一个模型，结果如表5所示。我们发现，VGG16 并不能成功将模型正确划分主题。

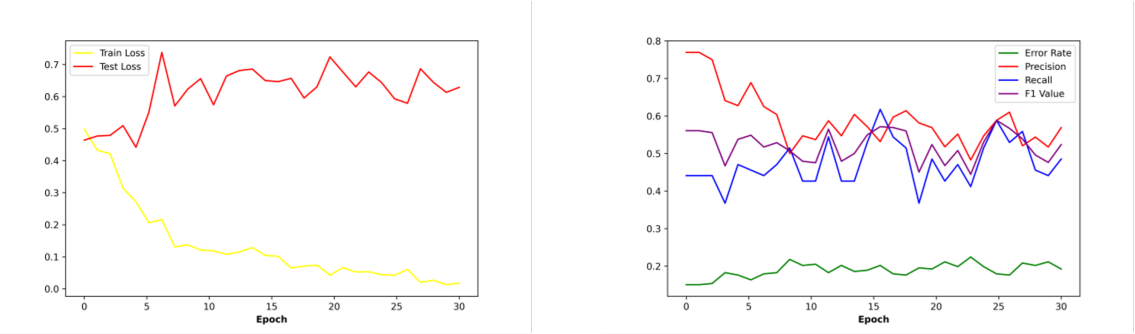
Table 5: VGG16 实验结果。

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

4.4 基于纯视觉预训练模型的沙盒图像场景理解实验结果

对于预训练好的 ResNet101 模型，我们采用了两种不同的微调方式，分别为：(1) 训练 ResNet101 的全部参数；(2) 只训练 ResNet101 最后一层的参数 (即分类器)，固定特征提取部分的参数

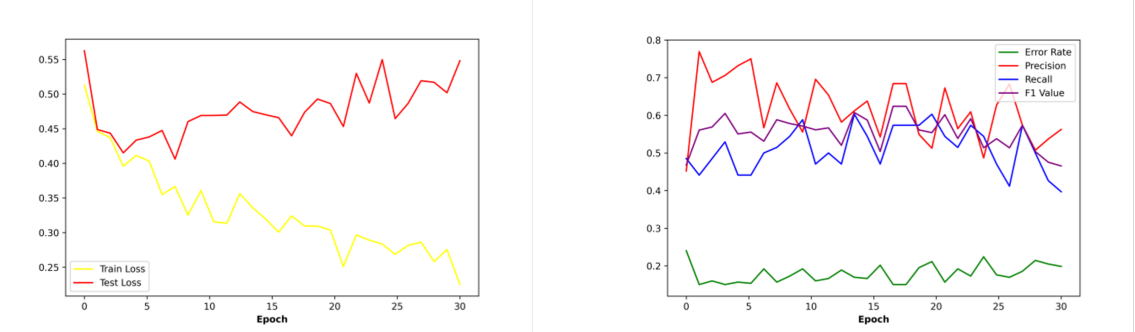
训练全部参数 由于 ResNet101 网络结构较为复杂，在样本量稀少的前提下，训练全部参数非常容易过拟合，在性能上更加的不稳定，图4给出了这种微调方式下的损失曲线和评价指标曲线。



(a) 损失曲线 (b) 评价指标曲线

Figure 4: 训练整个网络的损失曲线和评价指标曲线

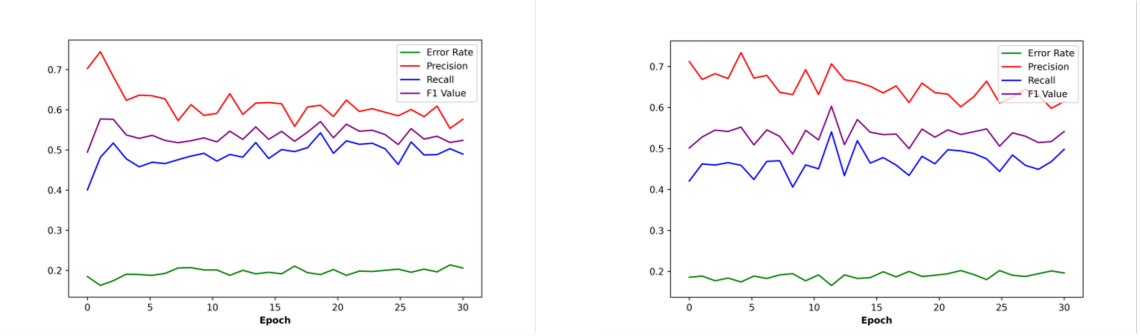
训练分类器参数 固定特征提取部分的参数，只训练分类器参数，可以在一定程度上抑制的过拟合的现象，性能和损失受过拟合的影响减小，但在 20 个 epoch 之后，过拟合现象仍然在逐渐加剧，在性能上有小幅度的提升。图5给出了这种微调方式下的损失曲线和评价指标曲线。



(a) 损失曲线 (b) 评价指标曲线

Figure 5: 训练分类器的损失曲线和评价指标曲线

交叉验证实验结果 为了进一步说明我们的实验结果，我们采用十折交叉验证，对每一个 fold 在相同的 epoch 下的各项指标取平均值。图 6是交叉验证下两种微调方式的评价指标曲线。交叉验证的结果依然说明训练全部参数的方法会使得模型更容易过拟合，网络的性能最差。由于 resnet101 模型过于的复杂，该实验的数据集规模过小，会使得模型泛化性能很小，因而不能进行准确的预测。通过固定了特征提取部分的参数，只训练分类器层，使得每张图片的特征并不受训练的影响，因此可以缓解过拟合的问题。



(a) 训练整个网络 (b) 训练分类器

Figure 6: 两种不同微调方式交叉验证结果的评价指标曲线

4.5 基于多模态预训练模型的沙盒图像场景理解实验结果

我们上述实验发现，沙盘图像具有丰富，且复杂的场景语义信息，而纯视觉模型由于缺乏对场景语义的理解而不能很好地解码沙盘图像。因而，为了能使我们的模型能更好地对解码复杂场景语义信息，我们采用 CLIP[8]，一种多模态视觉语言模型，来解码沙盘图像。要强调的是，我们使用了 CLIP 的 Vision Transformer 作为我们 Encoder。实验结果如表6所示。我们发现，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。

Table 6: 基于 CLIP Vision Encoder 的实验结果。

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

4.6 针对加权系数 p 的消融实验

加权系数 **p** 对模型训练非常重要，这里我们进行了一项额外的消融实验以证明 **p** 的有效性。我们依然采取 ResNet50 作为 backbone，损失函数去掉 **p** 一项并进行实验。实验结果如表7所示。与表3进行对比，我们发现 **p** 对于模型训练至关重要，没有加权系数无法训练出可识别模型。

Table 7: ResNet50 损失函数没有加权系数 **p** 的实验结果。

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

4.7 针对对比学习系数 β 的消融实验

针对加入的对比学习损失，我们对其权重进行消融实验。从 Table 8中我们可以看出：(1) 有监督的对比学习能够增强模型的整体性能。(2) 当对比学习损失的权重参数为 0.3 时，模型平均准确率达到最高。

Table 8: ResNet50 损失函数在不同对比损失加权系数 β 下的实验结果。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

5 结论

本文中，我们针对沙盒图像提出了一个基于视觉理解的沙盒主题多标签分类模型。我们仅用了很少的图像，123 张训练图像便很好的训练出来了一个可识别主题的模型。在我们的实验结果中，表明网络的深度是需要的，例如 ResNet18 与 VGG16 训练的模型无法有效识别沙盒主题。我们仍然需要注意对数据的标签进行有效加权，以防止数据分布不均衡对模型训练造成的偏见性影响。最后我们用 ResNet 验证了我们的模型，并证明了模型的有效性。我们也讨论了使用风格迁移的方式进行域自适应，以探索自然图像理解与沙盒图像理解的差异。我们讨论了使用基于预训练的卷积神经网络与 Vision-Transformer 的 CLIP 对模型进行微调。尽管数据集的数据量很稀缺，但是我们依然证明了深度网络具有对图像的理解能力，期待着更先进的视觉理解算法的提出。

Acknowledgement

本文由 6 个来自中国科学院信息工程研究所[1]的同学共同完成，同时感谢中国科学院自动化研究所智能系统与工程研究中心[2]的黄凯奇，陈晓棠和赵鑫老师的指导。本文每个同学的贡献如下。

姓名	学号	贡献
陈若愚	202118018629015	组长，组织安排；多标签模型 (Sec. 2) 与加权损失函数；论文初稿。
唐源民	202118018670068	PPT 第一版制作，多模态预训练模型设想 (Sec. 3.2) 与实验。
杨嘉瑞	2021E8018682140	ConvNet 预训练模型微调 (Sec. 3.1) 与实验。
涂皓钦	202128018670009	对比学习消融 (Sec. 2.3) 与实验。
刘康威	202118018670040	PPT 微调与润色，实验结果与内容核对。
颜　广	2021E8018682139	论文终稿撰写。

[1]中国科学院信息工程研究所：http://www.iie.ac.cn

[2]智能系统与工程研究中心：http://www.crise.ia.ac.cn/

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. IEEE, 2009.

[2] 黄凯奇, 张岩, and 丰效坤. 基于视觉分析的沙盘分裂主题识别系统, 方法, 设备, 2022.

[3] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[5] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[6] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673, 2020.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *Image*, 2:T2, 2021.