第 5 章 基于模型后处理的未知意图检测方法

5.1 引　　言

在本书的第一篇中详细介绍了对话未知意图发现的研究进展，发现未知意图的第一步是将已知意图与未知意图进行分离，使得分类器在正确识别已知意图的同时，也能识别超出其处理范围的未知意图样本。然而，现有方法皆需要对深度神经网络模型结构做一定的调整，才能进行未知意图检测。未知意图检测性能在很大程度上取决于分类器是否可以有效地对已知意图进行建模，而传统的分类模型（如支持向量机）对意图的高阶语义概念进行建模的能力有限，导致算法性能不佳。

为了解决上述问题，本章介绍了基于模型后处理的未知意图检测方法，该方法可使分类器模型具备未知意图检测的能力，而无须对模型结构进行任何修改，也可应用于任何深度神经网络分类器，充分利用其强大的特征提取能力来提高未知意图检测算法的性能。方法主要分为两部分：第一是通过所提出的 SofterMax 激活函数，对分类器输出的样本置信度进行校准，以获得合理的概率分布；第二是深度新颖检测模块，将深度神经网络学习的意图表示和基于密度的异常检测算法结合，进行检测。最后，将上述两部分所计算出的分数，通过 Platt Scaling 转换成概率，进行联合未知意图预测。在缺乏关于未知意图的先验知识及样本的情况下，所提出的方法仍然可以检测未知意图。

本章其余部分的安排如下。5.2 节介绍基于模型后处理的未知意图检测方法（SMDN），并对其子模块进行详细描述，包括所提出的 SofterMax 激活函数和深度新颖检测模块；5.3 节描述实验数据集、评价指标、实验结果与分析；5.4 节为本章小结。

5.2 基于模型后处理的未知意图检测方法

在本节中，将对基于模型后处理的未知意图检测方法 SMDN（SofterMax and Deep Novelty Detection）进行详细描述。首先，基于深度神经网络来训练一个已知意图分类

器。接着，通过温度缩放[98]来校准分类器的输出置信度，并收紧经过校准的 Softmax（SofterMax）激活函数的决策边界，以更好地检测未知意图。此外，将深度神经网络学习到的意图表示输入基于密度的异常检测算法，从不同角度来检测未知意图。最后，通过 Platt Scaling[99]将 SofterMax 的置信度分数和局部异常因子的新颖分数转换为新颖概率，并进行联合预测。

5.2.1 基于深度神经网络的意图分类器

后处理方法的关键思想是：在不修改模型结构的情况下，基于现有的意图分类器来检测未知意图。如果一个样本不同于所有已知的意图，将被视为未知意图。因为是基于已知意图分类器来检测未知意图，所以分类器的性能至关重要，分类器的性能越好，未知意图检测的效果就越好。因此，必须先实现接近最佳性能的单轮和多轮对话意图分类器，并在相同分类器下比较不同未知意图检测方法的性能。

1. 基于双向长短期记忆网络的单轮意图分类器

由于长短期记忆网络只考虑由左到右的序列输入，在对句子建模时可能会丢失由右到左的序列信息。双向长短期记忆网络（BiLSTM）通过前向与后向的长短期记忆网络，同时对由左到右和由右到左的序列输入建模，很好地弥补了这点不足。如图 5.1 所示，首先使用双向长短期记忆网络来建模单轮对话的意图，并将其用于后续的未知意图检测任务。

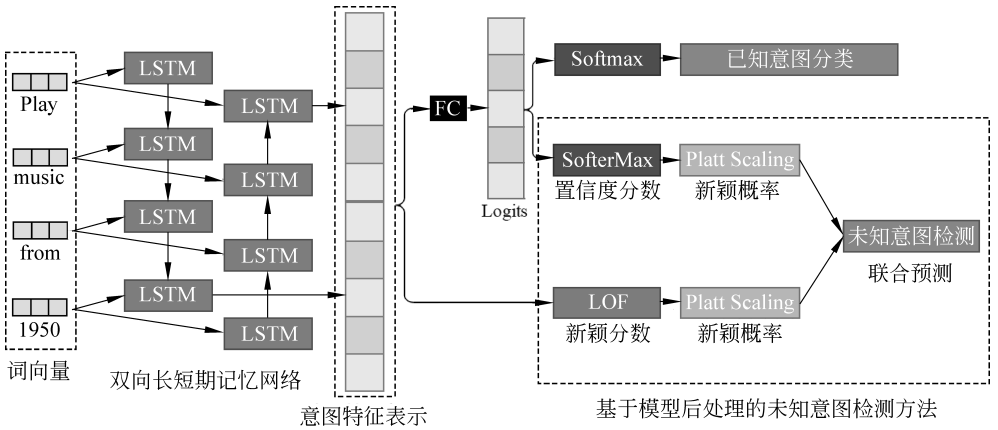

图 5.1 基于双向长短期记忆网络的单轮未知意图检测方法

给定一个输入语句及其最大序列长度 ℓ，将其单词序列 $\boldsymbol{w}_{1:\ell}$ 转换为 m 维的词向量序列 $\boldsymbol{x}_{1:\ell}$，输入到 BiLSTM 中获取意图表示 h：

$$\overrightarrow{\boldsymbol{h}_t}=\mathrm{LSTM}(x_t,\overrightarrow{c_{t-1}}) \tag{5-1}$$

$$\overleftarrow{\boldsymbol{h}_t}=\mathrm{LSTM}(x_t,\overleftarrow{c_{t+1}}) \tag{5-2}$$

$$\boldsymbol{h}=[\overrightarrow{\boldsymbol{h}_l};\overleftarrow{\boldsymbol{h}_1}],z=\boldsymbol{W}\boldsymbol{h}+b \tag{5-3}$$

其中，$\boldsymbol{x}_t\in\mathbf{R}^m$ 表示在时间步 t 时刻输入的 m 维词向量；$\overrightarrow{\boldsymbol{h}_t}$ 和 $\overleftarrow{\boldsymbol{h}_t}$ 分别是前向和后向 LSTM 的输出隐状态；$\overrightarrow{\boldsymbol{c}_t}$ 和 $\overleftarrow{\boldsymbol{c}_t}$ 分别是前向和后向 LSTM 的细胞状态。然后将前向 LSTM 的最后一个输出隐状态 $\overrightarrow{\boldsymbol{h}_l}$ 和后向 LSTM 的第一个输出隐状态 $\overleftarrow{\boldsymbol{h}_1}$ 视为句子表示，并将其拼接为意图表示 $\boldsymbol{h}$。在通过 Softmax 激活函数之前，logits $\boldsymbol{z}$ 是全连接层的输出，输出神经元的数量等于已知类的数量。最后将 h 作为深度新颖检测的输入，并将 $\boldsymbol{z}$ 作为 SofterMax 的输入。

2. 基于层次卷积神经网络的多轮意图分类器

在建模多轮对话意图时，当前用户的意图与其上下文有很强的关联性，而层次卷积网络可以对多个句子进行卷积操作，在建模意图时考虑多个句子，很好地弥补了这点不足。通过第一个卷积神经网络对句子进行建模，再通过第二个卷积神经网络对当前句子以及其上下文进行建模，以获得带上下文的多轮意图表示。如图 5.2 所示，使用层次卷积网络来建模多轮对话意图，并将其用于后续的未知意图检测任务。

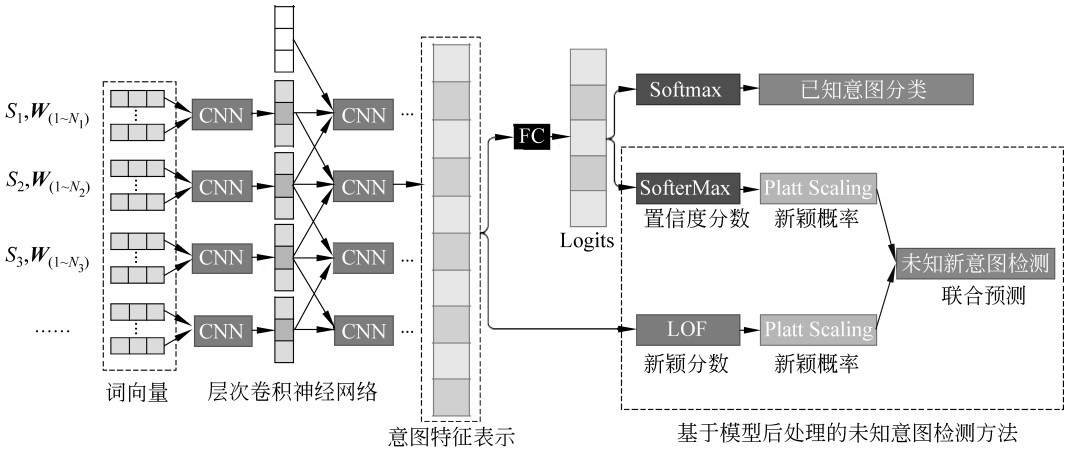

图 5.2　基于层次卷积网络的多轮未知意图检测方法

通过第一个卷积神经网络获得句子表示 $\boldsymbol{z}$ 后，即可进一步在窗口大小为 c 的句子上执行第二次卷积运算，生成目标句子带上下文的意图表示，计算过程如下所示：

$$\boldsymbol{Z}=[\boldsymbol{z}_{t-c-1},\cdots,\boldsymbol{z}_{t-1},\boldsymbol{z}_t,\boldsymbol{z}_{t+1},\cdots,\boldsymbol{z}_{t+c-1}] \tag{5-4}$$

$$\boldsymbol{h}_t = \mathrm{ReLU}(\boldsymbol{W}_f \cdot \boldsymbol{Z}_{t:t+n-1} + b_f) \tag{5-5}$$

其中，$\boldsymbol{Z} \in \mathbf{R}^{(2c-1)\times k_1}$ 表示对话中第 t 个句子的带上下文窗口大小 c 的意图表示；k_1 为句子级卷积核的个数；$\boldsymbol{h}_t$ 为卷积核 $\boldsymbol{W}_f \in \mathbf{R}^{n\times k_1}$ 在大小为 n 的连续窗口中执行卷积运算所产生的特征图。使用卷积核 $\boldsymbol{W}_f$ 在所有可能的句子窗口上进行卷积操作后，即可生成多个特征图 $\boldsymbol{h}$。

$$\boldsymbol{h} = [\boldsymbol{h}_{t-c-1}, \cdots, \boldsymbol{h}_{t-1}, \boldsymbol{h}_t, \boldsymbol{h}_{t+1}, \cdots, \boldsymbol{h}_{t+c-1}] \tag{5-6}$$

其中，$\boldsymbol{h} \in \mathbf{R}^{2c-1}$。通过在特征图上进行最大池化操作，即可获得卷积核 $\boldsymbol{W}_f$ 在特征图 $\boldsymbol{h}$ 上的有效特征 $\hat{\boldsymbol{h}}$：

$$\hat{\boldsymbol{h}} = \max\{\boldsymbol{h}\} \tag{5-7}$$

其中，$\hat{h}$ 是通过 W_f 学习到的标量特征。最后，通过 k_2 个上下文卷积核进行卷积运算，获得带上下文的意图表示 $\boldsymbol{r}$：

$$\boldsymbol{r} = [\hat{\boldsymbol{h}}_1, \hat{\boldsymbol{h}}_2, \cdots, \hat{\boldsymbol{h}}_{k_2}], \quad z = \boldsymbol{W}\boldsymbol{r} + b \tag{5-8}$$

其中，$\boldsymbol{r} \in \mathbf{R}^{k_2}$ 代表目标句子 k_2 维的带有上下文的意图表示。Logits $\boldsymbol{z}$ 是通过 Softma$\boldsymbol{x}$ 激活函数之前的全连接层的输出，其中神经元的数量等于已知类的数量。与 BiLSTM 模型相似，将 $\boldsymbol{r}$ 作为深度新意图检测模块的输入，并将 $\boldsymbol{z}$ 作为 SofterMax 的输入。如图 5.1 和图 5.2 所示，所提出的方法可以灵活地应用于各种深度神经网络分类器。

5.2.2 SofterMax 激活函数

在意图分类器的基础上，校准 Softmax(即 SofterMax)输出的置信度以获得更合理的概率分布，并收紧 SofterMax 的决策边界以拒绝未知样本。在图 5.3 中，可以看到 Softmax 和 SofterMax 之间的区别。DOC[37]表明，通过减少概率空间中的开放空间风险

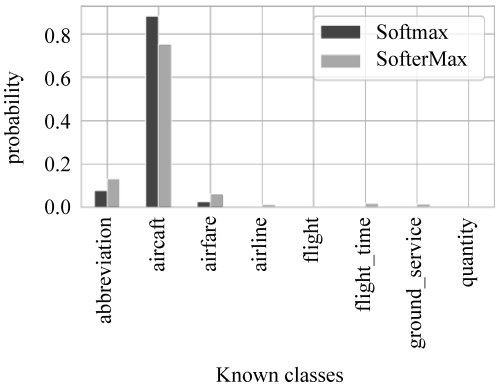

图 5.3 SofterMax 的效果示意图(与 Softmax 相比，SofterMax 输出的概率分布更为保守)

来拒绝未知类的样本是可行的。而深度神经网络分类器的 Softmax 输出概率倾向过于自信，暴露了太多的开放空间风险，将属于未知类的样本以高置信度被错误地分类为已知类，无法提供合理的后验概率分布。这是因为交叉熵损失函数在优化过程中将目标类别的预测概率最大化，并将其他类别的预测概率最小化，从而导致其他类别的输出概率接近零。这对计算决策阈值来检测未知意图的方法而言，是很不理想的性质。Hinton 等人在知识蒸馏任务中提出了温度缩放方法[98]，并将其应用于生成神经网络输出的软标签（即提高熵）。通过温度缩放来软化 Softmax 的输出，使模型的输出概率更加保守，进而降低开放空间风险，提高未知意图检测算法的性能。

1. 温度缩放

给定一个 N 分类的神经网络和输入 $\boldsymbol{x}_i$，以及网络的输出预测 $\hat{y}_i = \text{argmax}_n(\boldsymbol{z}_i)$，$\boldsymbol{z}_i$ 是网络的 logits 向量，Softmax 函数 σ_{SM} 和置信度分数 $\hat{p}_i$ 计算如下：

$$\sigma_{\text{SM}}(\boldsymbol{z}_i) = \frac{\exp(\boldsymbol{z}_i)}{\sum_{j=1}^{N}\exp(\boldsymbol{z}_i^{(j)})} \tag{5-9}$$

$$\hat{p}_i = \max_n \sigma_{\text{SM}}(\boldsymbol{z}_i) \tag{5-10}$$

通过将温度缩放应用于 Softmax 函数的概率输出后，定义 SofterMax $\hat{\sigma}_{\text{SM}}$ 和软化的置信度分数 $\hat{q}_i$ 如下：

$$\hat{\sigma}_{\text{SM}}(\boldsymbol{z}_i) = \sigma_{\text{SM}}(z_i/T) \tag{5-11}$$

$$\hat{q}_i = \max_n \hat{\sigma}_{\text{SM}}(\boldsymbol{z}_i) \tag{5-12}$$

其中，T 是温度参数。当 T 等于 1 时，σ_{SM} 是 $\hat{\sigma}_{\text{SM}}$ 的特例；当 $T>1$ 时，会产生更保守的概率分布；当 T 接近无穷大时，概率 $\hat{q}_i$ 接近 $\frac{1}{N}$ 并退化为均匀分布，这意味着熵达到最大值。

如何在狭窄的范围内选择合适的 T 是至关重要的[98]。温度参数 T 是一个经验性的超参数。在一般情况下，很难获得未知意图的样本，无法直接通过验证集对 T 进行调整。为了获得合适的 T，通过温度缩放[100]对模型执行概率校准，在缺乏未知意图样本的情况下，自动优化获得最佳温度参数 T。

2. 概率校准

当一个模型校准后的置信度接近其真实似然值时，认为该模型是经过良好校准的。通过温度缩放校准的输出概率，并将原始的输出置信度 $\hat{p}_i$ 转换为校准后的置信度 $\hat{q}_i$。给定标签的独热表示 t 和模型预测 y，样本的负对数似然表示如下：

$$\mathcal{L} = -\sum_{j=1}^{N} t_j \log\ y_j \tag{5-13}$$

通过验证集中的负对数似然，优化获得 SofterMax $\hat{\sigma}_{\mathrm{SM}}$ 的最佳温度参数 $\hat{T}$，以实现概率校准。如图 5.3 中所示，SofterMax 对于所有类别都保持相对保守的输出概率分布。T 在训练期间被设置为 1，在测试过程中被设置为 $\hat{T}$。此外，概率校准不会影响已知意图的预测结果。

3. 决策边界

通过计算每个类 c_i 的概率阈值，进一步降低概率空间中的开放空间风险，缩紧 SofterMax 输出的决策边界。首先。计算每个类的 $p(y=c_i \mid x_j, y_j=c_i)$ 的均值 μ_i 和标准差 σ_i，其中 j 表示 j 个样本。计算每个类 c_i 的概率阈值 t_i 如下：

$$t_i = \max\{0.5, \mu_i - \alpha\sigma_i\} \tag{5-14}$$

这里直观的解释是：如果样本的输出概率分数偏离平均值的 α 个标准差，则将其视为离群值。如果样本每个类别 c_i 的 SofterMax 输出置信度皆低于概率阈值 t_i，则将该样本视为未知意图。

为了比较不同样本之间的置信度，必须为每个样本计算一个可比较的置信度分数。从校准的置信度分数中减去每个类别的概率阈值 t_i，并取其最大值作为样本的单一、可比较的置信度分数。置信度分数越低，则样本就越可能为离群值。如果置信度分数低于 0，则认为该样本属于未知意图。对于每个样本，将其多个类别的置信度分数转换为单一置信度分数如下。

$$\mathrm{confidence}_{j,i} = p_{j,i} - t_i \tag{5-15}$$

$$\mathrm{confidence}_j = \max_i(\mathrm{confidence}_{j,i}) \tag{5-16}$$

由于 Softmax 是非线性变换，Softmax 经过温度缩放后的 logits 与原始 logits 为非线性相关。因此，当把相同的概率阈值方法应用于校准后的置信度分数时，即可获得不同的未知意图检测结果。SofterMax 在 Softmax 的基础上对 logits 进行温度缩放，并把输出概率减去每个类别的概率阈值，取其最大值作为置信度分数。后续将通过置信度分数来进行联合预测。

5.2.3 深度新颖检测模块

本节进一步将新颖性检测算法与深度神经网络学习的意图表示结合，以从不同角度检测未知意图。

OpenMax[36] 表明，减少特征空间中开放空间风险可以提高未知意图检测的性能。与 OpenMax 使用 logits 作为特征空间不同，将 logits 之前的隐层向量表示作为特征空间，并将其作为深度新颖检测算法的输入。由于此特征空间的维度远远大于 logits，因此其特征向量表示可包含比 logits 更多的高级语义概念。

接着，通过局部异常因子（Local Outlier Factor，LOF）[18]减少特征空间中的开放空间风险并发现未知意图。LOF 是一个基于密度的异常检测方法，通过计算局部密度来检测局部上下文中的未知意图。局部异常因子的计算如下。

$$\mathrm{LOF}_k(\boldsymbol{A})=\frac{\sum_{\boldsymbol{B}\in N_k(\boldsymbol{A})}\frac{\mathrm{lrd}(\boldsymbol{B})}{\mathrm{lrd}(\boldsymbol{A})}}{|N_k(\boldsymbol{A})|} \tag{5-17}$$

其中，$N_k(\boldsymbol{A})$表示 k 个最近邻的集合；lrd 为局部可达性密度，定义如下：

$$\mathrm{lrd}_k(\boldsymbol{A})=1/\left(\frac{\sum_{\boldsymbol{B}\in N_k(\boldsymbol{A})}\mathrm{reachdist}_k(\boldsymbol{A},\boldsymbol{B})}{|N_k(\boldsymbol{A})|}\right) \tag{5-18}$$

lrd 为目标 $\boldsymbol{A}$ 及其邻居之间可达距离的平均倒数。可达距离 $\mathrm{reachdist}_k(\boldsymbol{A},\boldsymbol{B})$的定义如下：

$$\mathrm{reachdist}_k(\boldsymbol{A},\boldsymbol{B})=\max\{k-\mathrm{distance}(\boldsymbol{B}),d(\boldsymbol{A},\boldsymbol{B})\} \tag{5-19}$$

其中，$d(\boldsymbol{A},\boldsymbol{B})$表示 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 之间的距离；k-distance 表示对象 $\boldsymbol{A}$ 到 k 个最近邻居的距离。如果样本的局部密度显著低于其 k 最近邻居的局部密度，则该样本更有可能被视为未知意图。然后将 LOF 分数视为新颖分数。新颖分数越高，则该样本越有可能属于未知意图。

5.2.4　Platt Scaling 联合预测

最后，将 SofterMax 的结果与深度新颖性检测算法结合，进行联合预测。由于 Softmax 计算的置信度分数和 LOF 的新颖分数的度量方式不同，无法直接合并，因此，使用 Platt Scaling[99]将分数统一转换为 0～1 的概率，以进行联合预测。

Platt Scaling 最初应用于支持向量机，目的是将样本到决策边界的距离转换为分类概率。Platt Scaling 通过逻辑回归模型来将分数映射为概率，计算过程如下：

$$P(y=1\mid\boldsymbol{x})=\frac{1}{1+\exp(\boldsymbol{A}f(\boldsymbol{x})+\boldsymbol{B})} \tag{5-20}$$

其中，$f(\boldsymbol{x})$表示分数。$\boldsymbol{A}$ 和 $\boldsymbol{B}$ 是通过算法学习的参数。Platt Scaling 的做法是：让决策边界附近的样本有 50%的概率被视为未知意图，并将剩余样本的分数缩放为介于 0～1 的概率。通过该方法归一化后，可以在相同的度量标准下估计 SofterMax 和 LOF 的新颖程度，从而进行联合预测。这种方法可以视为简单平均的集成学习策略，通过两种不同模型的视角进行检测，降低预测方差的同时提升算法的鲁棒性。

5.3 实　　验

实验分为 3 个部分介绍，包括任务与数据集、实验设置以及实验结果与分析。

5.3.1 任务与数据集

本章解决的任务是未知意图检测，目标是在正确识别出 n 类已知意图的同时，检测出不属于任何已知意图的未知意图样本。因此，将任务定义为 $n+1$ 分类，其中第 $n+1$ 类即为未知意图。为了研究该方法的鲁棒性和有效性，在 SNIPS、ATIS[95] 和 SwDA[101] 3 个公开的对话数据集上对其实验验证。表 5.1 是 SNIPS、ATIS 和 SwDA 数据集的统计信息。

表 5.1　SNIPS、ATIS 和 SwDA 数据集的统计信息

数据集	类别数	词表大小	训练集	验证集	测试集	轮次	数据分布
SNIPS	7	11 971	13 084	700	700	单轮	均衡
ATIS	18	938	4 978	500	893	单轮	不均衡
SwDA	42	21 812	162 862	20 784	20 146	多轮	不均衡

SNIPS 首先在 SNIPS 个人语音助手数据集上实验。该数据集包含 7 种不同领域的用户意图，例如播放音乐、询问天气、预订餐厅等，总共有 13 084 条训练数据、700 条验证数据和 700 条测试数据。每个类别中的样本数量相对均衡。

ATIS(Airline Travel Information System)航空公司旅行信息系统是对话意图研究中最经典的数据集，包含 18 种航空领域的用户意图，总共有 4978 条训练数据、500 条验证数据和 893 条测试数据。ATIS 中的类别高度不均衡，其中前 25%的类别占了训练集数据 93.7%的样本。

SwDA(Switchboard Dialog Act Corpus)是最经典的多轮对话数据集，包含 1155 条两人电话交谈记录，每组聊天内容皆围绕着特定主题，共有 42 种对话动作。用户意图可视为抽象版的对话动作，在这种情况下，需要验证现有的未知意图检测方法是否仍然有效。原始 SwDA 数据集并没有将其切分为训练、验证和测试集，遵循先前研究中[102]建议的数据切分方案，随机抽取 80%的对话数据作为训练集，10%的对话数据作为验证集，10%的对话数据作为测试集，总共有 162 862 条训练数据、20 784 条验证数据和 20 146 条测试数据，每组对话平均包含 176 个句子。此外，SwDA 中的类别高度不平衡，其中前 25%的类别约占 90.9%的训练集。

5.3.2　实验设置

本节将对实验设置进行详细介绍，包括基线方法、评价指标和模型超参数设定。采用与先前研究[31,37]相同的交叉验证设置，将数据集中的部分类别设置为未知意图，被视为未知意图的样本将不会参与模型训练，并将从训练和验证集中删除。将训练集内的25%、50%和75%的类别设置为已知意图，并使用所有类别进行测试。使用100%类别即为常规的意图分类任务。

为了证明该方法的分类器架构可以很好地对意图建模，在表5.2中报告了使用100%类别进行训练的分类结果，并与在该数据集上性能表现最优的模型进行比较。实验表明，该方法实现了接近最佳性能的对话意图分类器。后续实验将以这些意图分类器为基础，进行未知意图检测任务。

表 5.2　在所有类别都已知情况下的分类器性能　%

数据集	模型	原始准确率	复现准确率	Macro-F1
SNIPS	BiLSTM	97	97.43	97.47
ATIS	BiLSTM[28]	98.99	98.66	93.99
SwDA	层次 CNN[102]	78.45	77.44	50.09

此外，为了在类别不均衡的数据集上进行公平的评估，实验时使用加权随机不放回抽样法，来随机选择每次实验的已知意图。如果类别拥有更多样本，则更有可能被选为已知类别，而样本较少的类别仍然有机会以一定的概率被选中，其他未被选中的类别将被视为未知意图。最终报告所有实验运行10次的平均结果。

1. 基线方法

本章将所介绍的方法与其他的未知意图检测方法进行比较，包含了简单阈值、最先进的方法以及其变体方法。

1）Softmax($t=0.5$)

在Softmax输出上设置概率阈值作为最简单的基线方法，并将概率阈值设置为0.5。如果样本在每个类别的输出概率皆不超过0.5，则将该样本视为未知意图。

2）DOC[37]

DOC是目前在此类问题上效果最好的方法，通过把输出层激活函数设置为Sigmoid，再利用统计方法来计算每个类别的概率阈值，进一步缩紧其决策边界。如果样本在每个类别的输出概率皆不超过其概率阈值，则将该样本视为未知意图。

3）DOC(Softmax)

DOC的变体方法，用Softmax代替了Sigmoid激活函数。在相同意图分类器下，对

所有检测方法进行评估，以示公平。

2. 评价指标

使用 Macro-F1 作为评价指标，并对所有意图、已知意图和未知意图的分类结果进行评估。主要关注于未知意图检测的结果。给定一组类别 $C=\{C_1,C_2,\cdots,C_N,\}$，Macro-F1 分数计算如下：

$$\text{Macro-F1}=2\frac{\text{recall}\times\text{precision}}{\text{recall}+\text{precision}} \tag{5-21}$$

$$\text{precision}=\frac{\sum_{i=1}^{N}\text{precision}_{C_i}}{N},\text{recall}=\frac{\sum_{i=1}^{N}\text{recall}_{C_i}}{N} \tag{5-22}$$

$$\text{precision}_{C_i}=\frac{\text{TP}_{C_i}}{\text{TP}_{C_i}+\text{FP}_{C_i}},\text{recall}_{C_i}=\frac{\text{TP}_{C_i}}{\text{TP}_{C_i}+\text{FN}_{C_i}} \tag{5-23}$$

其中，C_1 代表类别 C 中的单个类别；Macro-F1 指标计算每个类别的精确率和召回率的调和平均数，其物理意义是对模型在精确率和召回率之间的最优平衡点，当模型的精确率和召回率越高，则 Macro-F1 越高。

对于概率校准，使用期望校准误差[103]（ECE）来评估温度缩放的有效性。主要思想是将置信度输出划分为大小相等的间隔 K 个箱子，并计算这些箱子的置信度和准确率之间差异的加权平均值。ECE 计算如下：

$$\text{ECE}=\sum_{i=1}^{K}P_{(i)}*\mid o_i-e_i\mid \tag{5-24}$$

其中，$P_{(i)}$ 表示落入第 i 箱的所有样本的经验概率；o_i 代表第 i 箱中正样本的占比（准确率）；e_i 是第 i 箱的平均校准后置信度。ECE 越低，代表模型校准得越好。

3. 超参数设置

使用 GloVe[14] 预训练词向量（包含 40 万个词，输出向量维度为 300 维）来初始化分类器的嵌入层，并在训练中通过反向传播进一步优化。对于 BiLSTM 模型，将隐状态输出维度设置为 128，dropout 率设置为 0.5。对于层次 CNN 模型中的句子 CNN 和上下文 CNN，将其上下文窗口大小设置为 3，内核大小设置为 1 到 3，卷积核特征图数量设置为 100。使用学习率为 0.001 的 Adam 优化器。对于 SNIPS 和 ATIS，最大训练迭代次数为 30 次；对于 SwDA，最大训练迭代次数为 100 次。对于 SNIPS 和 ATIS，将批处理大小设置为 128；对于 SwDA，将批处理大小设置为 256。将 ATIS、SNIPS 和 SwDA 的最大输入序列长度分别设置为 35、46 和 58。本章使用 Keras 框架实现所有的模型。