基于动态记忆归纳网络的小样本文本分类

本篇论文是北京阿里巴巴集团和加拿大皇后大学（Queen University）共同撰写的论文，发表于 2020 年 ACL 的顶级会议上

摘要部分

本篇论文提出的针对小样本文本分类的动态记忆归纳网络（Dynamic Memory Induction Networks-DMIN），这个模型利用动态路由为基于记忆的小样本学习提供更多的灵活性，以此更好的适应 support sets，而这对于小样本分类模型来说是至关重要的能力。在此基础上，本篇论文进一步研发了查询信息的归纳模型，目的是增强 meta-learning 的泛化能力。所提出的模型在 miniRCV1 and ODIC 数据集上获得了最新最好的结果，将此前的最好的效果提升了 2%-4%，进一步进行详细分析以显示每个组件的有效性

1. 总体介绍

小样本文本分类，要求模型在有限数量的训练实例执行分类，对于许多应用来说这是很重要的但是也是存在挑战的任务，早期的关于小样本学习的研究，采用数据增强和规则化技术来缓解由于数据稀疏性引起的过渡拟合问题，更多最近的研究利用 meta-learning 在 meta 片段中的 meta 任务之间抽取可转移的知识（transferable knowledge）。

对于小样本文本分类的一个关键性挑战是，从 support sets 中归纳出类级别的特征，在 meta-tasks 切换的过程中关键的经常会丢失掉。最近的解决方案是利用一个记忆组件来维持模型的学习经历，例如：通过从受监督阶段找到与未见过的类型相似的内容，从而获得一个不错的结果，但是这些记忆部分所占权重在推断期间是静态的，当面对一个新的类型的时候，模型的所展现能力仍然是有限的。另外一个突出的挑战是由于各种原因导致实例级别的多样性，从而难以为一个类找到一个固定的原形。最近的研究表明，模型可以从查询感知方法中获益。

本篇论文，我们提出动态记忆归纳网络(DMIN)来进一步解决上述挑战。DMIN 利用动态路由的功能，通过利用路由组件在训练中或者训练后自动调整耦合系数的能力，来为基于记忆的小样本学习提供更大的灵活性，从而更好的适用 support sets。在此基础上，我们进一步开发带有查询信息的归纳组件,来识别在 support sets 中的各种实例与查询更相关的样本向量，这两个模块是在 DMIN 中共同学习的。

所提出的模型在 miniRCV1 和 ODIC 数据集上获得了比较好的结果，将之前的最佳性能提高了 2%-4%的精确度。我们将展示更细节的分析，来进一步展示所提出的网络是如何实现改进的。

2.相关工作

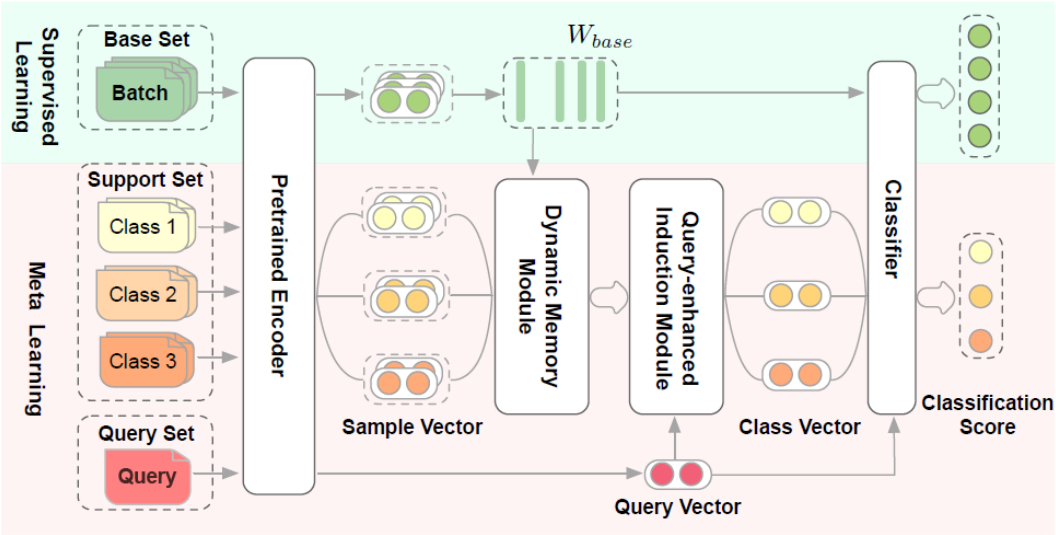
小样本学习在较早期的工作已经进行了研究，例如（……）并且还有更多最近的工作如（……），研究者也在各种 NLP 任务中进行了小样本学习的研究工作如（……），也包括文本分类如（……）。

记忆机制在许多 NLP 任务中都取得了不错的效果，如（……），在小样本的学习场景下，研究人员在每一个 meta episode 中应用记忆网络来存储编码后的上下文信息，如（……），尤其是（Qi 2018）and Gidaris and Komodakis（2018）构建一个两阶段训练步骤同时，把监督学习的类特征作为记忆的组成部分。

3.动态记忆归纳网络

3.1 总体架构

动态记忆归纳网络（DMIN）的架构图如下图所示



这是在构建于[Gidaris 和 Komodakis(2018)年]两阶段小样本学习框架之上，

在监督学习阶段(上部，绿色子图)，在训练数据选择部分类别作为基本数据集，由 C_{base}个基本类组成，这些被用来微调预训练句子编码器和训练分类器。

在 meta-learning 阶段(底部，橙色子图)，我们构造一个“片段(episode)”来计算梯度(gradients)，用来在每次训练迭代中更新我们的模型，对于C-way K-shot 问题，一个训练的片段（episode）是由随机从训练集中选择C个类型，并且从选择的C个类型的每一个类型中，选择 K 个样本来组成。用来当做 support sets $S = U_{c=1}^{C}\{x_{c,s}, y_{c,s}\}_{s=1}^{K}$ 。剩下样本的子集作为 query set $Q = \{x_q, y_q\}_{q=1}^{L}$。将 support set S 输入到模型中，然后更新它的参数以达到在 query set Q 中的最小化的 loss 来进行片段（episodes）的训练

3.2 预训练编码器

我们期望形成的小样本文本分类器应该得益于预训练模型的最新进展（Peters 2018，Devlin 2019，Radford），与最近的（Geng2109）研究工作不同的是，我们使用基于 BERT 对句子进行编码，该模型已经用于最近小样本学习模型（Bao 2109，Soares 2019）。（Devlin2019）BERT 模型架构模型是基于原始 Transformer 模型（Vaswani 2017）的一个多层双向的 Transformer 编码器。

将特殊分类嵌入（[CLS]）作为第一个标记，并添加特殊标记（[SEP]）作为最终标记。我们使用从[CLS] 输出的 d 维隐藏向量作为一个给定文本$\boldsymbol{x}$的特征$\boldsymbol{e}$:$\boldsymbol{e} = \boldsymbol{E}(\boldsymbol{x}|\boldsymbol{\theta})$。这预训练 BERT 模型提供了一个强大的上下文依赖的句子特征，可以用于各种目标任务，并且适应于小样本文件分类任务(Bao2019 Soares2019)。

我们在监督学习阶段微调预训练 BERT 编码器，对于每一个输入文本$\boldsymbol{x}$，这个编码器$\boldsymbol{E}(\boldsymbol{x}|\boldsymbol{\theta})$（$\boldsymbol{\theta}$为参数）将输出一个$\boldsymbol{d}$维的向量$\boldsymbol{e}$. $\boldsymbol{W}_{\boldsymbol{base}}$是一个矩阵，为每个基类维护一个类级别的向量，充当 meta-learning 的一个基本记忆。$\boldsymbol{E}(\boldsymbol{x}|\boldsymbol{\theta})$和$\boldsymbol{W}_{\boldsymbol{base}}$都将在 meta 训练过程中进一步调整，我们将在实验中表明，用预训练的编码器替换以前的模型的性能要优于相应的最新模型，并且所提出的 DMIN 可以对此进行进一步改进

3.3 动态记忆模块

在 meta-learning 阶段，从给定的 support sets 中归纳类级别的特征，我们基于在监督学习阶段学习阶段通过记忆矩阵$\boldsymbol{W}_{\boldsymbol{base}}$学习的知识，形成动态记忆模块（DMM），与静态记忆不同的是（Grdaris and Komodakis 2018），DMM 利用动态路由（Sabour 2017）为从基本集学习到记忆提供了更多的灵活性，能够更好的适用 support sets，路由部分能在训练期间和训练后自动调整耦合系数，这与生俱来就适合小样本学习的需要。

具体来说，在 support sets 中的实例首先通过 BERT 编码成为样本向量$\{\boldsymbol{e}_{\boldsymbol{c,s}}\}_{\boldsymbol{s=1}}^{\boldsymbol{K}}$和然后输入到下一个动态记忆路由过程

动态记忆路由过程
动态记忆路由算法称为 DMR，如下图所示，

Algorithm 1 Dynamic Memory Routing Process

Require: r, q and memory $M = \{m_1, m_2, ..., m_n\}$

Ensure: $v = v_1, v_2, ..., v_l, q'$

for all m_i, v_j **do**
  $\hat{m}_{ij} = squash(W_j m_i + b_j)$
  $\hat{q}_j = squsn(W_j q + b_j)$
  $\alpha_{ij} = 0$
  $p_{ij} = tanh(PCCs(\hat{m}_{ij}, \hat{q}_j))$
end for
for r iterations **do**
  $d_i = softmax(\alpha_i)$
  $\hat{v}_j = \sum_{i=1}^{n}(d_{ij} + p_{ij})\hat{m}_{ij}$
  $v_j = squash(\hat{v}_j)$
  for all i, j: $\alpha_{ij} = \alpha_{i,j} + p_{ij}\hat{m}_{ij}v_j$
  for all j: $\hat{q}_j = \frac{\hat{q}_j + v_j}{2}$
  for all i, j: $p_{ij} = tanh(PCCs(\hat{m}_{ij}, \hat{q}_j))$
end for
$q' = concat[v]$
Return q'

给定一个记忆矩阵 M（这里是W_{base}）和样本向量 $\boldsymbol{q} \in \boldsymbol{R^d}$, 算法目的是基于在监督学习阶段学习的记忆矩阵 M 来调整样本向量

$$\boldsymbol{q' = DMR(M, q)}$$

首先，对于每一个矩阵条目$\boldsymbol{m_i} \in \boldsymbol{M}$，在动态路由中标准的矩阵转换和压缩操作应用在输入上

$$\boldsymbol{\hat{m}_{ij} = squash(W_j m_i + b_j)}$$

$$\boldsymbol{\hat{q}_j = squash(W_j q + b_j)}$$

其中，转换权重矩阵$\boldsymbol{W_j}$和偏差$\boldsymbol{b_j}$在输入中共享，来适用小样本学习的场景

然后我们计算在$\hat{\boldsymbol{m}}_i$和$\hat{\boldsymbol{q}}_j$之间的皮尔逊相关系数 PCCs

$$\boldsymbol{p_{ij} = \tanh(PPCs(\hat{m}_{ij}, \hat{q}_j))}$$

$$\boldsymbol{PCCs = \frac{Cov(x_1, x_2)}{\sigma_{x_1}\sigma_{x_2}}}$$

上边给出的是关于向量x_1和x_2的 PCCs 通式，因为 PCCs 的值是在[-1, 1]之间的，所有它们可以用来上调或者下调路由参数。

路由迭代过程现在可以针对输入胶囊m_i，q 和较高层次胶囊的v_j来调整耦合系数，用参数d_i代表，

$$d_i = softmax(\alpha_i)$$

$$\alpha_{ij} = \alpha_{ij} + p_{ij}\hat{m}_i v_j$$

因为我们的目标是基于记忆开发动态路由机制来进行小样本学习，因此我们在每次路由迭代过程中都将带有路由协议的 PCCs 添加到其中，如下等式所示

$$\hat{v}_j = \sum_{i=1}^{n}(d_{ij} + p_{ij})m_{ij}$$

$$v_j = \mathbf{squash}(\hat{v}_j)$$

我们用等式 6 和等式 7 来更新耦合系数α_{ij}和p_{ij}，最终输出调整后的向量q'

动态记忆模型的目的是在记忆向量W_{base}使用的指导下，使用动态记忆路由过程（DMR）来调整样本向量$e_{c,s}$，也就是通过$e'_{c,s} = \mathbf{DMR}(W_{base}, e_{c,s})$ 计算，得到最终的调整后的样本向量

3.4 查询增强的归纳模块

在样本向量$\{e'_{c,s}\}_{s=1,\dots,K}$被调整并且用预训练的编码器对查询向量$\{e_q\}_{q=1}^{L}$进行编码后，我们现在结合查询来构建一个以查询为导向的归纳模块（QIM）。这样做的目的是为了在 support sets 被调整过的样本向量中，找到与查询比较相关的向量，以便构建类级别的向量来更好的对查询进行分类。因为动态路由可以自动的调整耦合系数来帮助增强相关的（例如相似度）查询和样本向量权重，并且减少不相关的查询和样本向量权重。QIM 通过把调整后的样本向量作为关于新颖类的背景知识的记忆来重新使用 DMR 过程，然后从与所受关注的查询更相关/相似的调整过的样本向量中归纳出类级别的特征

$$e_c = DMR(\{e'_{c,s}\}_{s=1,\dots,K}, e_q)$$

3.5 相似度分类器

在最终的分类阶段，然后我们把新类向量e_c和查询向量e_q输入到(在监督学习阶段所讨论的)分类器，得到分类得分。对于神经网络分类器的标准设置是在已经抽取特征向量$\mathbf{e} \in R^d$后，首先通过使用点积计算（$s_k = e^T w_k^*$ ）每一个分类$\mathbf{k} \in [1, K^*]$的原始分类得分s_k来评估分类可能的(概率)向量$\mathbf{p}$，然后对于所有的

K^*分类得分都进行softmax运算。然而，由于完全新颖的类别，这种类型的分类器不适应于小样本学习。在研究中，我们使用 cosine 余弦相似度运算来计算原始分类得分：

$$s_k = \tau \cdot \cos(e, w_k^*) = \tau \cdot \bar{e}^T \bar{w}_k^*$$

其中$\bar{e} = \frac{e}{||e||}$ 和 $\bar{w}_k^* = \frac{w_k^*}{||w_k^*||}$是$l_2$标准化向量，$\tau$是一个可学习的标量值，在基本分类器被训练后，所有属于同一类型的特征向量必须与该类别的单个分类权重向量非常紧密地匹配。因此，可以将在第一阶段训练出的基本分类向量$W_{base} = \{w_b\}_{b=1}^{C_{base}}$当做基本类的特征向量。

在小样本分类的场景下，我们把查询向量e_q和新类型向量e_c输入到分类器中，并以统一的方式获得分类得分

$$s_{q,c} = \tau \cdot \cos(e_q, e_c) = \tau \cdot \bar{e}_q^T \bar{e}_c$$

3.6 目标功能

在监督学习阶段，训练目标是在给定的输入文本x和它的标签y的情况下，最大程度减少C_{base}基本类型的交叉熵损失

$$L_1(x, y, \hat{y}) = -\sum_{k=1}^{C_{base}} y_k \log(\hat{y}_k)$$

其中y是正确标注标签的一个特征，$\hat{y}$是使用$\hat{y}_k = softmax(s_k)$计算的基本类型的预测可能性

在 meta 训练阶段，对于每一个 meta 片段（episode），给定 support set S 和查询数据集 $Q = \{x_q, y_q\}_{q=1}^{L}$，训练的目标是使在新类型C上的交叉熵的损失最小

$$L_2(S, Q) = -\frac{1}{C}\sum_{c=1}^{C}\frac{1}{L}\sum_{q=1}^{L} y_q \log(\hat{y}_q)$$

其中$\hat{y}_q = softmax(s_q)$ 是在这一个 meta 片段（episode）中，新类型C的预测可能性，$s_q = \{S_{q,c}\}_{c=1}^{C}$ 来自公式 12，我们把 support set S 输入到模型中，更新参数以达到最大程度减少每个 meta 片段（episode）中的查询数据集Q的损失

4 实验

4.1 数据集和评估度量

我们在 miniRCV1 和 ODIC 数据集上评估我们的模型。紧接着以前的工作

(Snell2017,Geng2019),我们使用小样本分类精度作为评估指标。我们从miniRCV1和ODIC测试集中平均获得了100到300个随机生成的meta-episodes，我们在每个episode中每个类抽取10个文本，来在1-shot和5-shot场景进行评估。

4.2 实现细节

我们使用谷歌预训练的 BERT-Base 模型作为我们的文本编码器，并在训练过程中对模型进行微调。在ODIC和miniRCV1上基本类型C_{base}数量分别被设置为100和20。DMR交互的数量为3，我们构建基于episode的meta训练模型，将$\mathbf{C} = [\mathbf{5}, \mathbf{10}]$和$\mathbf{K} = [\mathbf{1}, \mathbf{5}]$进行比较。除了使用K个样本文本作为support set之外，查询数据集还为每个训练片段（episode）中的C个采样的类，每一个提供10个查询文本。例如在一个训练片段（episode）中有10*5+5*5=75个文本，进行5-way 5-shot实验。

4.3 结果

我们将DMIN与各种基准和最新的模型进行比较：BERT微调(Devlin2019)，ATAML（Jiang2018）Rel.Net(Sung 2018),Ind. Net(Geng2019),HATT(Gao2019)和LwoF(Gidaris和Komodakis2018)。值得注意的是，我们用BERT句子编码器重新实现了他们以进行直接比较。

总体表现

模型的精确度和标准偏差展示在1和2表中，我们可以看到DMIN始终优于所有现有模型，并且在这两个数据集上均获得了最新最好的结果，在单侧配对t检验下，DMIN与其他模型在统计学上差异明显，为95%。

请注意，LwoF使用从监督学习阶段学习到的并且在meta-learning阶段使用的记忆模块来构建两阶段训练步骤，但是在训练之后该记忆机制是静止的，而DMIN 在训练后泛化到新的类型是使用动态记忆路由自动调整耦合参数，这明显优于LwoF，同时也注意到一些基准模型(Rel.Net and Ind. Net)的表1和2展示的表现高于在(Gen2019)，因为用BERT代替BiLSTM-based编码器，BERT编码器通过强大的上下文意义表达能力改进了基准模型，同时由于我们采用了动态记忆路由方法使得我们的模型进一步超过其他模型，尽管与这些较为强大的基准模型比较，我们所提出的DMIN在这两个数据集上始终优于它们

对比试验，在表3中，我们在数据集ODIC上分析DMIN不同组件的作用，具体的来说，我们删除DMM和QIM，并更改了DMR的迭代次数，我们看到通过3次迭代可以获得最佳性能，这结果表明了动态记忆模块和查询信息归纳模块的效能。

4.4 进一步分析

图二是在ODIC数据集上进行10-way 5-shot试验，在DMM之前和在DMM之后支

持样本向量的 t-SNE 可视化效果，我们从 ODIC 测试集中随机选择一个包含 50 个文本（每个类 10 个文本），获得 DMM 之前和之后的样本向量，也就是 $\{\boldsymbol{e}_{c,s}\}_{c=1,\ldots 5,s=1\ldots 10}$和$\{\boldsymbol{e}'_{c,s}\}_{c=1,\ldots 5,s=1\ldots 10}$ 。我们可以看到由 DMM 生成的向量可以更好地分离，这表明 DMM 在利用监督学习经验对低级实例特征和高级类特征之间的语义关系进行编码，以进行小样本文本分类的有效性

5.结论

我们为了小样本文本分类提出了动态记忆归纳网络，该网络建立在具有动态路由外部工作记忆的基础上，利用后者来追踪以前的学习经历，而前者来适应和泛化以更好地支持 support sets，同时因此而更好的适应未见过的类型。这个模型在 miniRCV1 和 ODIC 数据集获得了最新的最佳效果，因为动态记忆可能是比我们小样本学习已经使用过的更加广泛的学习机制，我们将在其他学习问题上研究这种类型的模型

致谢

本文的作者 ACL-2020 的组织者和审稿人的有益的建议。同时最后一位作者的研究得到了加拿大自然科学与工程研究委员会（NSERC）的支持