RETFound自监督学习在OCT中的运用

技术报告

小组成员：冯泽欣，李子豪，黄子勖，张伟祎，陈奕冲，段柱材
TA:肖尊杰

Introduction

一、SSL

1. 定义

自监督学习(Self-supervised learning) 是机器学习中的一种方法，它在没有或很少有标注数据的情况下，通过训练模型来理解数据的内部结构。这种方法通常用于处理大量未标注的数据，尤其适用于那些获取 标注数据成本高昂或者困难的场景。自监督学习的核心在于，模型通过预测数据的某些部分或属性来学习数 据的表示。

无监督和自监督学习的训练数据都是无标签，但区别在于：自监督学习会通过构造辅助任务来获取监督信息，这个过程中有学习到新的知识；而无监督学习不会从数据中挖掘新任务的标签信息。

它旨在对于无标签数据 ，通过设计辅助任务（Proxy tasks）来挖掘数据自身的表征特性作为监督信息， 来提升模型的特征提取能力。

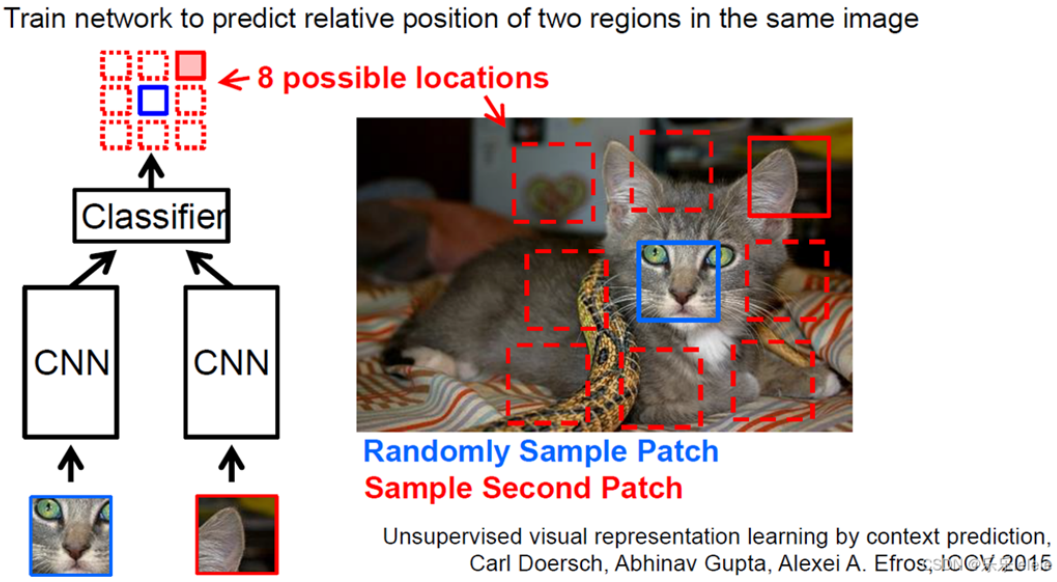
2. 工作原理

在自监督学习中，通常会对原始数据进行一定的处理，创建一个任务（就是**辅助任务**，这是自监督学习最关键的内容），让模型在没有显式标签的情况下学习。这种处理可以是遮挡部分数据、预测数据序列的下一个元素等。模型在这个过程中学习到的表示可以被用于各种下游任务，比如分类、检测或者更复杂的理解任务

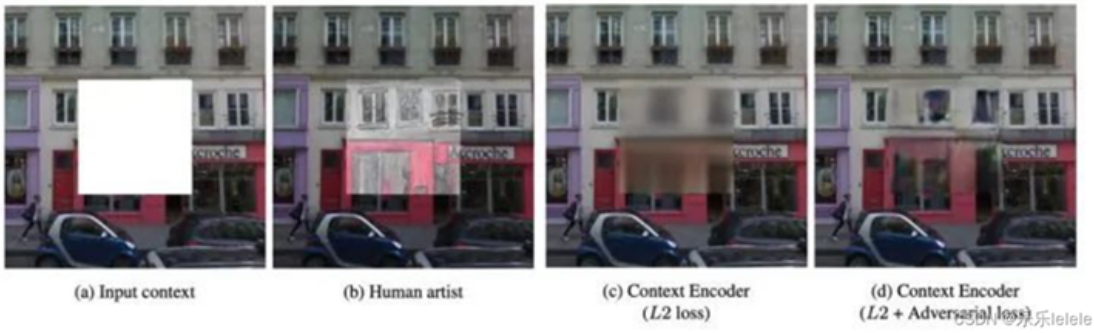
3. 例子

在图像处理中，一个常见的任务是图像的重建。模型可能只被给予图像的一部分，并被要求预测缺失的部分。通过这种方式，模型学习到了图像的关键特征和结构。

- **图片重组**：将图像分为不同的patch，比如九宫格，然后让网络预测不同patch的相对位置信息。这个过程中可以提高模型的局部特征提取能力以及全局空间信息提取能力。



- **图片修复**：对图像进行随机裁剪，训练网络修复图像



4. 总结

- **优点**
  - **减少对标注数据的依赖**：自监督学习可以在没有大量标注数据的情况下训练有效的模型。
  - **更好的泛化能力**：通过学习数据的内在结构，自监督学习模型通常能够更好地泛化到新的、未见过的数据。
- **缺点**
  - **任务设计的挑战**：设计有效的自监督任务可能很具挑战性，需要对数据有深入的理解。
  - **可能需要更多的计算资源**：为了有效地从未标注的数据中学习，自监督学习模型可能需要更多的计算资源和更长的训练时间

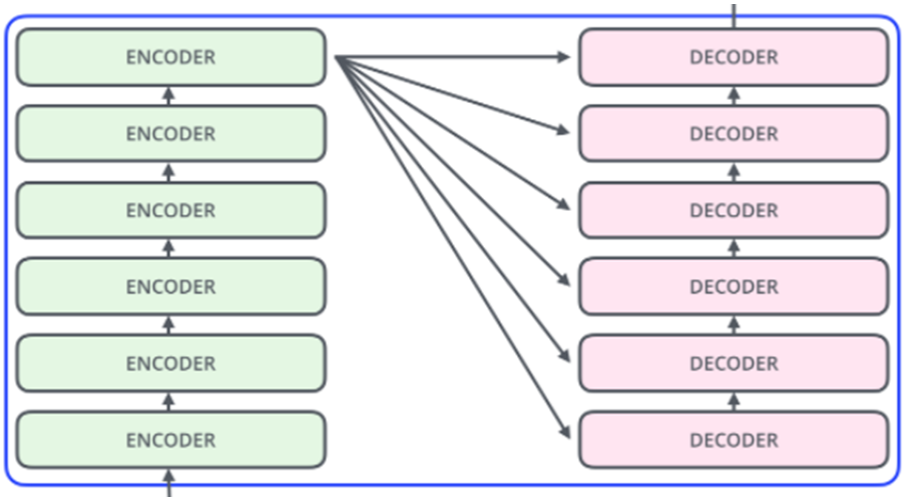
二、Transformer & ViT

1. 概述

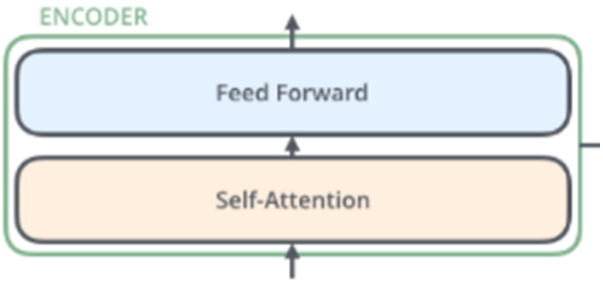
Transformer是由Vaswani等人在2017年提出的一种深度学习模型，主要用于处理序列数据，如自然语言处 理（NLP）。它的主要创新在于使用了“自注意力（Self-Attention）”机制，这使得模型能够在处理序列数据 时更加高效和灵活。

2. 架构

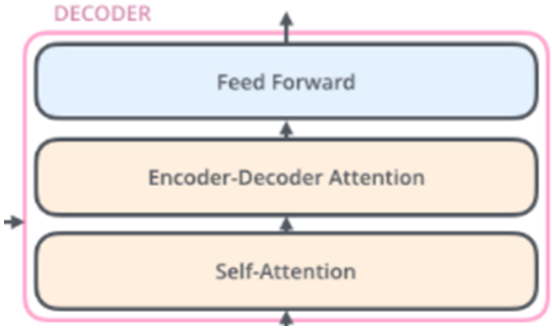
Transformer 本质上是一个 Encoder-Decoder 架构。主要由编码组件和解码组件组成。其中，编码组件由多层编码器组成，解码组件也是由相同层数的解码器组成。



- **编码器**：每个编码器由两个子层组成：自注意力和前馈网络层

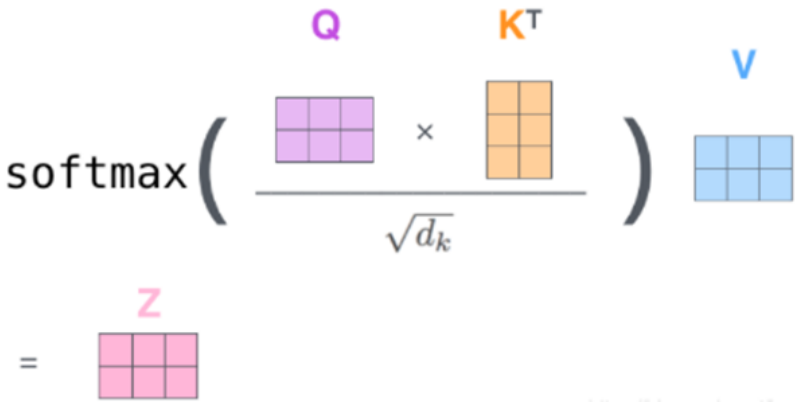


- **解码器**：在编码器结构基础上加入Encoder-Decoder Attention 层。

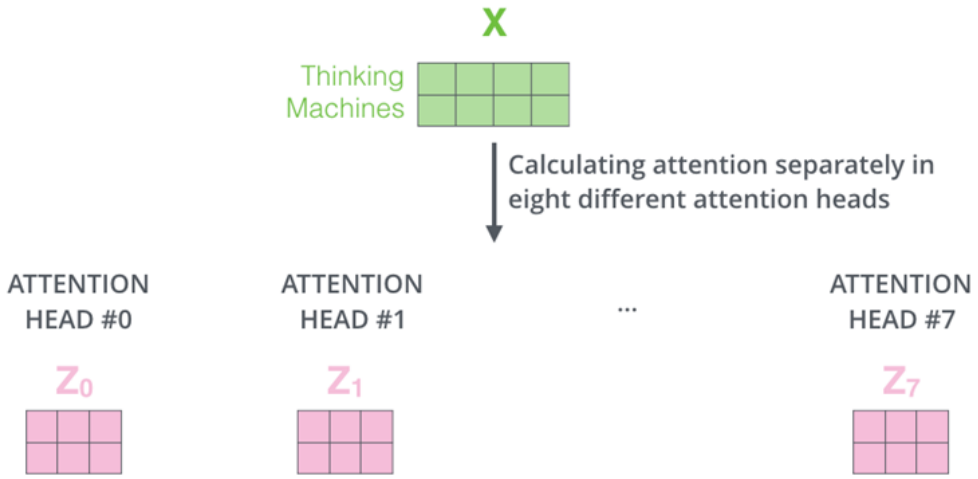


3. 核心机制

- **自注意力机制**：允许输入序列中的每个元素都与序列中的其他元素进行交互，从而捕获序列内的全局依赖关系，使模型可以更直接地处理数据中的长距离依赖问题。



- **多头注意力机制**：进一步完善自注意力层，将注意力分为多个头，每个头学习输入数据的不同方面。用于将输入映射到不同的子表示空间，使得模型可以在不同子表示空间中关注不同的位置。



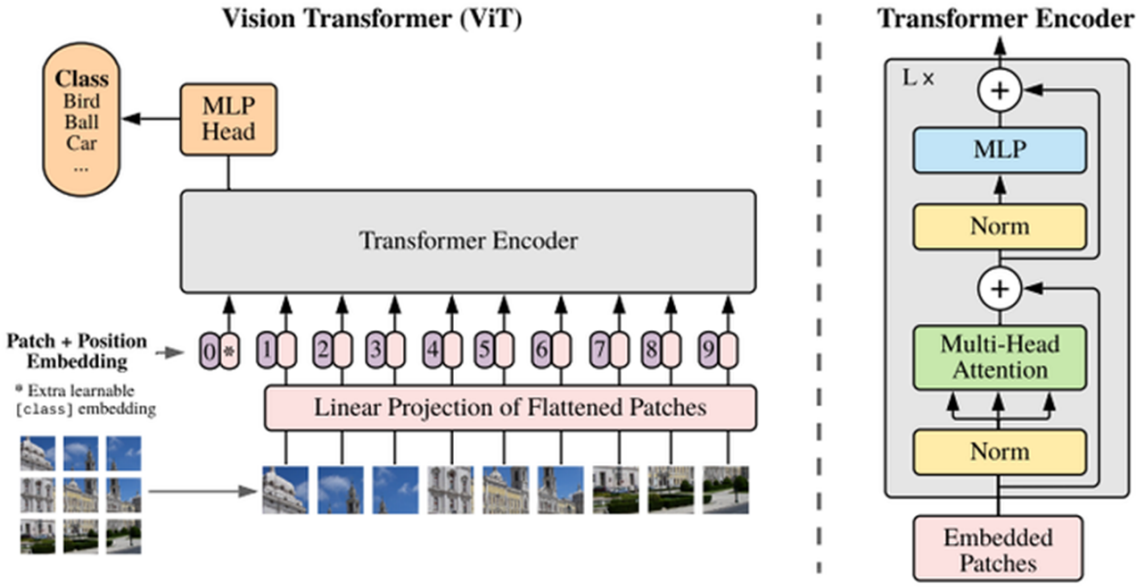
4. ViT（Vision Transformer）概述

Vision Transformer是一种应用于计算机视觉任务的模型。 ViT 只使用了 Transformer 的编码器部分，ViT将Transformer架构——最初用于自然语言处理——成功应用于图像识别任务中。 其核心思想是将图像划分为一系列小块（patches），然后将这些小块处理成类似于自然语言处理中的序列数据

5. ViT核心组件

- **图像划分（Image Patching）：**解决图像直接输入复杂度过大的问题
  - 图像被划分为大小相等的小块（例如，16x16像素的小块）
  - 这些小块被展平并转换成一维的向量序列embedding
- **线性嵌入（Linear Embedding）：**
  - 将图像块的一维向量通过线性层转换，以生成固定大小的嵌入向量（补全维度）
  - 这些嵌入向量被送入Transformer模型
- **位置编码（Positional Encoding）：**
  - 生成CLS符号的TokenEMB

- **类别嵌入（Class Token）**：在输入序列的最前面添加一个特殊的类别标记（通常称为[CLS] token），它经过 Transformer 编码器的所有层，其最终状态被用作图像表示
  - 生成所有序列的**位置编码**
    - 位置编码的重要性：并行输入，标志位置
  - token + 位置编码



6. 总结

Transformer通过其独特的自注意力机制和编码器-解码器架构，在处理复杂的序列数据（尤其是文本）时显示出显著的效率和效果。Vision Transformer是计算机视觉领域的一个重要突破，它将Transformer架构有效地应用于图像处理任务中。尽管存在对大量训练数据和计算资源的依赖，但其在图像识别等任务上展现出的强大性能和潜力使其成为当前计算机视觉研究的热点

三、MAE

1. 概述

MAE全称为掩码自编码器(masked autoencoders)，是一种自监督学习模型。 方法过程为随机MASK住图片里的一些块（patch），然后再去重构这些 被MASK住的像素，令得到的预测结果与真实的imagepatches之间的误差作为 损失。

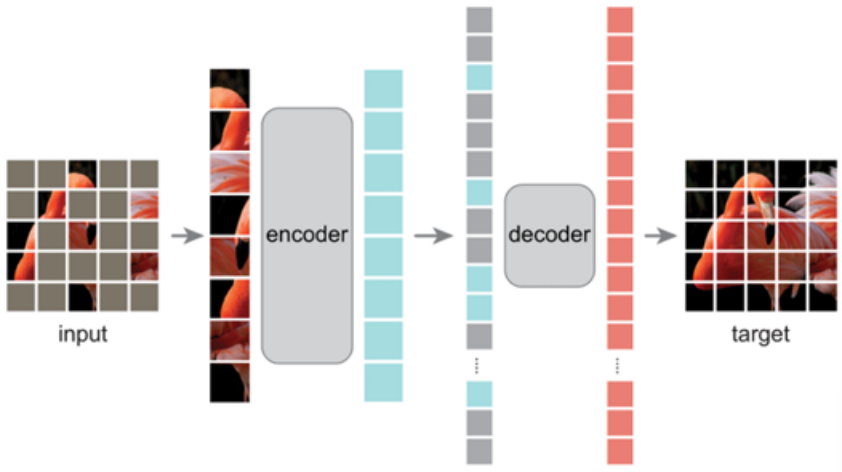


Figure 1: MAE学习架构

2. 架构

预训练阶段一共分为四个部分，MASK，Encoder，Decoder，还有Reconstruction Target

Mask

一张图片作为输入，首先把其切块切成一个一个的小块，按格子切下来。其中要被MASK住的这一块就是涂成一个灰色，然后没有MASK住的地方直接提取出来。采取随机采样，而不是中心采样，网格的采样，局部采样等方式，这部分在实验里对比过。这里比较符合认知的解释是，可以防止引入类似中心归纳偏好等特定bias，随机是最公平的。

Encoder

MAE Encoder 采用 ViT 架构，但只会作用于 unmasked images。和 ViT 思路一样，MAEEncoder 会先通过 Linear Projection 编码图片，再加上位置编码，随后送入一堆连续的Transformer Block 里面。但是编码器只对整个图片 patches 集合的一个小子集 (例如25%) 进行操作，而删除 masked patches。这 里和BERT 做法不一样，BERT 使用对于mask 掉的部分使用特殊字符，而 MAE 不使用掩码标记。

Decoder

MAE Decoder 采用 Transformer 架构，输入整个图片 patches 集合，不光是 unmasked tokens (Figure 1中蓝色色块)，还有被 mask 掉的部分 (Figure 1中灰色色块)。每个mask tokens 都是一个共享的、学习的向量，它指示了这里有一 个待预测的tokens。作者还将位置嵌入添加到这个完整image patch 集合中的 所有tokens 中，位置编码表示每个patches 在图像中的位置的信息。

MAE Decoder 仅用于预训练期间执行图像重建任务。因为自监督学习的特点就是只用最后预训练好的 Encoder 完成分类任务。因此，可以灵活设计与编码器设计无关的解码器结构。作者用比编码器更窄更浅的很小的解码器做实验。在这种非对称的设计下，tokens就可以由轻量级解码器处理，这大大缩短了预训练的时间。

Reconstruction Target

Decoder 的最后一层是一个 Linear Projection 层，其输出的 channel 数等于图像的像素(pixel) 数。所以 Decoder 的输出会进一步 reshape 成图像的形状。 损失函数就是MSELoss，即直接让 reconstructed image 和 input image 的距 离越接近越好。

还有另外一种损失函数，就是先计算出每个patch 的像素值的 mean 和 deviation，并使用它们去归一化这个 patch 的每个像素值。最后再使用归一化 的像素值进行MSELoss 计算。但是发现这样做的效果比直接MSELoss好。

3. 训练过程

以下是MAE的具体实现方法：

1. 首先通过 Linear Projection 和位置编码得到 image tokens。
2. 随机 shuffle 这些 tokens，按照 masking ratio 扔掉最后的一部分。
3. .把 unmasked patches 输出到 Encoder 中，得到这些 tokens 的表征。
4. 把 Encoder 的输出，结合 masked tokens (可学习的向量)，执行 unshuf f le操作恢复顺序，再一起输入到Decoder 中。

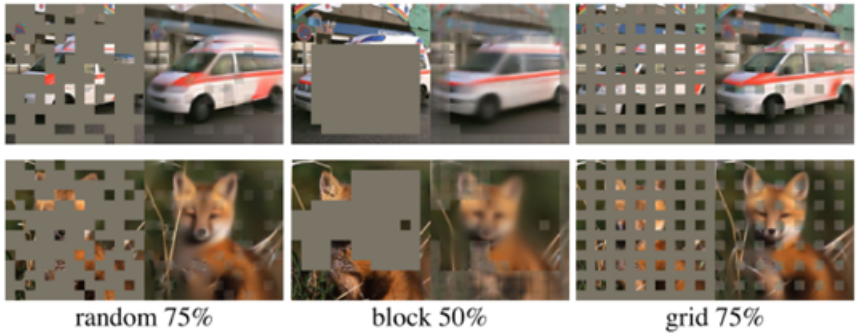


Figure 2: 不同mask策略的影响

4. 优势

- **拓展性**
  - Encoder只操作可见patches，把mask tokens给本身参数就不多的decoder去运算，大大降低了计算量，尤其当mask的比例很高的时候，大大减少了预训练时间，让MAE可以很轻松的scale到更大的模型上（enabling us to easily scale MAE to large models），并且通过实验发现随着模型增大，效果越来越好
- **容量和泛化性**
  - 使用MAE预训练方法，可以训练很大的model，比如ViT-Large/Huge，当把预训练好的ViT-Huge迁移到下游任务时，模型表现非常好，甚至超过了使用监 督预训练的相同模型（achieves better results than its supervised pre-training counterparts），这说明MAE预训练学习到的表示可以很好的泛化到下游 任务（these pre-trained representations generalize well to various downstream task）

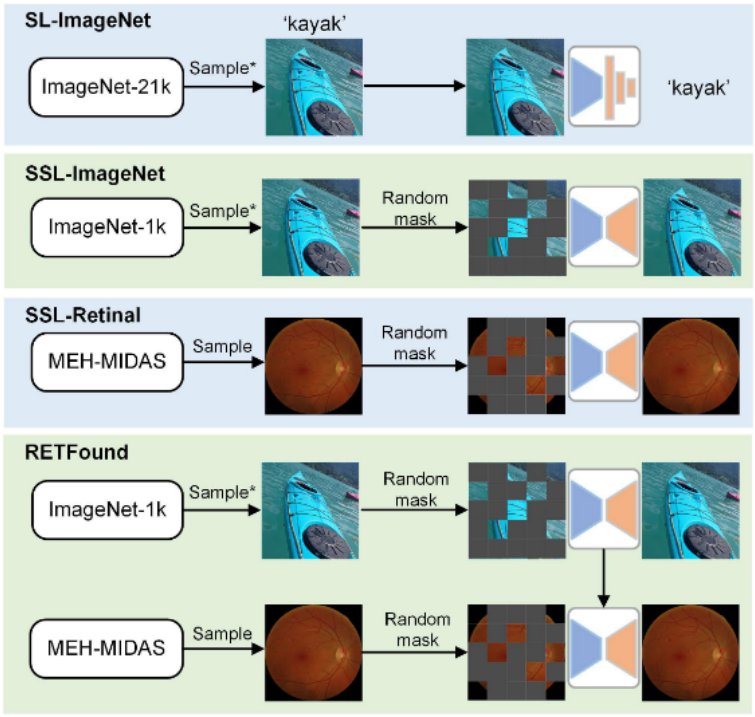
Experiment

一、论文分析

论文题目：《A foundation model for generalizable disease detection from retinal images》

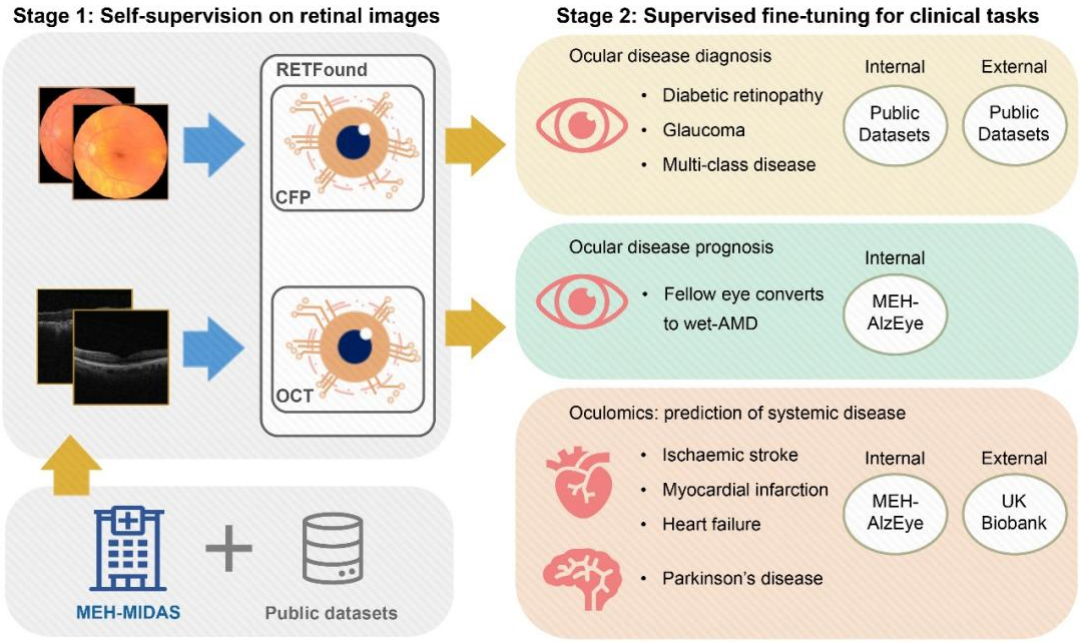
论文链接：[A foundation model for generalizable disease detection from retinal images | Nature]

医学人工智能能识别和关注眼科扫描图像中的特征性信息，从而推进眼科疾病和系统疾病的诊断。然而，开发高性能AI模型过程中，训练和测试需要庞大的、高质量的带标注的数据集，医学专家的标注容量已经无法满足呈指数级上涨的模型开发需求，导致大量医疗数据因无标签而未得到充分利用。通过自监督学习，AI模型可以获得强大的表征学习能力，并在下游任务中提高微调的性能，例如诊断糖尿病黄斑水肿。



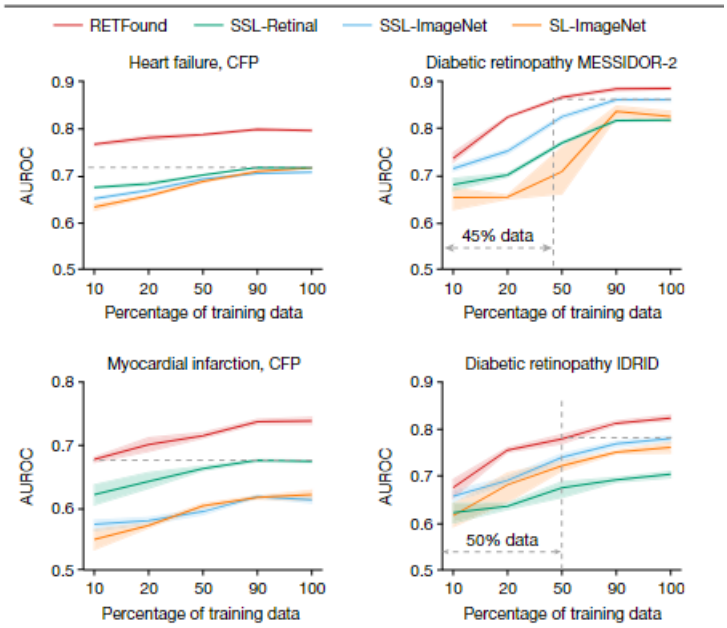
在论文中，研究人员通过自监督训练，在大规模未标注的视网膜图像上构建基础模型RETFound，并用它来促进多种疾病的检测。该研究使用自监督技术Mask Autoencoder依次在自然图像（ImageNet-1k）和160万的视网膜图像上进行训练。开发了两个的RETFound模型，一个用于彩色眼底摄影（colour fundus photograph），另一个用于光学相干断层扫描（OCT），并通过微调RETFound来适应一系列下游疾病检测和预测任务。

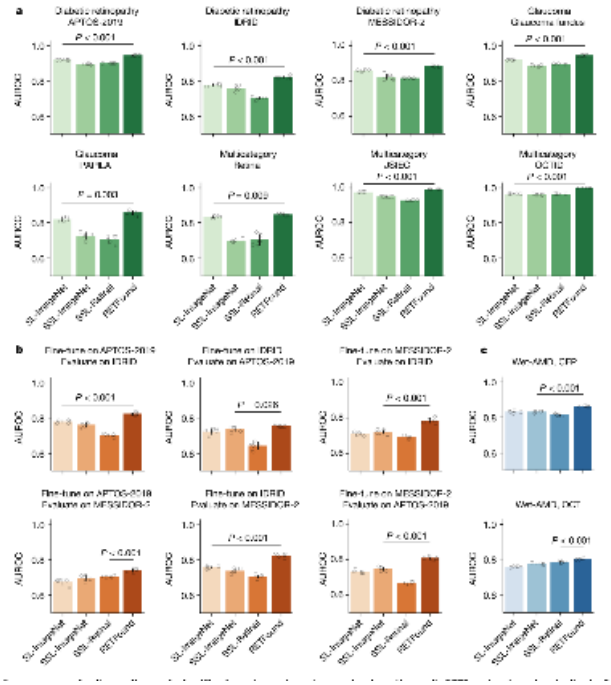
在训练阶段，Encoder包含24个ViT-large，嵌入向量大小为1024；Decoder包含8个ViT-small，嵌入向量大小为512。同时，在训练的过程中，CFP的遮掩率为0.85，OCT的遮掩率为0.75，batch_size的大小为1792， 共计800个训练周期，学习率预热阶段为15epoch。



在验证中，研究人员首先研究了眼部疾病的诊断分类，包括糖尿病视网膜病变和青光眼。研究人员在8个眼科公开数据集进行实验并观察到了性能的显著提升。为了研究更加广泛的应用，研究人员使用AlzEye数据集构建了眼科疾病预后任务，即在老年黄斑病变人群中，预测未治疗眼（fellow eye）在一年内转化为新生血管性老年黄斑病变（wet AMD）的概率，和"oculomic"（利用视网膜图像观测系统疾病，包括缺血性中风，心肌梗塞，心衰，和帕金森病）。

为了验证模型在不同数据集上的泛化性，研究人员使用英国生物样本库（UK Biobank）作为外部测试集。RETFound在内部和外部验证上都显示了最佳性能。





通过论文中的数据，RETFound与传统的迁移学习预训练模型和其他自监督模型相比，在适应下游任务时表现出一致的优秀性能和标签效率，例如在心梗和心衰预测上，RETFound只需要10%的标注数据就可以达到对比方法的最佳性能。我们可以推知，通过mask遮蔽重建任务，模型增加了对图形的某种理解，并将这种理解带到了下游任务，使得模型在下游任务的训练中收敛速度更快，预测准确率得到提高。

该研究验证了RETFound在适应多种医疗应用中的功效和效率，展示了在检测眼部疾病方面的高性能和泛化能力，以及在预测系统疾病方面的显著改进。通过克服当前临床AI应用的障碍，尤其是标注数据的规模和性能以及泛化能力的限制。

二、实验复现

1. 数据集搭建

在训练阶段，我们从多个途径搜集了OCT视网膜图像数据，最终我们的数据集总量为182980张。

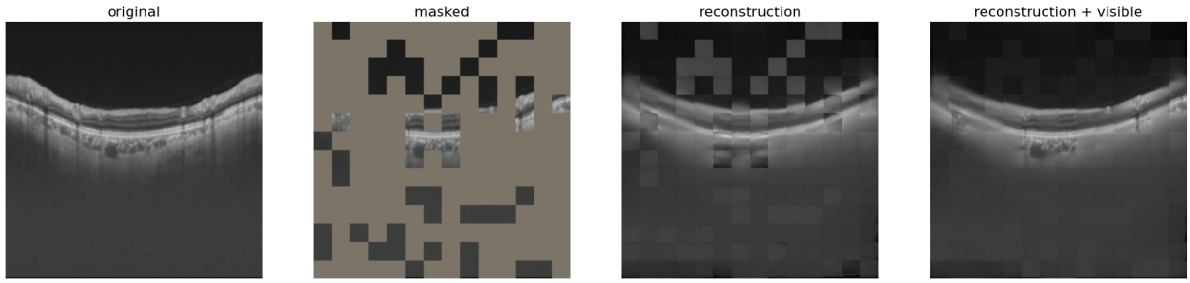
OCT	Total	Train	Validate	Test	Segmentation Labels	Classification Labels
Baidu Goals	200	100	0	100	1	GC_Label
DukeOCT2015	610	610	0	0	0.5（未能导出）	无
Topcon	1280	768	256	256	1	无
OCTA500	180800	180800	0	0	1	很多

2. 训练

较大的 `batch_size` 需要更多的内存来存储模型参数、梯度和中间计算结果，因此受显存限制，与论文中1792的`batch_size`相比，在我们的实验中`batch_size`只取到了16。

参数	参数值
batch_size	16
world_size	1
epochs	50
lr	5.00E-03
layer_decay	0.65
weight_decay	0.05
drop_path	0.2
input_size	224

mask重建结果展示：

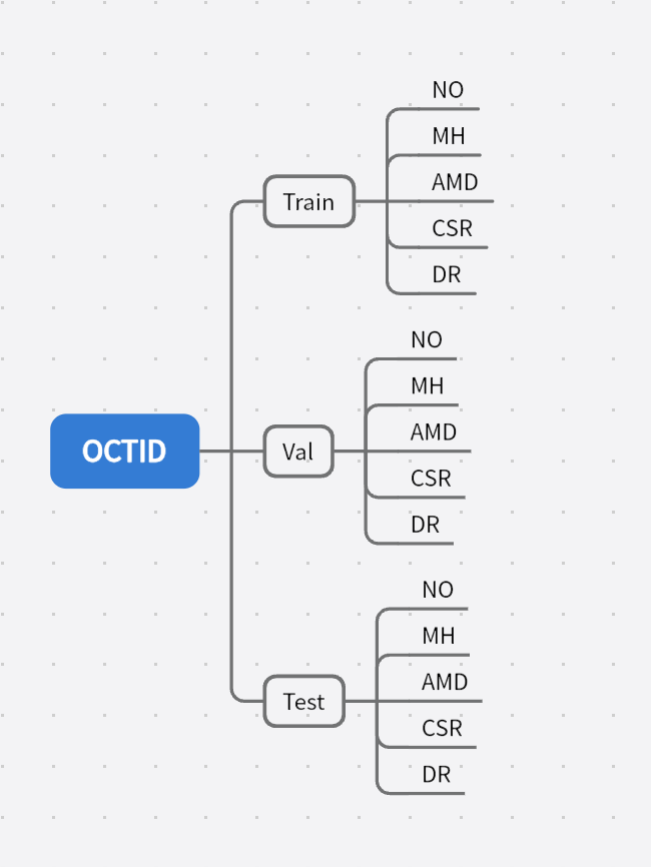


三、下游任务验证与比较

1. 数据集的选择与处理

在下游任务验证阶段，我们使用的OCT图像数据集是OCTID，该数据库包含按不同疾病分类的OCT图像，由超过500个光谱域OCT体积扫描组成，包括五个类别：正常(NO)，黄斑孔(MH)，年龄相关性黄斑变性(AMD)，中心性浆液性视网膜病变(CSR)和糖尿病视网膜病变(DR)，共577张。

为了方便直接使用论文所附代码进行测试，我们将数据集按照论文中所提供的架构进行组织：其中train：val：test = 6：2：2。



测试阶段参数：

参数	参数值
batch_size	16
world_size	1
epochs	50
lr	1.00E-03
layer_decay	0.65
weight_decay	0.05
drop_path	0.2
input_size	224

2. 对比试验

我们比较了以下四种训练模型在OCTID分类上的性能：

(1). 在没有任何预训练模型上测试

(2). 在RETFound使用自然图像预训练所得模型上测试

(3). 在RETFound使用oct图像进行预训练所得模型上测试

(4). 在复现所得模型上测试

我们使用了LabelSmoothing CrossEntropy作为损失函数：

```
criterion = torch.nn.CrossEntropyLoss()
```

Label Smoothing CrossEntropy 在交叉熵损失的基础上引入了标签平滑（Label Smoothing）的概念。它的主要优势在于对模型的泛化性能和过拟合的处理上有一些改进。标签平滑可以被看作是一种正则化的形式，它使得模型对于训练数据中的噪声不敏感，从而减轻过拟合的程度，使模型更加鲁棒。

同时我们使用了慢启动的AdamW作为优化器：

```
optimizer = torch.optim.AdamW(param_groups, lr=args.lr)
```

慢启动（Warmup）是一种优化算法中的一项技术，它在训练初始阶段使用较小的学习率，并逐渐增加学习率。慢启动的目的是在训练初期保持模型参数的稳定性，防止在学习率较大的情况下模型参数波动过大，影响收敛。AdamW 在训练中使用了权重衰减（Weight Decay）来防止过拟合。慢启动可以确保在训练初始阶段，权重的更新不会太大，从而维持模型参数的稳定性，帮助模型更好地收敛。

3. 结果分析

	ACC	sensitivity	specificity	precision	AUC_roc	AUC_pr	F1	mcc	metric_logger.loss
RETFound Without pre-train	0.803097345	0.178571429	0.913043478	nan	0.776046901	0.450879341	nan	nan	1.4085 (1.6316)
RETFound With SSL-imagenet1k-pre-train	0.969026549	0.860281385	0.984345351	0.896825397	0.98794333	0.952160324	0.875914634	0.859583958	0.2637 (0.3847)
RETFound With OCT-image-pre-train	0.973451327	0.872186147	0.986767112	0.922754329	0.985130122	0.95331405	0.892085984	0.879930776	0.2218 (0.2130)
Our Model	0.880530973	0.547835498	0.932328191	0.740051615	0.914913794	0.767524694	0.533444816	0.526865889	0.6454 (0.6363)

注：性能评价指标含义解释：

混淆矩阵：

	正例	负例
预测正	真正例 (true positive, TP)	假正例 (false positive, FP)
预测负	假负例 (false negative, FN)	真负例 (true negative, TN)

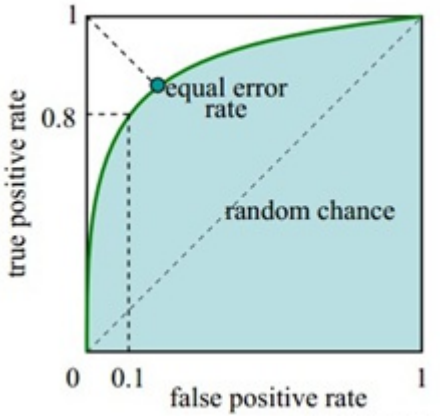
① ACC: 准确率(Accuracy) 判断正确的结果与所有观测样本之比。

② Sensitivity: 灵敏度(或称召回率/真阳率) 模型预测正确的个数占真实值为positive的比例。

③ Specificity特异度(或称选择率/真阴率) 判断正确的个数占真实值为Negative的比例。

④ precision: 精确率(或阳性预测值) 判断正确的结果占预测为positive的比例

⑤ AUC_roc: 由TPR-FPR曲线（ROC）包围的面积
ROC曲线是通过改变分类的阈值，进而得到一系列的（TPR,FPR）的点，然后根据阈值从小到大得到的点绘制成TPR-FRP曲线，这条曲线称之为ROC曲线，然后计算曲线包围的面积，当面积越大时，说明性能越好。即AUC越大性能越好



⑥ AUC_pr: Precision和Recall组成的曲线，跟ROC曲线类似，改变阈值，得到一系列的RECALL和PRECISION点，绘制成的曲线。

P-R曲线包围的面积称之为AP，AP越大性能越好
⑦ F1: 平衡F分数，被定义为精准率和召回率的调和平均数。

⑧ mcc: 马修斯相关系数
用于评估二元分类模型性能的指标，特别适用于处理不平衡数据集。它考虑了真正例（TP）、真反例（TN）、假正例（FP）和假反例（FN），提供一个能够总结分类质量的单一数值。

由表中数据可知，我们复现所得模型在下游任务分类任务中的准确率比未经过预训练的模型提高0.08%，表现优于未经过预训练的模型；但与论文中提供的RETFound在自然图像上和在OCT图像上进行预训练的模型相比，准确率出现了下降。

四、结论

在本次复现任务中，我们对MAE自动编码器的探索也相对成功的，遮蔽复原的图像较好地还原了原有图像的重要特征；但对于MAE的目的，即通过预训练得到有效模型以增强下游任务的准确率，结果并不理想，我们复现得到的模型仅优于未经过预训练的模型，而比在自然图像和在OCT图像上进行预训练的表现都差。也就是说，我们对SSL-imagenet1k训练过模型进行OCT预训练，反而对模型下游任务的表现产生了干扰。

Discussion

在这次项目中，我们对数据集的处理、模型的整体架构、下游任务的目的和处理方式有了更多的理解和实践。通过这次项目，我们认识到接下来我需要在实验中深刻理解各个参数的现实意义及调整这些参数带来的影响，从而进一步能够在遇到不同问题时合理调整参数以实现更精准的预测；也应提高代码水平，对神经网络模型的结构有更全面的认识。

自监督学习虽然优点众多，但是在实际训练和使用的过程中，需要的训练、参数技巧非常复杂，在本课程的学习中我们了解到此学习方法并尝试进行了训练。受制于设设备和数据集。我们的样本量远小于RETFound论文中的数据（18w比70w）。显存原因batch size也有所调整。这些参数造成了最终训练效果的不理想。在课程后我们也会继续改进，如：增大数据集、继续更改训练策略等。