文章编号：2095-2163(2023)04-0027-06 中图分类号：TP391 文献标志码：A

文本阅读理解的快速多粒度推断深度神经网络

王思语[1,2]，程 兵[1]
（1 中国科学院 数学与系统科学研究院，北京 100049；2 中国科学院大学，北京 100049）

摘 要：机器阅读理解任务(MRC) 是自然语言处理领域的重要研究方向，通过深度学习网络来进行机器阅读理解课题研究已成为目前的主流方法。考虑到深度网络中的计算冗余与同质性现象，本文提出了一个快速多粒度推断深度神经网络(FMG)。FMG 模型在纵向上以卷积神经网络和注意力机制为基本底层架构，横向上以多粒度的文章文本表征与问题表征分层交互融合，共同实现答案的推断。实验结果表明，多粒度推断机制在提高模型表现上具有一定的有效性，且相比于经典循环神经网络，模型实现了训练速度上的进一步提升。

关键词：机器阅读理解；深度学习；多粒度推断；卷积神经网络

Fast multi-granularity inference deep neural networks for text reading comprehension

WANG Siyu[1,2], CHENG Bing[1]
(1 Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100049, China;
2 University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】 Machine reading comprehension (MRC) is an important research direction in the field of natural language processing. Applying deep learning network to research machine reading comprehension has become the mainstream methods at present. Considering the computational redundancy and homogeneity in depth network, a fast multi-granularity inference deep neural network (FMG) is proposed in this paper. The FMG model takes convolutional neural network and attention mechanism as the basic underlying architecture vertically, and multi-granularity passage representation and question representation are interacted in a hierarchical interactive way horizontally, so as to jointly realize the inference of answers. The experimental results show that the multi-granularity inference mechanism is effective in improving the performance of the model, and the model improves the training speed compared with the classical recurrent neural networks model.

【Key words】 machine reading comprehension; deep learning; multi-granularity inference; Convolutional Neural Networks

0 引 言

对机器阅读理解(MRC)的研究致力于提高机器的智能水平，从大量的文本信息中提取最关键的部分。由于其在各个领域的广泛应用，MRC 已成为人工智能前沿研究中最热门的方向之一。应用于机器阅读理解的模型在训练时长上存在不足，且存在同质性的问题。

为克服模型训练极慢的问题，Weissenborn 等学者[1]提出的 FastQAExt 模型使用了轻量级的架构，节约了时间成本，但本质上仍是 RNN 类模型，无法并行化。Zhang 等学者[2]提出的 jNet 模型，对文章与问题的相似度矩阵进行池化，过滤不重要的文章表征。微软亚研的 R-Net 模型使用门限注意力机制，屏蔽掉一些无关信息，是首个在某些指标中接近人类的深度学习模型[3]。谷歌提出的 QANet，将卷积神经网络应用于文本特征的提取，克服了 RNN 网络无法并行的问题[4]。

同质性现象指随着层数的增加，更深层、更抽象的一些表征被抽取出来，这就使得某一位置与其他位置相似的可能性增加。Dai 等学者[5]分析了近年来特征融合存在的问题，提出了注意力特征融合机

基金项目：科技创新 2030—“新一代人工智能”重大项目(2021ZD0111204)。

作者简介：王思语(1996-)，女，硕士研究生，主要研究方向：自然语言处理；程 兵(1963-)，男，博士，研究员，博士生导师，主要研究方向：金融统计、人工智能、自然语言处理。

通讯作者：程 兵 Email：bc2@amss.ac.cn

收稿日期：2022-05-18

制(AFF)。Chen 等学者[6]设计了链式 LSTMs 推断模型，特征融合时计算差和点积,体现两特征的差异性。Wang 等学者[7]基于此方法,对模型中部分表征进行融合,且有研究表明在 TriviaQA 数据集上表现最好。Chen 等学者[6]基于 BERT 模型,采用自适应的方法直接将各编码器的输出表征加权加和,超过了原模型的表现。Huang 等学者[8]提出历史单词的概念,将文章与问题的表征通过注意力函数融合起来得到历史单词,从而帮助推断出答案。

本文针对抽取式机器阅读理解任务,提出一个快速多粒度推断深度神经网络(FMG)。以 CNN 网络和注意力机制为基层网络,融合函数在纵向上的使用,使得浅层表征贯穿整个模型,参与到最终答案的推理中,横向上,交互模块打破以往模型一贯采用单交互模块的方式,以多个交互模块接受问题和不同层次文章表征,形成多个不同层次问题导向的文章表征,以指导答案推理,这种机制可称为多粒度推断机制。

1 一般方法论

研究中,首先给出抽取式机器阅读理解的形式化定义。

定义 1 **抽取式** MRC(Extraction-based MRC) 给定三元组 (P, Q, A) 的样本,其中 P 表示文章,Q 表示问题,A 表示答案,通过训练学习 $(P, Q) \rightarrow A$ 的映射关系,即学习函数 $f(\cdot, \cdot)$ 使得 $f(P,Q)=A$。具体地,$A=(a_{start}, a_{end})$,这里 a_{start} 表示答案在文章中开始的位置,a_{end} 表示答案在文章中结束的位置。

对于一篇文章 P 与一个问题 Q,经过监督学习训练得到答案 A 的过程如下:

(1) P 与 Q 分别经由嵌入处理,把维数为所有词的数量的高维空间嵌入到一个维数低得多的连续向量空间中,每个单词被映射为实数域上的向量,得到文章嵌入向量 $\boldsymbol{V}_P$ 与问题嵌入向量 $\boldsymbol{V}_Q$。

(2) $\boldsymbol{V}_P$ 与 $\boldsymbol{V}_Q$ 向量分别经由特征提取模块,将文本数据转换为可用于机器学习的数字特征,得到特征向量 $\boldsymbol{U}_P=g(\boldsymbol{V}_P)$ 与 $\boldsymbol{U}_Q=g(\boldsymbol{V}_Q)$,其中 $g(\cdot)$ 是编码函数。

(3) $\boldsymbol{U}_P$ 与 $\boldsymbol{U}_Q$ 进行交互处理,将问题的信息整合融入到文章的特征向量中,得到文章与问题之间的交互信息后的向量 $\boldsymbol{X}=[\boldsymbol{U}_P; Att(\boldsymbol{U}_P, \boldsymbol{U}_Q)]$,其中 $[\cdot;\cdot]$ 表示粘接,$Att(\cdot,\cdot)$ 表示注意力机制。

(4) 交互向量 $\boldsymbol{X}$ 经由预测器处理,解码得到预测出的答案 $A=Pred(\boldsymbol{X})$。

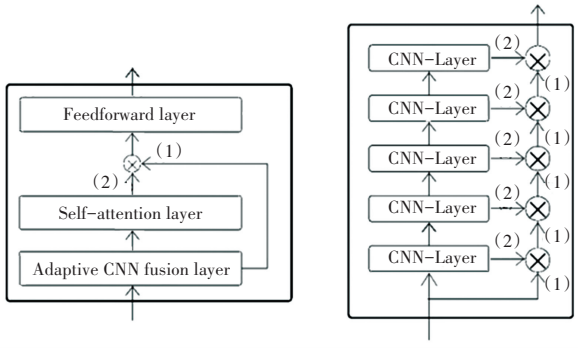
嵌入处理最早的方法为独热(One-hot) 字词嵌入,而后发展出 Word2Vec、Glove 等方法,字符嵌入、命名实体信息、词性特征等语义关系规则也在后续发展中被考虑到嵌入过程中。一般情况下,特征提取模块常用的方法有 CNN、RNN、Transfomer 等。其中,著名的预训练模型 BERT 就是以 Transfomer 为基本模块。从文本交互方法上,交互处理过程一般有单向的注意力机制交互以及双向的注意力机制交互;从迭代轮次方面而言,一般分为单跳交互与多跳交互,主要对应着单个问题与连续提问的情况。答案预测过程常用的方法有 Ptr-Net、Memory Networks 等,预测类型对应于 4 种不同的 MRC 任务,分别为单一单词预测、多项选择、范围抽取、答案生成。

2 快速多粒度推断机制

本文提出以 CNN 网络和注意力机制为基层网络,融合函数在纵向上的使用,使得浅层表征贯穿整个模型,参与到最终答案的推理之中。横向上,交互模块打破以往模型一贯采用单交互模块的方式,以多个交互模块接受问题和不同层次文章表征,形成多个不同层次问题导向的文章表征,以指导答案推理,这种机制可称为快速多粒度推断机制。纵向上的特征提取以 CNN-A 模块来实现,横向上的交融采用 P-Q 多粒度交融方式完成。

2.1 CNN-A 模块

在一个 CNN-A 模块(CNN-A Block)中,包含自下而上的自适应 CNN 融合层 (Adaptive CNN fusion layer)、自注意机制层(Self-attention layer) 以及前馈-前向层 (Feedforward layer),如图 1 所示。图 1 中,"⊗"表示特征融合操作,(1) 表示输出矩阵为传入融合操作的第一矩阵,(2) 表示输出矩阵为传入融合操作的第二矩阵。



(a) CNN-A 模块　　(b) 图(a)中自适应 CNN 融合层具体结构

图 1 CNN-A 模块

Fig. 1 CNN-A Block

自下而上的自适应 CNN 融合层为图 1 中子图(b)，共由 5 个 CNN-Layers 堆叠而成。每个 CNN-layer 由下层的 layernorm 层标准化，再传入至深度可分离卷积层[8]，深度可分离卷积的使用，减少了大量参数，提高了训练速度。在此基础上，采用融合操作对相邻两层的特征进行融合。融合操作公式具体如下：

$$Fuse(I_1, I_2) = \lambda \cdot \tanh(\boldsymbol{W}_f[I_1; I_2; I_1 \circ i_2; I_1 - I_2]) + (1 - \lambda) \cdot I_1 \quad (1)$$

其中，“∘”表示元素逐点乘积，λ 是待训练的参数，函数经由投影算子 $\boldsymbol{W}_f$ 与输入的原始输入矩阵的大小保持一致，这里 $\boldsymbol{W}_f \in R^{d \times 4d}$。

浅层表征、即靠前的 CNN-Layer 得到的表征，作为融合操作的第一输入矩阵。相应地，深层表征作为第二输入矩阵，充分捕捉了文本表征的局部信息并加以纵向融合。

自下而上的自适应 CNN 融合层结果传入自注意机制层，捕捉到全局的信息。自注意机制层对上层特征先进行 Layernorm 层标准化，再计算自注意力。

研究中将自注意机制得到的表征（第二输入矩阵）与自适应 CNN 融合层得到的表征（第一输入矩阵）再次进行融合传入前馈-前向层中。

2.2　P-Q 多粒度交互

交互层旨在将文章与问题的信息进行交融，生成以问题为导向的文章表征。在早期的问答系统模型中，通常将文章信息与问题信息整合成为一个特征向量，BiDAF（Seo 等学者，2016）[9] 的提出，为模型交互层制定了一个新的标准，即每个时间步骤的注意力向量、以及来自前一层的嵌入，均被传输到随后的建模层，这就减少了早期汇总造成的信息损失。同时 BiDAF 从以下 2 个方向计算注意力向量：

（1）问题关于文章方向的注意力（Passage-to-question Attention）：度量对每个文章词而言，问题中哪些词与其相关程度较高，从而依据相关程度给问题中词分配权重，得到关于这一文章词的问题词向量的加权和，每个文章词均得到一个向量。

（2）文章关于问题方向的注意力（Question-to-passage Attention）：度量哪个文章词与问题中某个词有更强的相关性，从而对问题的回答更加重要。对问题中每个词，寻找出与其相似度最高的文章词，再将整个问题所匹配的这些文章词进行加权求和，得到一个向量，随后将这一向量进行平铺，与每个文章词对应。

本文中提出的 FMG 在文章与问题的交互上采用类似 BiDAF 的模式，从 2 个方向分别计算注意力向量。但本文的模型出于对早期汇总造成信息损失的考量，将文章在纵向上捕捉到的 3 个层级的特征，分别与问题特征进行横向上的交互，交互层包含 3 个横向的、彼此不相连的交互模块，分别接收文章不同层次粒度的表征与问题的表征。

FMG 中每个交互模块接收到关于文章的表征 $\boldsymbol{H} \in R^{d \times T}$ 与关于问题的表征 $\boldsymbol{U} \in R^{d \times T}$，交互信息的计算首先确定相似度矩阵 $\boldsymbol{S} \in R^{T \times J}$，其中 S_{ij} 表示第 t 个文章词与第 j 个问题词之间的相似度，可由式（2）进行描述：

$$S_{ij} = \alpha(\boldsymbol{H}_{:T}, \boldsymbol{U}_{:j}) \in R \quad (2)$$

其中，α 是一个待训练的尺度函数，编码 2 个向量间的相似度；$\boldsymbol{H}_{:T}$ 是矩阵 $\boldsymbol{H}$ 的第 t 列向量；$\boldsymbol{U}_{:j}$ 是矩阵 $\boldsymbol{U}$ 的第 j 列向量，此处选择 $\alpha(h,u) = \boldsymbol{w}^{\mathrm{T}}[h;u;h \circ u]$，这里 $\boldsymbol{w} \in R^{3d}$ 是待训练的权重向量，[;] 表示向量粘接串联。得到相似度矩阵 $\boldsymbol{S}$ 后，便可从 2 个方向计算注意力向量：

（1）问题关于文章方向的注意力（Passage-to-question Attention）。通过 *softmax* 函数对相似度矩阵 $\boldsymbol{S}$ 的每一行进行归一化，得到矩阵 $\bar{\boldsymbol{S}} \in R^{T \times J}$，$\bar{\boldsymbol{S}}$ 中第 t 行 $\bar{\boldsymbol{S}}_{t:} = softmax(\boldsymbol{S}_{t:}) \in R^J$，表示所有问题词关于第 t 个文章词的关注权重，由此问题关于文章方向的注意力将被计算为：

$$\boldsymbol{A} = \boldsymbol{U} \cdot \bar{\boldsymbol{S}}^{\mathrm{T}} \in R^{d \times T} \quad (3)$$

（2）文章关于问题方向的注意力（Question-to-passage Attention）。FMG 网络采用了一种更加简单有效计算此方向注意力的方法。与上述 $\bar{\boldsymbol{S}}$ 类似，对 $\boldsymbol{S}$ 进行列归一化，得到矩阵 $\bar{\bar{\boldsymbol{S}}} \in R^{T \times J}$，第 j 列 $\bar{\bar{\boldsymbol{S}}}_{:j} = softmax(\boldsymbol{S}_{:j}) \in R^{\mathrm{T}}$，表示所有文章词关于第 j 个问题词的关注权重，最终 Q2P attention 的计算为：

$$\boldsymbol{B} = \boldsymbol{H} \cdot \bar{\bar{\boldsymbol{S}}} \cdot \bar{\boldsymbol{S}}^{\mathrm{T}} \in R^{d \times T} \quad (4)$$

得到 2 个方向的 attentions 之后，将这些结果结合起来作为双向注意流机制的输出 $\boldsymbol{G} \in R^{4d \times T}$，其中 $\boldsymbol{G} = [\boldsymbol{H}_{:t}; \boldsymbol{A}_{:t}; \boldsymbol{H}_{:t} \circ \boldsymbol{A}_{:t}; \boldsymbol{H}_{:t} \circ \boldsymbol{B}_{:t}] \in R^{4d}$。

3　FMG 模型架构

本文提出的 FMG（Fast multi granularity inference deep neural networks）模型包含以下 5 层（详见图 2）：

（1）嵌入层（Embedding Layer）：将每个单词通

过词嵌入与字符嵌入转化为一个向量。

(2)编码层(Encoder Layer):通过 CNN-A 模块搭建而成的 3 个 CNN-A 编码块提取文章的不同深浅层次的语义特征。

(3)交互层(Interaction Layer):不同粒度的文章向量分别与问题向量利用双向注意流进行匹配。

(4)建模层(Modeling Layer):遍历不同深浅层次的上下文信息并融合。

(5)输出层(Output Layer):得到文章每个位置为答案开始与结束位置的概率,给出问题的答案。

至此可知,模型的创新点包含 CNN-A 编码器与模型编码器中自适应融合机制、交互层中多粒度交互模式、输出层特征融合操作。

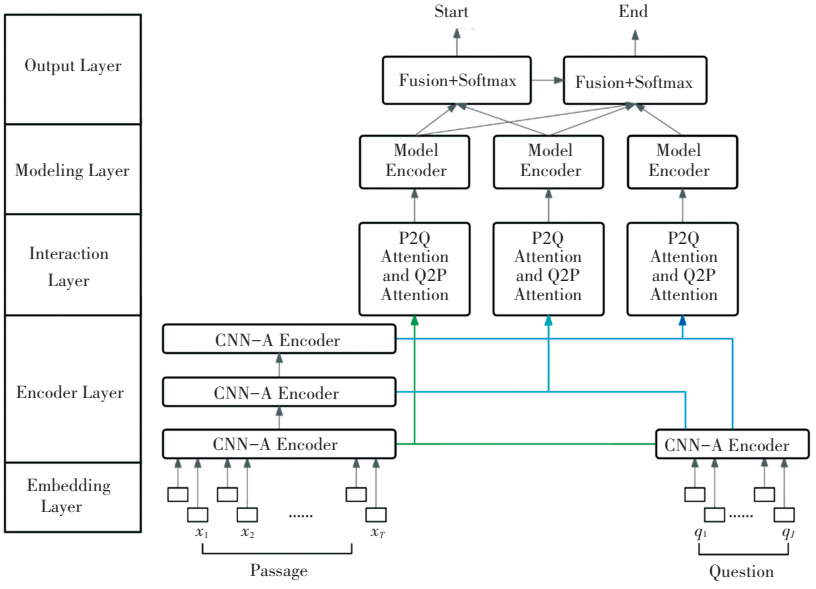


图 2 快速多粒度推断深度神经网络

Fig. 2 Fast multi-granularity inference deep neural networks

3.1 嵌入层

首先将每一个单词转化为 $p_1 = 300$ 维的向量,此处采用 Glove 的单词预训练向量,对于袋外的词汇,用<UNK>表示,并将其随机初始化,作为单词嵌入。字符嵌入将每个单词映射到一个高维向量空间。每个单词的长度被截断或填充为 16,每个字符表示为 $p_2 = 200$ 维的可训练向量,每个单词可以视为其每个字符的嵌入向量的跨行粘接,得到的矩阵每行取最大值,从而得到每个单词固定大小的词嵌入向量表示。将单词嵌入与字符嵌入粘接,再采用与 BiDAF 相同的方法,使用两层 Highway 网络得到整嵌入层的输出,此阶段过后,得到 2 个矩阵,分别是关于文章输入结果的矩阵 $\boldsymbol{P}' \in R^{d\times T}$,以及关于问题输入结果的矩阵 $\boldsymbol{Q}' \in R^{d\times J}$,其中 T, J 分别是文章和问题的长度。

3.2 编码层

FMG 模型的编码层采用 3 个 CNN-A 模块堆叠而成的 CNN-A 编码器(CNN-A encoder)构建。kenrel 大小设置为 7,filters 的数量 $d = 128$。自注意力机制的多头个数为 $h = 8$。为利用不同粒度级别的文章信息,与问题信息进行交互融合,文中用 3 个 CNN-A 编码器串联编码文章,1 个 CNN-A 编码器编码问题。因此,研究得到关于文章嵌入 $\boldsymbol{P}'$ 的矩阵 $\boldsymbol{H}_1, \boldsymbol{H}_2, \boldsymbol{H}_3 \in R^{d\times T}$,关于问题嵌入 $\boldsymbol{Q}'$ 的矩阵 $\boldsymbol{U} \in R^{d\times J}$。

3.3 交互层

FMG 模型出于对早期汇总造成信息损失的考量,将文章在纵向上捕捉到的 3 个层级的特征,分别与问题特征进行横向上的交互,见图 2, P2Q Attention and Q2P Attention 模块、即为交互模块,交互层包含 3 个横向的、彼此不相连的交互模块,分别接收文章不同层次粒度的表征与问题的表征。每个交互模块分别利用双向注意流方法计算文章关于问题方向的注意力和问题关于文章方向的注意力。

P-Q 多粒度交互方法得到问题关于文章方向的注意力分别是 $\boldsymbol{A}_1, \boldsymbol{A}_2, \boldsymbol{A}_3 \in R^{d\times T}$,文章关于问题方向的注意力分别为 $\boldsymbol{B}_1, \boldsymbol{B}_2, \boldsymbol{B}_3 \in R^{d\times T}$。研究将这些结果结合起来作为整个交互层的输出 $\boldsymbol{G}^i \in R^{4d\times \mathrm{T}}$,其中 $\boldsymbol{G}^i_{:t} = [\boldsymbol{H}^i_{:t}; \boldsymbol{A}^i_{:t}; \boldsymbol{H}^i_{:t} \circ \boldsymbol{A}^i_{:t}; \boldsymbol{H}^i_{:t} \circ \boldsymbol{B}^i_{:t}] \in R^{4d}$, $i = \{1, 2, 3\}$。

3.4 建模层

本层模型编码块(Model Encoder)的结构由 3

个 CNN-A 模块堆叠而成。

3.5 输出层

研究中对于开始与结束概率的预测，所采用的不是同样的上下文最终编码矩阵，而是采用融合方法，将 M_1 与 M_2 融合并计算开始位置概率的最终编码，此结果再与 M_3 融合并计算结束位置概率的最终编码，即：

$$\boldsymbol{p}^s = softmax(\boldsymbol{w}_s^{\mathrm{T}} Fuse(M^2, M^1)) \tag{5}$$

$$\boldsymbol{p}^t = softmax(\boldsymbol{w}_t^{\mathrm{T}} Fuse(M^3, Fuse(M^2, M^1))) \tag{6}$$

其中，$Fuse(\cdot,\cdot)$ 为2.1节中的融合操作，线性化算子为$\boldsymbol{w}_s \in R^d$ 是待训练参数向量，线性化得到一个 T 维的向量，对应文章的 T 个位置；接着，对线性化结果进行归一化，使用的仍旧是归一化函数 $softmax(\cdot)$；最后，归一化的结果作为每个位置对应的开始概率，输出概率最高的位置作为预测答案开始位置。同理，每个位置结束概率以及预测答案结束位置的产生过程可依上述过程给出，在此不再赘述。

答案区间的得分是其开始位置与结束位置概率的乘积，研究中的损失函数定义为所有示例正确答案开始与结束位置所对应预测概率的负对数和的平均，即：

$$L(\theta) = -\frac{1}{N}\sum_{i}^{n} \log(p_{r_i^s}^s) + \log(p_{r_i^t}^t) \tag{7}$$

其中，θ 表示所有待训练参数的集合（包括 $\boldsymbol{W}_f$，$\boldsymbol{b}_f$，λ，$\boldsymbol{w}$，$\boldsymbol{w}_s$，$\boldsymbol{w}_t$，CNN 过滤器的权重与偏差，自注意力机制层的投影矩阵等）；N 是参与训练的样例数；r_i^s, r_i^t 是第 i 个样例的正确答案开始位置与正确结束位置。

模型预测过程中，得到答案区间 $A = (a_{start}, a_{end})$，区间开始与结束位置的选择原则为使得 $p_{a_{start}}^s \cdot p_{a_{end}}^t$ 最大且 $a_{start} \leqslant a_{end}$，动态规划方法下可以在线性时间内得到结果。

4 实验结果与分析

SQuAD 数据集为斯坦福大学发布的机器阅读理解数据集，包含 107 700 个基于 536 篇维基百科文章的问题-答案对，其中 87 500 用于训练，10 100 用于验证，另外的 10 100 用于测试。每个问题均由人工标注出其答案，答案均来源于文章段落中的一个序列。每篇文章的长度大约为 250 个词，答案的长度一般不超过 10 个词。

首先，仿真中使用了 L_2 权重衰减，衰减参数 $\mu = 3 \times 10^{-7}$。在词嵌入间的 *dropout* 概率是 0.1，字符嵌入的 *dropout* 概率为 0.05，层间的 *dropout* 概率为 0.1。在相邻的 CNN-A Block 与模型编码块间采用随机深度层 *dropout*。在训练中，使用 AdamW 优化器，$\beta_1 = 0.8$，$\beta_2 = 0.999$，$\epsilon = 10^{-7}$。在初始阶段使用预热技术（warm-up），前 1 000 个 *steps* 学习率以负指数速率从 0 增长到 0.01，后续每 3 个 *epochs* 学习率变为先前的 0.5 倍。

本文提出的 FMG 模型，有 FMG 标准模型（Standard FMG）和 FMG 超轻模型（Utral lightweight FMG）两种形式，两者的模型结构一样，只是为了训练效率，将 FMG 标准模型的一些结构简化，得到 FMG 超轻模型，后文的表 3 中列出了两模型在层级结构上的数量选择。

4.1 主要结果

研究提出的 FMG 超轻模型在验证集上获得了 70.96%/81.54% 的 EM/F_1 分数，标准模型获得了 73.72%/81.15% 的 EM/F_1 分数，其中标准模型在验证集上的表现，超越了其他对比模型，见表 1。表 1 中，第一列表示 Dev set，第二列表示 Test set。

表 1　SQuAD 数据集上不同模型的表现

Tab. 1　The performances of different models on SQuAD datset

单模型	EM/F_1	EM/F_1
LR Baseline（Rajpurkar 等，2016）	40.0/51.0	40.4/51.0
Match-LSTM with Ptr（Wang 等，2016）	64.1/73.9	64.7/73.7
FastQA（Weissenborn 等，2017a）	-/-	68.4/77.1
BiDAF（Seo 等，2016）	68.0/77.3	68.0/77.3
RasoR（Lee 等，2016）	66.4/74.9	-/-
FastQAExt（Weissenborn 等，2017b）	70.8/78.9	70.8/78.9
Ruminating Reader（GongDeng ，2017）	70.6/79.5	70.6/79.5
jNet（Zhang 等，2017）	-/-	68.7/77.4
R-Net（Group，2017）	71.1/79.5	71.3/79.7
DCN+	-/-	74.9/82.8
SLQA	-/-	74.5/82.8
FusionNet（Huang 等，2017）	-/-	76.0/83.5
QANet（Yu 等，2018）	73.6/82.7	-/-
CGDE and FGIn（Cao 等，2021）	66.4/77.6	-/-
FMG（Ultra lightweight）	70.96/81.54	
FMG（Standard）	**73.72/83.15**	

4.2 提升效果

文中在同样的环境下测试了 QANet 与 FMG 模型的训练速度。研究结果见表 2，分析发现，FMG 标准模型的训练速度是 QANet 的 1.2 倍，而这发生在本次实验的参数量多于 QANet 的情况下。FMG 超轻模型的训练速度是 QANet 的 8.7 倍。FMG 标准

模型在训练速度为 1.2 倍的情况下,获得了高于同为 CNN 模型的 QANet 2.33%/1.78% 的 EM/F_1 分数,在训练速度大于 4 倍的基础上,获得了高于基线模型 3.56%/3.97% 的 EM/F_1 分数。FMG 超轻模型在训练速度大于 24 倍的情况下,获得了高于基线模型 1.23%/1.83% 的 EM/F_1 分数。

表 2 SQuAD 数据集上 FMG 模型与基线模型的速度对比

Tab. 2 Speed comparison between FMG model and baseline models on SQuAD datset

模型	速度/(样本数·s^{-1})	参数量/万	速度倍数(以 QANet 为基准)
BiDAF(Seo 等,2016)	-	-	$\frac{1}{4.3}$×(结果来源于(Yu 等,2018))
QANet(Yu 等,2018)	80	138	1.0×
FMG(Standard)	96	155	1.2×
FMG(Ultra lightweight)	696	71	8.7×

4.3 模型简化测试

多粒度推断结构的效应对比结果见表 3。表 3 中,Model1 与 Model2 模型是去除本次研究多粒度推断结构的模型,也即在 CNN-A Blocks 中去除 Fusion 操作,交互层将横向的多粒度交互模式改为编码层最终输出的交互模式,最终结果输出时不融合表征而是直接将建模层后 2 个编码块的输出线性化归一化。整个信息的流向是纵向的。在层数相同的情况下,本次研究的多粒度推断机制显然提高了 EM/F_1 分数,FMG(Standard) 获得高于 Model2 0.91%/1.54% 的分数,FMG2 获得了高于 Model1 0.72%/1.03%的分数。甚至在 CNN 层数条件弱于 Model2 的条件下,FMG2 模型也凭借着多粒度推断机制的作用,超越了 Model2 模型的表现。

表 3 多粒度推断结构的效应比对

Tab. 3 Effect comparison of multi-granularity inference structure

模型	*CNNs*	*Blocks*	参数量/万	*EM*	F_1
FMG(Standard)	5	3	155.2	**71.50**	**81.27**
FMG(Ultra lightweight)	1	1	78.0	69.23	79.13
Model1	3	3	133.9	69.91	79.48
Model2	5	3	154.1	70.59	79.73
FMG2	3	3	134.8	70.63	80.44

注:Model1 与 Model2 是在 FMG 的基础上去除多粒度推断机制得到的模型

本文提出的 FMG 超轻模型,在满足一些准确率需求的情况下,是值得推荐的短时间内完成文本推断过程的模型,该模型的训练速度,大约是 RNN 模型的 24 倍之多。FMG 标准模型,准确率高于 FMG 超轻模型,训练速度上是 RNN 类模型的 4 倍之多。

5 结束语

本文提出了一个快速多粒度推断深度神经网络 FMG。FMG 模型在纵向上以卷积神经网络和注意力机制为基本底层架构,横向上以多粒度的文章文本表征与问题表征分层交互融合,共同实现答案的推断。实验结果表明,多粒度推断机制在提高模型表现上具有一定的有效性。基于本文的成果,下一步的工作拟将本文提出的多粒度推断机制应用于 BERT 模型,对 CNN-A 模块进行微调,来替换 BERT 模型中的编码器和解码器,以探索本文多粒度推断机制的更多可能性。

参考文献

[1] WEISSENBORN D, WIESE G, SEIFFE L. Making neural QA as simple as possible but not simpler[C]//Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017). Vancouver, Canada: ACL, 2017: 271-280.

[2] ZHANG Junbei, ZHU Xiaodan, CHEN Qian, et al. Exploring question understanding and adaptation in neural-network-based question answering [J]. arXiv preprint arXiv:1703.04617,2017.

[3] Natural Language Computing Group. R-net: Machine reading comprehension with self-matching networks [C]//Annual Meeting of the Association for Computational Linguistics. Macao: ACL, 2017:1-11.

[4] YU A W, DOHAN D, LUONG M T, et al. QANet: Combining local convolution with global self-attention for reading comprehension [J]. arXiv preprint arXiv:1804.09541,2018.

[5] DAI Y, GIESEKE F, OEHMCKE S, et al. Attentional Feature Fusion [J]. arXiv preprint arXiv:2009.14082, 2020.

[6] CHEN Nuo, LIU Fenglin, YOU Chenyu, et al. Adaptive bi-directional attention: Exploring multi-granularity represaentations for machine reading comprehension [J]. arXiv preprint arXiv: 2012.10877,2021.

[7] WANG Wei, YAN Ming, WU Chen. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering [J]. arXiv preprint arXiv: 1811.11934, 2018.

[8] HUANG H Y, ZHU Chengguang, SHEN Yeyong, et al. FusionNet: Fusing via fully-aware attention with application to machine comprehension [J]. arXiv preprint arXiv:1711.07341,2017.

[9] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension [J]. arXiv preprint arXiv:1611.01603,2016.