深度学习在医学影像诊断中的应用

人工智能 91 利友诚

January 2023

1 引言

1.1 乳腺超声简介

乳腺癌近年来成为女性第一大恶性肿瘤 [1]，而乳腺癌的筛查对预防乳腺癌有巨大作用。根据 BI-RADS 指南 [2]，目前主流的乳腺癌筛查方法有乳腺 X 射线摄影（Mammography），乳腺超声（Ultrasound）和乳腺核磁共振成像（MRI）。其中乳腺超声以其实时性、低成本、无创性等优点，成为临床常规诊断乳腺病变的一种常用成像手段。乳腺超声检查有助于确定病灶的位置、大小和良恶性，特别是对于致密型腺体中的肿块有更高的敏感度，使其成为乳腺筛查中的重要一环。检查时医生移动超声探头寻找病灶，并依据病灶的形状、深度、边缘情况、后方回声情况等确定病灶的良恶性。

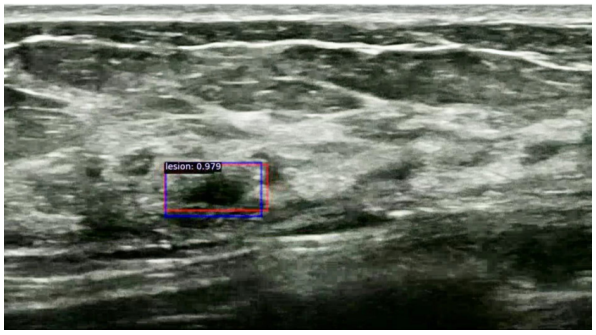

图 1: Breast Ultrasound Detection

与乳腺 X 射线摄影相比，超声城乡在更大程度上依赖于超声师的能力。由于噪声和伪影的干扰，解读超声视频需要非常丰富的经验。而中国城乡医

疗水平差距显著，乡县一级的地方医院往往缺乏合格的超声师。计算机辅助诊断（CAD）有助于减少“操作者依赖性”对超声成像的影响，帮助超声师更好地进行病变检测。本文的工作是在超声视频中进行病灶检测，以辅助医生提高超声检查的召回率和准确率。

乳腺超声视频的病灶检测是一个全新的领域，鲜有前人的研究可供参考。故本文主要从自然视频的检测工作出发，设计对乳腺超声视频特点的模型。超声视频中的病灶检测与一般视频中的物体检测有较大的区别，主要体现在以下三个方面。

超声检测要求实时性，不能利用未来较远的信息辅助当前帧的检测，且对检测模型的速度要求较高；

超声视频中的病灶会随着探头的移动发生形态剧烈变化，病灶的边缘较模糊，更难以确定病灶的真实轮廓；

肿块可能与其他正常的低回声结构和良性病变相似。

由于隐私保护、标注难度大等原因，目前没有大规模数据集，也鲜有实际应用的产品。本文希望通过设计乳腺超声视频病灶检测模型，实现真正落地应用的乳腺超声检测辅助产品，降低乡镇医院乳腺癌漏检率、误检率。

1.2 文献综述

许多视频检测模型是通过图像检测模型（即单帧检测器）叠加帧间信息的结合实现的，所以本文首先介绍经典的单帧检测器。在单帧检测模型的基础上，介绍两种比较有代表性的多帧信息结合方式。最后，介绍前任对乳腺超声视频的检测尝试。

1.2.1 单帧检测

双阶段网络 Faster R-CNN [3] 是最经典的双阶段单帧检测模型。文章创新型地采用了 Region Proposal Network （RPN），通过一个 CNN 从预先设计好的密集的锚定框（anchors）中选取前景概率最大的 k 个作为后续处理的候选框（Region of Interest，RoI）。随后通过 RoI Align 操作将所有 RoI 插值为 $h \times w \times c$ 的特征图，再进行 bounding box 回归和前-背景分类的操作。文章的关键在于提出了 RPN 网络从密集的 anchors 中选取后续处理需要的稀疏的候选框，较之前用传统技术（如 Selective Search [4]）选取的方法高效且准确率更高。

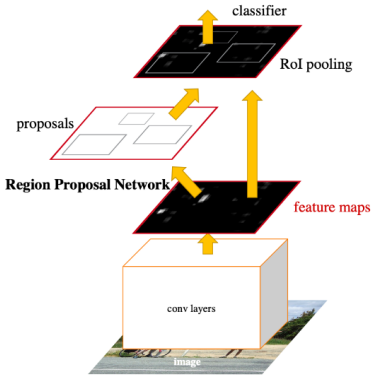

图 2: Faster R-CNN

然而 anchor 是为了给 RPN 提供一个可选择的“锚点”，直觉上生物视觉的产生不会依赖于密集的 anchor。同时，anchor 的尺寸，密集程度等是与数据集相关的超参数，普适性较差。

Anchor Free 方法 FCOS [5] 是第一个 anchor free 的单阶段检测模型。FCOS 将经过 backbone 得到的 $H \times W \times C$ 的特征图上的 $H \times W$ 个点作为“预测者”，即每一个 C 维特征都需要给出一个预测框的便相对“预测者”的上下左右距离 (l, t, r, b) 和对应的置信度，最后通过设置置信度阈值来给出预测结果。FCOS 舍弃了 RPN 的设计，使得检测流程更加整洁。

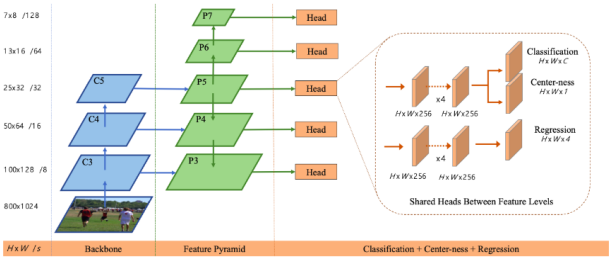

图 3: FCOS

FCOS 只让标注框中心附近的“预测者”来预测这个标准框，若一个预测者接近某个标注的中心，则它更有可能是前景。为了利用这一先验知识，FCOS 设计了一个新的分支来预测“中心值”，$centerness^* = \sqrt{\frac{min(l^*,r^*)}{max(l^*,r^*)} \times \frac{min(t^*,b^*)}{max(t^*,b^*)}}$。“中心值”越大则“预测者”对应的标注框中心越近，测试时在置信度分数上乘以“中心值”作为最终的置信度。另外，标

注框的大面积重叠可能导致同一个“预测者”需要预测两个不同的标注框。FCOS 采用 FPN [6] 结构，并根据物体尺度分配到不同层次的特征图来最大程度地避免标注框重叠。如果仍有重叠，则直接选取面积最小的标注框作为匹配的对象。

FCOS 的预测框仍然是密集的，并且需要非极大值抑制的后处理（Non Maximum Suppression，NMS），即对重合较高的检测框只保留置信度最高的。NMS 的存在会导致无法检出前后遮挡较大的两个物体，即检测模型的上限小于 100%。根源在于计算 loss 时，FCOS 将与标注框重合度较高的检测框都匹配为正样本，并进行梯度回传（one-to-many label assignment）。DETR [7] 创造性地提出将检测作为一个 set prediction 问题，开创了一对一匹配（one-to-one label assignment）的先河。

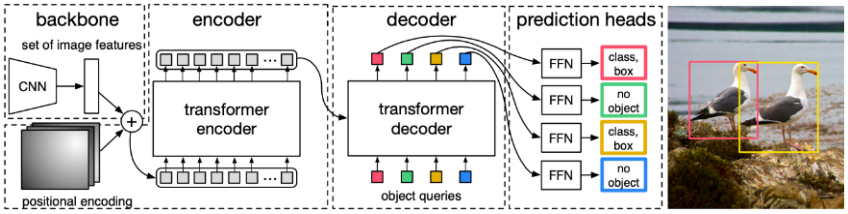

图 4: DETR

One-to-one label assignment DETR 将通过 backbone 得到的特征图叠加位置编码，将 $H \times W$ 个 tokens 通过 Transformer Encoder，作为密集的 keys。再将 k 个训练时可学习，测试时固定 object queries 通过 Transformer Decoder 与 keys 结合得到 k 个预测框和对应的置信度。将预测框与标注框做带权重的二分图匹配，匹配上的即为正样本。k 个 object queries 是稀疏的，作为一种隐式的位置编码起到了询问“某个位置是否有物体”的效果，直觉上也更符合人脑识别物体的模式。DETR 的每个标注框只与一个预测框匹配，这样模型学习到的就是尽量去预测不同物体的框，起到了 NMS 的效果。

虽然 DETR 的效果没有超越当时效果最好的检测器，但是 sparse query 和 one-to-one label assignment 滋养出一批后续工作，如 Sparse R-CNN [8] 和 DeFCN [9]。

Relation 模块 DETR 中采用的 Transformer [10] 应用了 self-attention 模块在 NLP 和 CV 领域都取得了巨大的成功。RelationNet [11] 试图改进原有的注意力机制，通过学习一张图片内毒物体之间的关系来提升物体检测的

效果。作者将物体的特征分为语义特征(appearance)和位置特征(geometry), Relation 模块通过结合同一张图中不同 candidates 的相对位置信息和语义信息得到注意力机制的参数。将 candidates 集合 R_1, R_2 通过 Relation 模块得到的结果记为 $R = Relation(R_1, R_2)$，其中 R_1 作为注意力机制 queries。由于结合了其他 candidates 的信息，Relation 模块也能起到 NMS 的作用。本文第三章将详细介绍 Relation 模块。

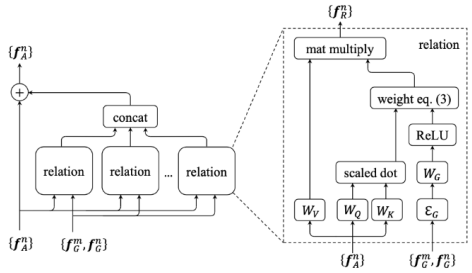

图 5: Relation Module

1.2.2 视频检测

由于 3D-CNN 在视频检测中的效果并不理想，目前的视频检测模型基本采用在单帧检测模型的基础上结合多帧信息的框架，这就产生了两个问题：如何设计单帧检测模型，以及如何结合多帧信息。

RDN [12] 选择 Faster R-CNN 作为单帧检测器，并利用 Relation Encoder 模块来结合多帧的信息。对于需要预测的帧（reference frame）I_t，在 $[t-T, t+T]$ 中随机选取两帧作为“支持帧”（support frames）。将这三帧通过 Faster R-CNN 的 RPN 和 RoI Align 后得到不同的 RoI 特征图，作为后续操作的 candidates。在三帧中分别选取置信度最高的 k 个 candidates，组成大小为 $3k$ 的集合，记为 R^s，其中专属 reference frame 的 candidates 集合记为 R^r。经过初步的 “Basic Stage” 和更为精细的 “Advanced Stage” 得到充分结合 support frames 信息的 candidates。

“Basic Stage” 中，R^s 和 R^r 通过 M 层 Relation Encoder 模块 [11] $Relation_{basic} = f^1_{rm} \cdots f^M_{rm}$ 得到 $R^{r1} = Relation_{basic}(R^r, R^s)$。此时 R^{r1} 经过 “粗糙” refine 的 reference frame 的 candidates 集合。“Advanced Stage” 中，首先从 R^s 中选取置信度前 $r\%$ 组成 R^{sa}，R^{sa} 与 R^s 经过一个 Relation Encoder 得到 $R^{r2} = Relation(R^{r1}, R^a)$。$R^2$ 是经过“精细”refine 的 reference

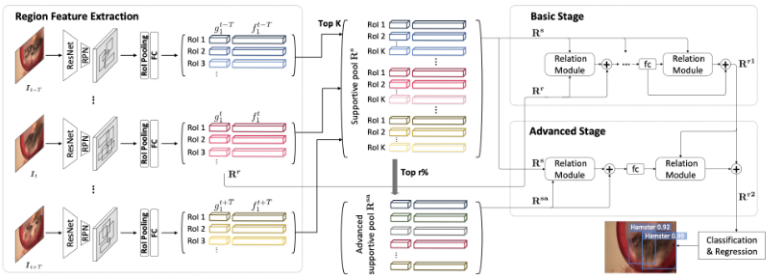

图 6: RDN

frame 的 candidates 集合，进行后续的分类和回归任务。

MEGA [13] 同样选择 Faster R-CNN 作为单帧检测器，并利用了 Relation Encoder 模块来结合多帧的信息。其核心思想是尽量结合所有帧提供物体的语义信息，并更好地利用相邻帧的位置信息。在关键帧 I_t 附近 $[t-T_l, t+T_l]$ 内的帧作为局部（local）信息源，所有的 RPN proposal 记为 L。在所有帧中随机选取 T_g 帧，取其中所有的 proposals 记为全局（global）信息源 G。用不加位置信息的多层 Relation Encoder 模块将 G 的语义信息结合到 L 中，结果记为 $L^g = N_{relation}(L, G)$。将 L^g 用上位置信息的 Relation 模块结合位置信息，结果记为 L^l。

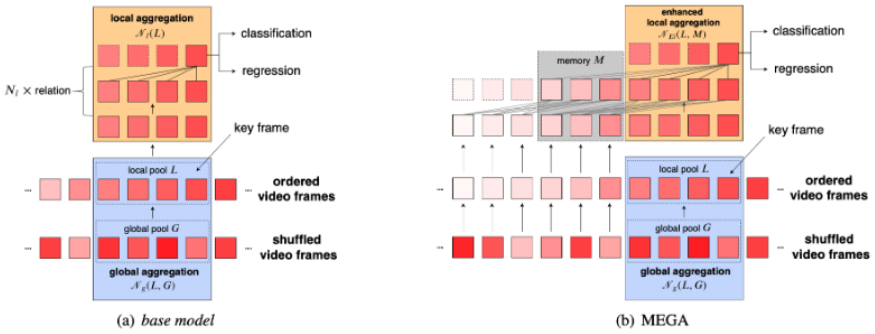

图 7: Relation Module

然而这样做的局限性在于只能结合 T_g 帧的全局信息如何结合更多的全局信息？对于已经完成推理的帧，不把计算好的特征丢弃，而是存在一个大小为 T_m 的缓存，但是由于 $t-T_m$ 帧结合了 $t-2T_m$ $t-T_m$ 帧的信息，所以增加的局部感受野是 $2T_m$。同时每个 L^l 中特征都结合了随机选取的 T_g 帧全局信息，这就大大增加了全局感受野。MEGA 通过巧妙地运用 Relation

Encoder 模块结合帧间信息，达到了目前自然视频检测的最好效果。

1.2.3 乳腺超声视频检测

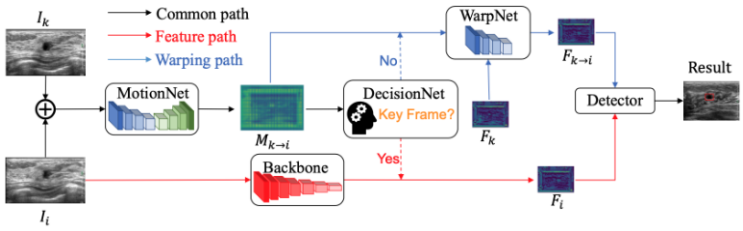

图 8: STCD

STCD [14] 提出了一种半监督学习的方法来进行超声视频的检测。由于对视频进行标注成本过高，文章探究了如何从有标注的超声图片中学习监督信息，并应用到整个视频中。首先在所有有标注的帧上训练一个单帧检测器。对于关键帧 I_k，通过预训练好的单帧检测器 backbone 得到特征图 F_k，对于输入的当前帧 I_i，与 I_k 一起通过一个 MotionNet 得到从关键帧到当前帧的运动信息 $M_{k\to i}$。然后将 $M_{k\to i}$ 通过一个 DecisionNet 来预测一个 S_i 分数，表征 I_i 与 I_k 的相似度。如果 I_i 与 I_k 的相似度低于某个阈值，则将 I_i 设置为新的关键帧。否则 I_i 为普通帧，通过一个 WarpNet 融合 F_k 与 $M_{k\to i}$ 得到 F_i，并进行后续的检测框回归和前-背景分类。

STCD 构思巧妙，但本文探究的任务是监督学习，所有超声视频的所有帧均有标注。监督信息比只有超声图片标注强很多，所以采用的方法也截然不同。乳腺超声食品检测是一个全新的领域，没有太多可以参考的文献，希望本文能为超声视频检测社区提供基于监督学习的新思路。

1.3 初步拟定执行方案

毕业设计活动拟分为四个阶段执行。第一阶段为文献调研阶段，主要工作内容为调研乳腺超声视频目标检测算法，学习乳腺超声常用诊断依据；第二阶段为算法设计阶段，主要工作内容为根据乳腺超声视频数据集统计结果，设计超声视频目标检测算法；第三阶段为实验验证阶段，主要工作内容为在乳腺超声视频中验证算法，并作消融实验，与主流视频目标检测算法比

较，得出实验结论；第四阶段为论文撰写阶段，主要工作内容为根据之前工作，整理为毕业设计论文。以下为各个阶段的起止时间和详细内容。

文献调研阶段（从 2022 年 11 月至 2023 年 1 月）在这一阶段，主要工作内容为调研乳腺超声视频和通用视频目标检测算法，了解最新算法的表现。同时，需要学习超声视频成像原理和乳腺癌超声视频的人工诊断标准。这一阶段的成果主要在毕业设计论文中的导言部分展示。

算法设计阶段（从 2023 年 1 月至 2023 年 3 月）在这一阶段，主要工作内容为统计乳腺超声视频理化指标、设计乳腺超声视频目标检测算法，包括骨干网络、检测头、损失函数和其他模块。这一阶段的成果主要在毕业设计论文中的模型设计部分展示。

算法验证阶段（从 2023 年 3 月至 2023 年 4 月）在这一阶段，主要工作内容为在乳腺超声视频数据集上验证提出的算法，作消融实验，并与主流视频目标检测算法进行比较，最终得出结论。这一阶段的成果主要在毕业设计论文中实验结果部分展示。

论文撰写证阶段（从 2023 年 4 月至 2023 年 5 月）在这一阶段，主要工作内容为整理之前工作，撰写毕业设计论文。

参考文献

[1] Siegel, R., Ma, J., Zou, Z. and Jemal, A. (2014), Cancer statistics, 2014. CA A Cancer Journal for Clinicians, 64: 9-29. https://doi.org/10.3322/caac.21208

[2] Spak DA, Plaxco JS, Santiago L, Dryden MJ, Dogan BE. BI-RADS® fifth edition: A summary of changes. Diagn Interv Imaging. 2017 Mar;98(3):179-190. doi: 10.1016/j.diii.2017.01.001. Epub 2017 Jan 25. PMID: 28131457.

[3] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[4] Uijlings J R R, Van De Sande K E A, Gevers T, et al. Selective search for object recognition[J]. International journal of computer vision, 2013, 104(2): 154-171.

[5] Tian Z, Shen C, Chen H, et al. Fcos: Fully convolutional one-stage object detection[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 9627-9636.

[6] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[7] Carion N, Massa F, Synnaeve G, et al. End-to-end object detection with transformers[C]//European conference on computer vision. Springer, Cham, 2020: 213-229.

[8] Sun P, Zhang R, Jiang Y, et al. Sparse r-cnn: End-to-end object detection with learnable proposals[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 14454-14463.

[9] Wang J, Song L, Li Z, et al. End-to-end object detection with fully convolutional network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 15849-15858.

[10] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[11] Hu H, Gu J, Zhang Z, et al. Relation networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3588-3597.

[12] Deng J, Pan Y, Yao T, et al. Relation distillation networks for video object detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7023-7032.

[13] Chen Y, Cao Y, Hu H, et al. Memory enhanced global-local aggregation for video object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10337-10346.

[14] Chen S, Yu W, Ma K, et al. Semi-Supervised Breast Lesion Detection in Ultrasound Video Based on Temporal Coherence[J]. arXiv preprint arXiv:1907.06941, 2019.