1　舌象数据集制作

本研究选定的基础数据集来自如图 2-1 所示的飞桨网站的 Elastic Heart 的数据集。

图 1-1 数据集网站界面

其中包括五种舌象特征类别，分别是 Mirror-Approximated，Thin-White，White-Greasy，Yellow-Greasy，Grey-Black。对应五种舌象分别是红舌，薄白舌，厚白舌，黄舌，黑灰舌，如图 2-2 所示。该数据集包含训练集 941 例、验证集 236 例、测试集 295 例，一共 1472 例舌头图像数据。

（a） 红舌（b） 薄白舌（c） 厚白舌（d） 黄舌（e） 黑灰舌

图 1-2 五种舌象

经过专业中医的分析，该数据集的数据质量整体较差，不同分类下的质量和数量参差不齐，有大量错误标注的数据，还存在同一张图像同时出现在不同分类下的情况。虽然质量很差，但由于公开的数据集过少，只能在此基础上进行优化。

首先在专业中医的指导下对该数据集进行数据清洗，要对图像质量筛选，删除舌体露出部分过小，清晰度过差，拍摄角度不正，以及图像亮度过高或过暗的图片。这一步是为了确保剩下的舌像质量良好且适合进一步处理。经过数据清洗后的数据集，再根据图像中的舌象对所有的图像重新进行分类。

分类完成后，该数据集又出现了各分类下图像数量严重不均匀的情况，所以又从互联网中重新收集了更多高质量舌头图像并对其进行分类。由于收集到的舌像在各分类下的数量依然有较大差距，为了降低分布不均匀导致模型在训练过程中发生过拟合现象，对数据集又进行了数据增强。数据增强主要采用旋转图像的方法，使得部分分类的舌像数量得到增加，让各个分类下的数据量大小基本相当。数据增强在数据集划分成训练集和测试集之后，以防止出现两张相似的图像同时出现在训练集和测试集。在数据增强前先使用随机程序将舌象数据集按照 6:2:2 的比例进行划分，然后再进行数据增强，通过旋转图像，增加了部分舌象分类下的图像。

1.1　舌像标注

因为重新制作的舌象数据集没有舌象的分类标注数据，所以使用舌象数据集标注

软件如图 2-3 所示，对新增的舌像数据进行分类标注，格式为 YOLO 格式。

图 1-3 舌象数据集标注软件

经过标注后，得到了可以适用于 YOLO 模型训练的舌象数据集，数据集格式树状图如图 2-4 所示。

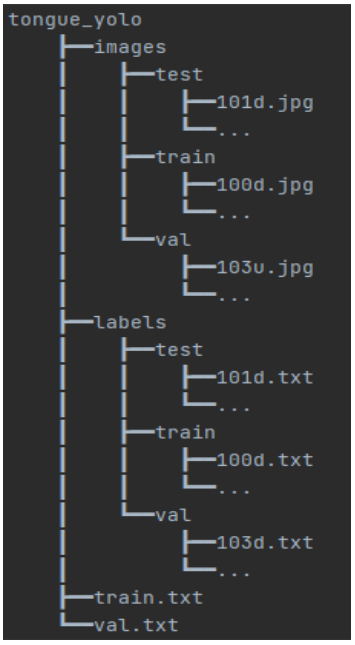

图 1-4 数据集格式树状图

所有数据标注完成后，各分类下的数据量基本一致，最终得到了 480 例训练集，160 例验证集，160 例测试集，共计 800 例舌像数据集。各分类下舌头图像的数据集参数如表 2-1 所示。

表 1-1 数据集参数

	Train	Val	Test	All
红舌	87 张	31 张	29 张	147 张
薄白舌	101 张	33 张	35 张	169 张
厚白舌	92 张	28 张	32 张	152 张
黄舌	102 张	36 张	36 张	174 张
黑灰舌	98 张	32 张	28 张	158 张
总体	480 张	160 张	160 张	800 张

1.2 CEF 数据集

由于 CEF 模型所使用的训练的图像需要先提取图像的色相分布，为了方便后续 CEF 模型的训练，提前将需要训练的数据进行色相分布提取。

首先将图像大小统一调整为 640*640 大小，以便后续处理。由于图像中常用的 RGB 色彩模式数值范围大，变化较多，不方便计算色相分布，所以利用 OpenCV 库将图像转换为 HSV 颜色空间，以便更有效地提取颜色信息。在此过程中，因为亮度对颜色分析的贡献较小，将其删除还可以使需要计算的数据量大大减小。将图像的三维数组整合为二维数组，每行包括三个元素（色相、饱和度、明度）。再从展平后的数组中删除第三个维度（即 HSV 中的 V 值），然后再获取数据中的独特颜色及其出现的次数，并存储在个长度为 1*65026 的整数数组中，未出现的颜色次数为 0，其中索引表示颜色的唯一标识值，值表示该颜色在图像中出现的次数，并将正确的分类选项序号添加在列表末尾，最终将转换的数据保存在 json 文件中，具体实现代码可参考附录 A。在该数据集下的训练集，验证集和测试集与 YOLO 训练的数据集保持一致。

通过将列表转换成了大小为 255*255*1 的黑白色相分布图，效果如图 2-5 色相分布图所示，图中的像素点越亮代表该点对应的颜色数量越多。

图 1-5 色相分布图

2 对舌象智能分类的研究

模型首先接收用户的舌头图像的输入，将图像传输给 CEF 模块与 YOLOv7 进行分类检测，CEF 通过计算分析图像中的色相特征，对舌象类型进行限制，将限制结果插入 YOLOv7 的检测模块，最终经由 YOLOv7 对舌象类型做出判断。

2.1 颜色特征提取

在舌象分类中，舌象分类的准确性对于提供精确的诊断建议至关重要。尽管 YOLOv7 模型在目标检测领域具有出色的性能，但在舌象分类任务中，由于其复杂的背景和多样的舌象特征，单一的模型往往难以达到完美的准确率。在标注舌象数据集的过程中，可以观察到大部分舌象的颜色区分明显，基于这一观察，为了进一步提升 YOLOv7 在舌象分类的准确性，本研究提出了 CEF（Color-based Extraction Feature）模型，旨在通过舌色信息对舌象提高舌象分类模型的准确性。CEF 模型的设计使得舌色在舌象分类中起到了辅助作用。如图 3-10 所示，该模型主要由几个关键部分组成有图像预处理、特征提取、非线性引入、特征降维和最终分类。

根据 2.2 节中描述的 CEF 数据集制作方法，首先需要对输入的舌象图像进行预处理。预处理步骤包括图像的标准化、噪声去除以及舌色分布的初步提取。这一步骤不仅为后续的特征提取奠定了基础，还可以通过预先准备好的数据集，模型在训练时可以直接跳过对图像的色相分布实时提取，从而大大节省了计算资源和时间。

首先，CEF 模型利用多层卷积操作对舌象图像的色相数据进行深入的特征提取。卷积操作能够有效地捕获图像中的局部模式，并在不同层次上提取出越来越抽象的特征。这些特征对于舌色分类至关重要，因为它们包含了舌色的细微变化和分布信息。在特征提取之后，CEF 模型通过 ReLU 函数引入非线性。ReLU 函数是一个简单的激活函数，它能够将模型的输出限制在正值范围内，并允许模型学习复杂的非线性关系。这对于提高模型的泛化能力和分类准确性非常重要。

然后，CEF 模型使用全局平均池化层对提取的特征进行降维。这种操作不仅能够减少模型的计算量，还能够提高模型的鲁棒性，使得模型对特征图的空间位置不敏感。

最后，CEF 模型通过全连接层进行最终的分类预测。全连接层将降维后的特征向量映射到舌色分类空间，输出各个舌色的概率。在本研究中，根据实际需求设置了多个输出节点，分别对应不同的舌象类别，通过增加不同数量的输出节点，决定了舌色类型的数量。

虽然 CEF 模型在单独使用时对舌象分类的效果比较有限，尤其是当舌象之间的色相区分较小时。但是将 CEF 模型与 YOLOv7 模型相结合，利用 CEF 模型提取的舌色特征来限制 YOLOv7 的检测结果。这种组合方法充分发挥了 CEF 模型在舌色特征提取上的优势，同时也利用了 YOLOv7 在目标检测方面的强大能力。实验结果表明，这种组合方法能够显著提高舌象分类的正确率。

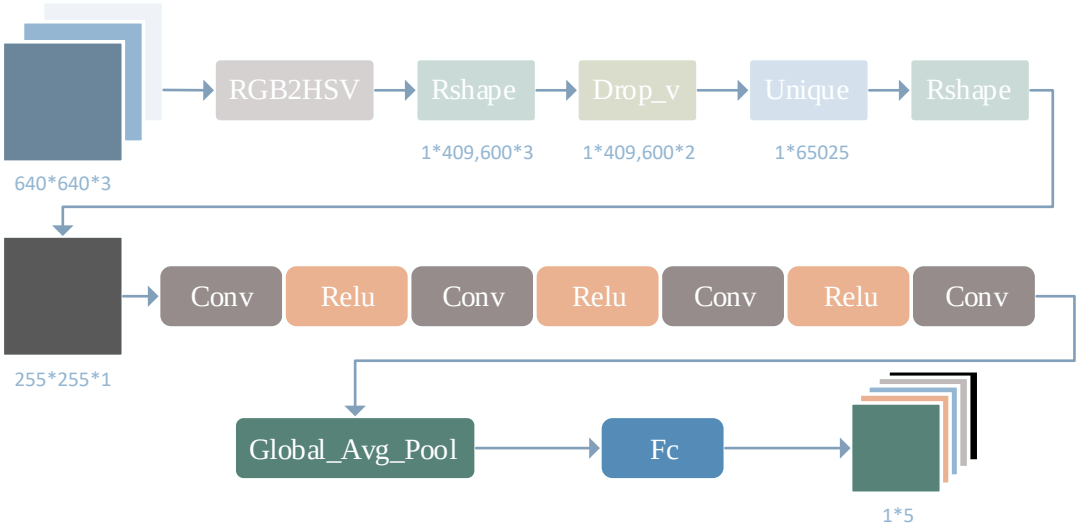

图 2-1 CFE 模型

在得到分类结果后，并不能直接作为舌象分类的结果，只是用来限制 YOLO 模型，降低对非正确选项的概率，通过式(3-1)，首先对 CEF 模型得到的每种分类 i 的 p[i]的概率结果取 0.3 的权重，再通过式(3-2)，将 YOLOv7 的检测结果概率乘以 0.7 与 CFE 结果对应分类下的概率相加得到最终的各分类下的结果，过式(3-3)求出最终分类结果 R。

$$p[i] = p[i] * 0.3 \quad (3\text{-}1)$$

$$p[i] = p[i] + q * 0.7 \quad (3\text{-}2)$$

$$R = \mathrm{Max}(p) \quad (3\text{-}3)$$

通过对 YOLOv7 的检测结果进行修正和限制，基于舌色的分类信息，可以显著降低检测错误的风险，同时又不完全依赖于单一的检测模型，提高了整体分类的准确性和鲁棒性。

3 关于多模态的舌诊对话系统的研究

舌诊对话系统结构流程如图 4-1 所示，用户通过上传舌象图片，经由舌象分类模块将对舌象的判断结果传输至 KBQA 智能医疗诊断对话系统，实现智能舌诊多模态。在系统展示舌象分类结果的同时，还提供针对用户舌象问题的智能解答和合理建议。用户可以在这个交互过程中再输入文本问题，系统会即时解答用户的疑问，从而实现了用户与智能医疗系统之间的自然、有效的交流。

智能舌诊对话系统不仅能为用户提供准确的舌诊信息和诊断结果，还能够满足用户的日常交流需求，使得实用性进一步提高。

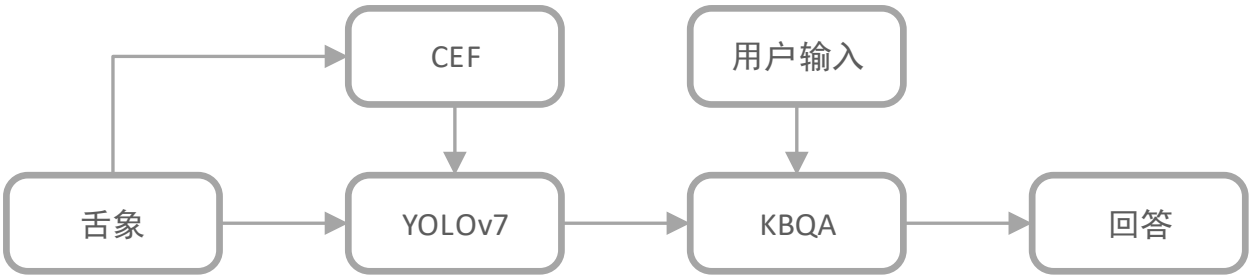

图 3-1 舌诊对话系统结构流程

3.1 KBAQ 系统

本研究中的智能对话部分是基于 Github 网站上 wangle1218 的 KBQA-for-Diagnosis 智能医疗诊断对话系统项目进行扩展的，KBQA 结构如图 4-2 所示。为了更精准地匹配用户输入的问题，并提供相应的答案，模型采用了语义解析的方法对用户问题进行处理。首先，模型使用 jieba 对问题进行分词，然后利用基于 BERT 预训练的意图识别模型对用户输入的意图进行分类。这些意图主要分为两类：闲聊意图和诊断意图。

对于闲聊意图，系统针对问候、告别、肯定和拒绝等四种情况进行了相应的回复。而对于诊断意图，系统会根据意图的置信度进行不同的回答。当诊断意图被确认后，继续利用 BiLSTM-CRF 模型来识别用户输入文本中的病症名称，同时利用 BERT intentre cognition 模型进一步识别诊断意图中的 13 种意图，这 13 种意图与后续 Neo4j 数据库中的各个实体关系相互对应，其他。之后系统会将识别到的命名实体在图数据库中进行查询，将查询结果存储到 json 文件中，最后根据用户意图从 json 查询返回与用户所询问病症相关的问题。

系统通过将查询到的病症相关属性信息存储到 json 文件中，就可以实现在用户对同一病症提出进一步问题时，系统可以直接从 json 文件中提取数据，而无需再次询问用户所需了解的病症信息。这使得对话系统能够实现简单的多轮对话，并更加方便用户获取所需信息。

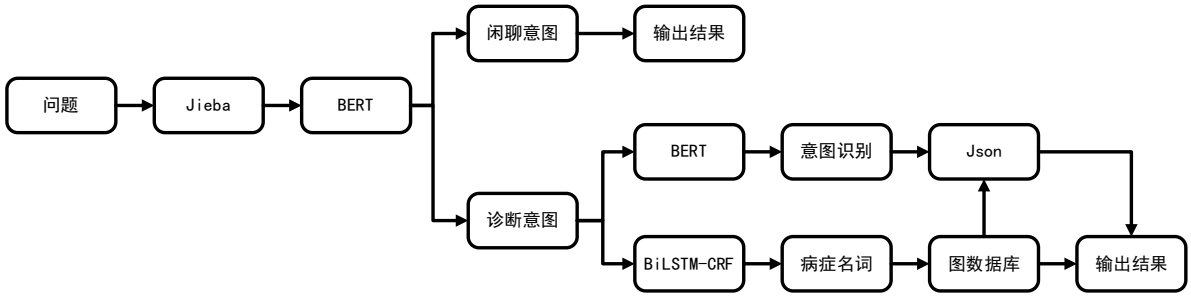

图 3-2 KBQA 结构

3.1.1 BERT 模型

BERT 是由谷歌公司在 2018 年发布的语言模型，基于 Transform 模型的 encoder 部分，通过堆叠 N 层 encoder 层而成，模型通过在不同的超大规模的语料数据集上训练得到了各类可以实现不同文本处理任务的 BERT 预训练模型，可以方便模型在之后的针对意图分类任务进行使用。

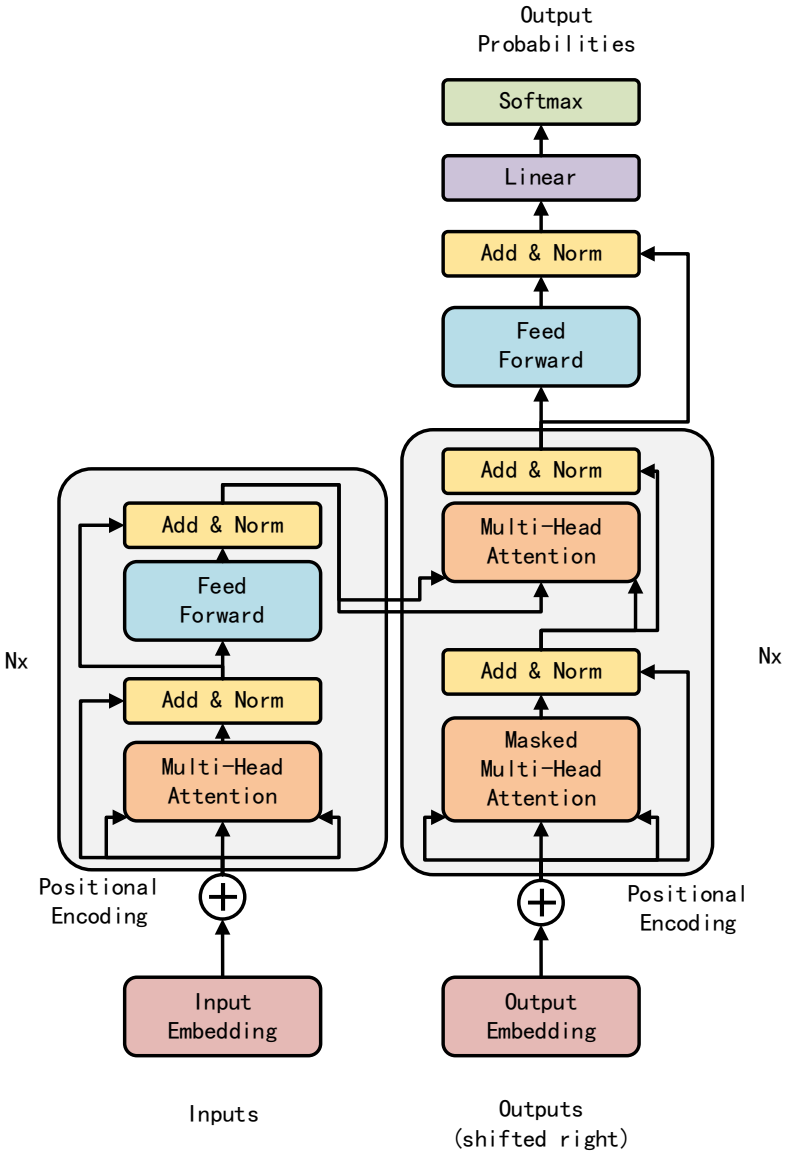

图 3-3 Bert 模型结构

Bert 模型结构如图 4-3 所示，每一个 encoder 层都是由 multi-head-Attention 层 +，Layer Normalization 层， feedforword 层+，Layer Normalization 层堆叠而来，堆叠的 encoder 层数决定了 Bert 模型的规模大小，本研究通过使用 Bert 模型在意图数据集上训练实现了判断用户意图是两种意图中的哪一种，如果是诊断意图，后续还会继续判断是诊断意图中的哪一种详细意图，原项目已经提供预训练模型，因此不需要再额外进行训练。

3.1.2 BiLSTM-CRF 模型

BiLSTM-CRF 模型结构如图 4-4 所示，模型实现代码可参考附录 D。其中 BiLSTM 模型是由两个方向相反的 LSTM 组成，使得模型可以充分考虑句中的上下文信息，捕获词语的前向和后向语义信息，输出每个词语对应于每个标签的得分概率，并将这些得分概率作为 CRF 层的输入。

CRF 层是一种无参数的概率图模型，通过获取 BiLSTM 的概率结果，计算不同词

语的转移概率值以及这些词语的关系，最终输出预测的标签序列。

当 Bert 意图模型判断出用户的对话意图为诊断意图，就会将对话文本传递给该模型，由该模型对对话文本中的病症名称进行提取，所有的病症名称都存储在 json 文件中，通过模型对对话文本进行编码提取，从存储疾病名称的 json 文件中匹配最相似的疾病命名实体，原项目已经提供预训练模型，因此也不需要再额外进行训练，如果需要添加新的疾病，只需要将疾病名称添加进存储所有疾病名称的 json 文件即可，不需要再对模型进行重新训练，大大地减轻了后续添加新数据的负担。

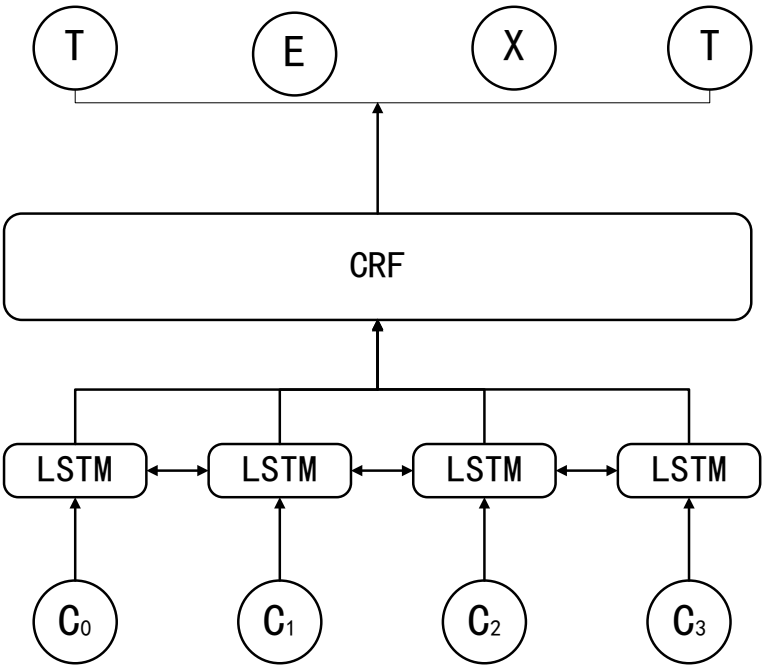

图 3-4 BiLSTM-CRF 模型结构

3.2 图数据库

该 KBQA 系统的图数据库结构如图 4-5 所示，通过构建 8 个实体型，包括疾病的名称，相关症状，在医院应该做检查的科室，需要做的检查，可以治疗的药品，药品常见的生产厂家，推荐和不推荐吃的食物，推荐的菜谱。

8 个实体型共有 11 种实体型关系，使得实体型互相关联，包括与疾病关联的症状、常见药物和特效药物、科室、检查、菜谱、相关疾病、推荐和不推荐吃的食物，大科室关联的下属科室，药企关联的生产药品。

通过以上的设计，实现对各种病症的知识图谱关系的构建，即可实现了解决大部分医疗相关的对话问题。

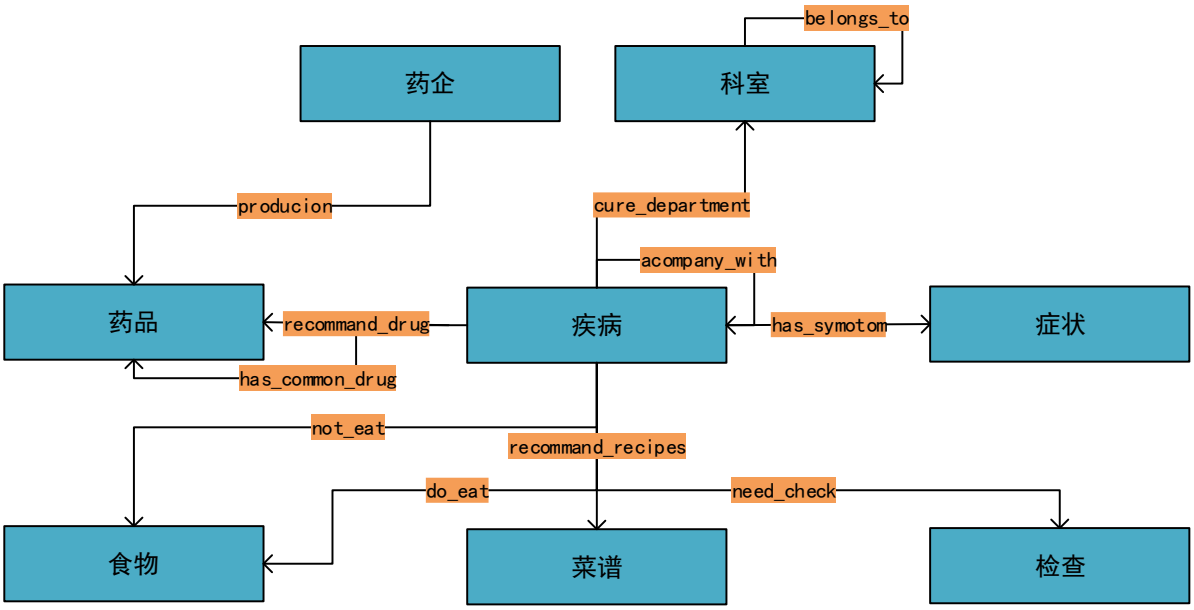

图 3-5 图数据库结构

3.3 图数据库数据扩充

本研究使用的图数据库是 Neo4j 图数据库管理系统，数据库在本地运行后，可以直接在网页上对数据进行基本数据库的各类基本操作，neo4j 图数据库交互界面如图 4-6 所示。

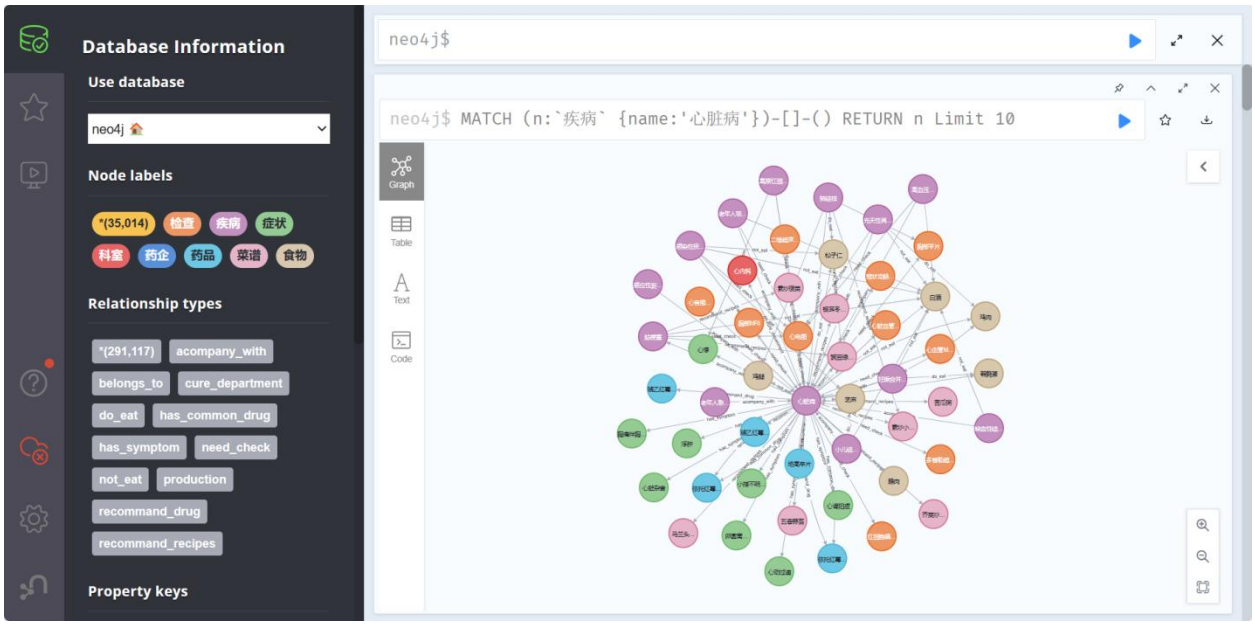

图 3-6 neo4j 图数据库交互界面

由于原有的知识图谱缺少中医舌诊相关的数据，无法针对舌象进行专业的解答，所以本研究在专业中医指导下对图数据库中的数据进行了扩充，新增了舌诊相关的数据，使得对话系统能够对中医舌诊相关问题进行解答，图数据库中舌诊知识图谱可视化如图 4-7 所示，可以清晰地看到添加的各类舌象与其相关联的各项数据。经过扩充后的图数据库中包含 35014 个实体，291117 个关系，涵盖了目前大部分的病症信息，图数据库的数据量越多，对话系统能回答的病症问题也就越多。

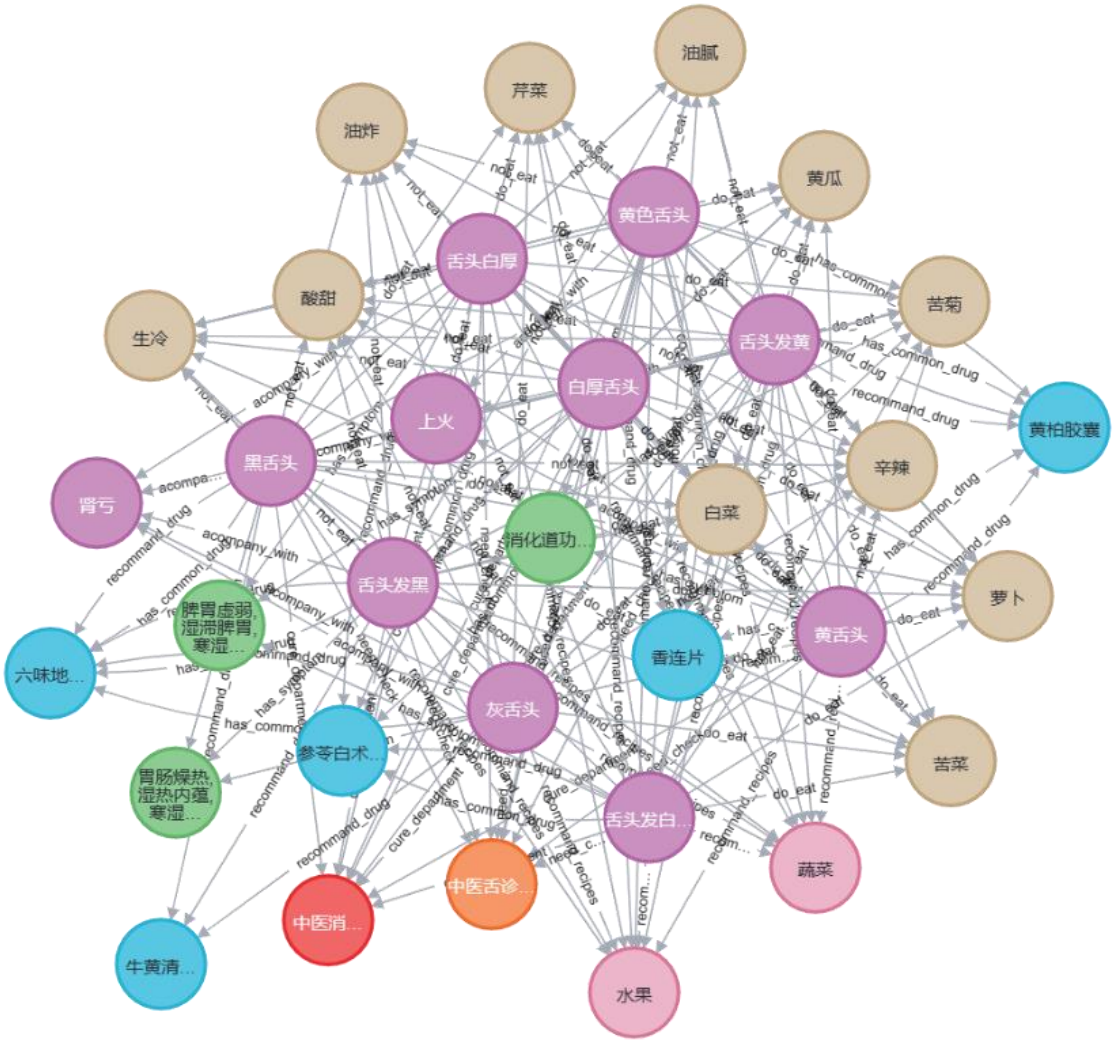

图 3-7 图数据库中舌诊知识图谱可视化

3.4 舌诊对话软件界面设计

直接使用模型对于大部分人来说是有较高的门口，需要学习大量相关知识才能够使用，为了提升用户对智能舌诊对话系统的使用便捷性，本研究深入研究了市面上常用的聊天软件，特别是中老年喜欢使用的即时通讯软件微信，从中汲取了其优秀的设计理念，并将这些元素融入了智能舌诊对话界面的设计中。如图 4-8 所示，这一界面以简洁、直观为核心设计理念，致力于为用户提供一个流畅、无障碍的舌诊体验。

当舌象图片成功上传后，系统的模型会立即对图像进行舌象判断和分析。这一过程通过前几章中介绍的舌象分类技术和中医舌诊理论实现，能够准确地识别出舌象类型，并以此对用户的舌象进行健康状况的初步判断。与此同时，通过与上一章的对话模型结合，实现了用户与智能舌诊模型进行问诊对话的功能。用户可以根据自己的需求，提出关于舌诊和健康问题的询问。对话模型会根据用户的问题和舌象信息，给出智能的回答和建议。这种交互方式不仅增强了系统的智能性，还使得用户能够更加深

入地了解自己的健康状况。

在界面的布局上，采用了清晰、直观的排版方式。各个功能区域划分明确，用户可以直观地找到所需的按钮和选项。同时，还注重了界面的美观性和舒适性，使得用户在使用过程中能够感受到愉悦和舒适。在界面的底部，设置了一个醒目的“上传图片”按钮。用户只需通过点击该按钮，就能快速选择并上传自己的舌象图片。这一步骤简单明了，极大地降低了用户的操作难度。

这样的设计旨在为用户提供一个用户友好的舌诊应用界面。通过简化操作流程、优化算法和增强交互性，可以使得用户能够轻松、快捷地进行舌诊诊断和与智能舌诊系统交流。界面的简洁性和直观性降低了用户的学习成本，使得不同年龄和背景的用户都能够轻松上手，享受到智能舌诊带来的便利和好处。

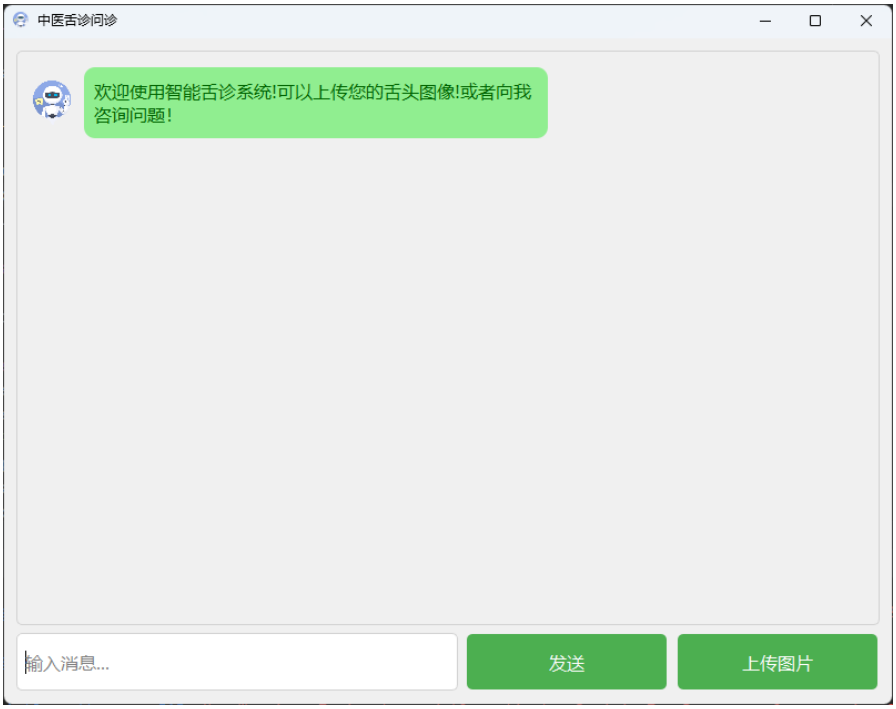

图 3-8 智能舌诊对话界面

3.5 本章小结

本章详细介绍了多模态舌诊对话系统的研究与实现。该系统结合了上一章提出的舌象分类模型与自然语言处理技术，为用户提供智能的舌诊服务。系统流程包括用户上传舌象图片，通过舌象分类模块将结果传输至 KBQA 智能医疗诊断对话系统，进而实现智能舌诊的多模态交互。本章还详细介绍并实现了专门的对话界面方便用户对话，为用户的智能舌诊提供更便利的服务。

模型好坏的评判，需要经过合理的实验验证，下一章将会针对各个模型进行多项标准严格的实验验证本研究提出的模型的优劣。

4 实验设计

4.1 实验环境

本研究使用的模型训练环境参照表 5-1，模型训练均在 torch=1.9.0+cu111 环境下训练。实验采用的 Python 版本为 3.7，其余依赖的包环境已在项目文件中的 requirements.txt 中写明，实验采用的显卡为 NVIDIA Geforce 2060(6G) ，如果显存低于 4G，将无法训练本研究使用的模型，运行内存为 32GB。

表 4-1 模型训练环境

项目	内容
操作系统	Windows 11 23H2
相关库版本	PyTorch 1.9.0 / cuda 11.1.0/ Python 3.7
图形处理器	NVIDIA Geforce 2060(6G)
中央处理器	Intel(R) Core(TM) i7-10875H
内存	RAM 32GB
存储设备	NVME 1TB

4.2 实验设置

为了确保训练参数对结果影响不会过大，以及最后评判的公平性，所有模型的训练 epoch 统一设置为 300 轮，为了防止因显存不够导致模型无法在统一参数下进行训练，Batch _size 大小统一设置为 4。所有模型训练测试的数据集均统一。YOLOv7 使用的预训练模型为 yolov7_training.pt。

4.3 评价指标

为了评价模型的性能，不能只考虑准确度的高低，防止因为样本不均匀导致的假高性能，还应该考虑模型判断的精确率和召回率，同时为了直观的对比性能高低，本研究还引入了 F1 分数，方便对比各个模型之间的性能差距。虽然本项目是多分类问题，依然可以依据混淆矩阵如表 5-2，计算各个评分标准下的性能得分。

表 4-2 混淆矩阵

真实类别	预测类别	
	正例	反例
正例	TP	FP
反例	FN	TN

准确度 A 是指分类正确的样本总数除以样本总数，由式（5-1）计算：

$$Accuracy = \frac{\mathrm{TP} + \mathrm{TN}}{TP + TN + FP + FN} \quad (5\text{-}1)$$

精确率 P 与召回率 R 分别如式（5-2）和式（5-3）所示：

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad (1\text{-}2)$$

$$Recall = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \quad (5\text{-}3)$$

F1 得分是对精确率和召回率的调和平均，由式（5-4）计算：

$$F1 = \frac{\mathrm{PR}}{\mathrm{P} + \mathrm{R}} \quad (5\text{-}4)$$

4.4 结果分析

4.4.1 分类方法性能比较结果

表 4-3 对于舌象不同分类方法的评价指标

方法	新数据集				旧数据集			
评价指标	Acc	P	R	F1	Acc	P	R	F1
Resnet18	76.25%	77.54%	78.58%	78.06%	69.91%	47.34%	60.12%	52.97%
VIT	75.00%	82.20%	74.11%	75.30%	72.03%	28.87%	33.87%	31.09%
YOLOv3	69.37%	83.67%	83.04%	83.36%	58.47%	67.04%	58.81%	62.65%
YOLOv5	83.75%	85.29%	85.12%	85.20%	72.88%	70.55%	65.20%	67.77%
YOLOv7	86.87%	88.92%	88.57%	88.74%	74.15%	68.37%	66.34%	67.34%
YOLOv8	86.87%	87.65%	86.32%	86.77%	66.94%	60.15%	52.92%	55.59%
YOLOv7CFE	92.50%	93.54%	93.03%	93.28%	76.27%	76.623%	60.18%	67.41%

对于舌象不同分类方法的评价指标，如表 5-3 所示，所有模型在原数据集上，表现的效果都很差，没有模型在评价指标下超过 80%的评分，在各项评价指标中都有好有坏，而使用本研究重新制作的数据集训练后，各个模型的各项评价指标也都有了显著提升，几乎都提升了 20%以上。并且本项目的模型在本研究中重新制作的数据集上训练后所有评价指标下都取得了最好的成绩。

在旧数据集上训练模型，整体得分较低，而在新数据集上训练的模型整体得分显著提高，这表明图像质量对模型性能影响较大，良好的数据集是训练优秀模型的关键。本研究中所创建的数据集质量相对较高。

尽管 ResNet18 和表现较差，但由于模型简单，资源消耗最少。YOLOv3 是早期的图像识别模型，其正确率远低于其他模型，因为对置信度的要求较高，因此精确率和召回率相对较高。VIT 作为基于 transform 的图像分类模型整体表现并不佳。而 YOLOv5、YOLOv7 和 YOLOv8 的模型结构仍在更新，它们之间的性能差距较小。由于测试集规模较小，各项评分基本相当，相对而言，YOLOv7 更适合本项目。因此，在 YOLOv7 的基础上引入 CEF 模型以提升性能。加入 CEF 模型后，各项评价指标有明显提升，这表明 CEF 模型在结果方面发挥了一定的作用。尽管加入 CEF 模型导致 YOLOv7 的检测速度减慢，但在舌象分类中，正确性比速度更为重要，因此用时间换取正确率是值得的。

4.4.2 不同分类下结果

表 4-4 本项目舌象分类模型对于舌象不同分类下的评价指标

分类	A	P	R	F1
红舌	94.67%	76.31%	100.00%	86.56%
薄白舌	94.37%	100.00%	74.28%	85.24%
厚白舌	98.15%	91.42%	100.00%	95.52%
黄舌	98.75%	100.00%	94.44%	97.14%
黑灰舌	99.37%	100.00%	96.42%	98.18%

本项目舌象分类模型对于舌象不同分类下的评价指标，如表 5-4 所示，本研究的模型在各个舌象分类下准确率都已高达 96%以上的准确率，精确率得分最高 100%，最低 81%。召回率得分最高 93%，最低 79%。

模型在不同舌象分类下表现良好，尤其在厚白舌和黑灰舌分类上性能较为优异。但由于红舌和薄白舌的色相相近，CEF 模型在这两类舌象的分类上表现不佳，导致对 YOLO 输出的结果限制不强。

4.4.3 舌诊对话系统性能评估实验结果

图 4-1 对话系统使用实例

经过对话实验，对话系统使用实例在图 5-1 中可以清晰地观察到，KBQA 系统通过扩充舌诊知识，对话系统能够更全面地回答用户提出的舌诊相关问题。此外，系统还具备实现多轮对话的能力，并且在回答问题时不会产生无法预料的错误回答，这进一步提升了系统的智能程度和用户体验。

值得注意的是，随着图数据库知识的不断扩充，对话系统对于疾病相关问题的回答将变得更加全面。而且对系统进行新知识的增加也无需重新训练模型，这为系统的持续优化和知识更新提供了便利。

因此，通过以上实验结果表明，KBQA 系统在智能对话方面有着较高的潜力，具有一定的智能性，并且具备了简单方便的持续进化和优化的能力。