基于深度学习的机器阅读理解技术研究综述

孙相会

学号：1971654

摘要：　本文主要针对机器阅读理解领域的经典数据集和模型做综述。机器阅读理解 (MRC) 是自然语言处理 (NLP) 领域非常具有挑战性的任务，给定一段文本和与文本有关的几个问题，要求模型能够根据这段文本回答这些问题，也就是让机器向人一样理解文章回答问题。近年来随着深度学习以及 NLP 领域预训练语言模型的发展，在 MRC 方向上的研究已经取得了很大的突破，在很多数据集上模型已经超过了人类的水平。本篇论文是对 MRC 领域做一个概述，主要涉及 (1) 机器阅读理解的任务定义以及不同形式的阅读之间的区别，(2) 一些经典的模型以及目前流行的基于预训练方式的模型 (3) 机器阅读理解领域未来的研究趋势。

关键词: 机器阅读理解；自然语言处理；预训练语言模型

中图分类号: 100　　**文献表示码**: A

A Survey in Machine Reading Comprehension

Xianghui Sun

id: 1971654 (NEU)

Abstract: This article summarizes recent and classic dataset and model in machine reading comprehension.

Key words: machine reading comprehension; what `https://blog.csdn.net/`

1　引言

机器阅读理解 (MRC) 是自然语言处理领域十分重要也是具有挑战性的研究方向，这项任务的目的是衡量计算机理解自然语言文本的能力。具体的就是给定一篇文章和相关的问题，要求计算机通过阅读理解这篇文章后能够正确的回答这些问题。早期的 MRC 系统主要是基于规则和模式匹配的方法，而且数据集规模比较小，系统难以获得期望的性能也不能实际的应用。随着深度学习的兴起，词嵌入技术的发展，注意力机制应用在 NLP 领域 [7] 以及大规模阅读理解数据集如（CNN/Daily Mail [11]，SQuAD [15]，RACE [27]，MS MARCO [32]，CoQA [31] 等）的发布，这些推动了 MRC 领域的发展，越来越多的学者采用神经网络构建 MRC 模型，也叫神经机器阅读理解，效果上显著的优于传统的机器学习方法并且在 SQuAD [15] 数据集上逐渐的接近人类的阅读理解水平。

自 2018 年，随着 ELMo [36]、GPT [37]、BERT [38] 等预训练语言模型的出现，再一次提升了机器阅读理解的水平，特别是 BERT [38] 在 SQuAD 数据集上首次超过了人类的表现。

本篇论文主要从具体任务，数据集，经典的神经机器阅读理解模型，目前流行的基于预训练的模型，对于不同任务的不同的评估指标以及 MRC 领域目前新的趋势几方面对机器阅读理解领域做阐述。

2 机器阅读理解任务概述

机器阅读理解任务是为了使得计算机具有对自然语言文本理解的能力，像人类一样阅读并且理解一篇文章。具体的就是给定一篇文章 P 和一些与文章 P 相关的问题 Q，要求模型通过阅读 P 之后给出 Q 的正确答案 A，即建模给定 P 和 Q 的条件下预测 A 的概率：

$$P(A|P,Q) \tag{1}$$

根据答案形式的不同，任务也是多种多样的，大致可以概括为 4 类：完形填空、多项选择、片段选择和自由回答。下面对这四种任务分别进行叙述并介绍相关的数据集。

2.1 完形填空

完型填空型阅读理解是指给定一篇文章 P 和一个与文章相关的问题 Q，Q 是通过删除掉句子中某一个单词构成，要求模型根据 P 能够正确的填写出 Q 缺失的单词。CNN 和 Daily Mail 数据集 [11] 是由 Google DeepMind 和牛津大学发布于 2015 年，这是第一个较大规模的阅读理解型数据集。从 CNN[1] 中收集 93k 篇文章，从 Daily Mail[2] 上收集 220k 篇文章。每一篇文章的作者都为这篇总结出一些具有概括性的句子，这些句子涵盖了这篇文章的要点。于是把这些概括性的句子删去其中的一个单词，以此作为问题，构建了 (文章-问题-答案) 的三元组形式的语料库作为填空式的阅读理解任务。

2.2 多项选择

多项选择型这类问答任务是对于给定的文 P，以及和 P 相关的问题 Q 和多个候选答案 $A = \{A_1, A_2, \cdots, A_n\}$，从中选择正确的答案，即 $P(A_i|P,Q)$，其中 $A_i \in A$。相关的数据集如 RACE[3]，这个数据集是从中国中学生的英语考试题中建立的数据集。共有将近 28000 篇文章以及 100000 个问题，答案并不是简单的限制于文章中的单词，而且答案和问题中单词可能从没有在文中出现过，因此简单的利用单词匹配方式并不能达到很好的效果。这些问题和候选的答案都是由出题专家生成的，因此更加的接近真实世界的语义。另外 RACE 数据集中文章主题的覆盖度比其它的数据集更广泛，比如 CNN/Daily [11] 所有文章全都是来源于 CNN 新闻，SQuAD [15] 数据集所有的文章全都是来源于维基百科。而 RACE 数据集涵盖多个领域如新闻、故事、广告、传记等等。由于其类型的多样性因此可以更好的评估机器的阅读理解能力。

2.3 片段选择

这类问答任务是 MRC 领域较为流行的研究方向，给出 P 和问题 Q，问题的答案是 P 中的一段连续的单词构成，答案的长度不固定，可以表示为 $P(A|P,Q)$，其中 $A = \{t_i, t_{i+1}, \cdots, t_{i+k}\}(1 \leq i \leq i+k \leq n)$，$n$ 代表 P 中单词的个数。抽取式问答任务最为广泛使用的数据集是 SQuAD 1.1 [15] 和 SQuAD 2.0 [28]。其中 SQuAD 1.1 是 Stanford 问答数据集的第一个版本，由众包工人在维基百科上面的文章中给出问题，答案来源于文章中某段连续的单词。SQuAD 1.1 含有 536 篇文章，总计 107785 个问题-答案对。在 SQuAD 2.0 中又在原有的数据集中加入了 50000 多个没有答案的问题，准确的说这些问题的答案不在相应的文章中。

2.4 自由回答

简单的从文章中摘取一段文本可能并不能回答问题需要的答案，自由回答型任务的答案是自由形式的，不局限于文章中的某些单词，语法上往往是更加的灵活。可以表示为 $P(A|P,Q)$，其中 $A \subseteq P$ 或 $A \nsubseteq P$。从文章中概括提炼出问题的答案也是更加的符合人类的阅读方式的，基于这些原因，自由回答式问答的数据集也因此公布出来并且受到广泛的关注，相关的数据集如 MS MARCO [32]。MS MARCO 是由微软通过在必应搜索引擎的日志上收集用户提出的问题，对于文本段落是来源于必应搜索引擎的返回的搜索结果。具体的就是对于每一个问题，给出 10 个最相关的查询结果的文本段落，然后由标注人员从这 10 个文本段落中找出那些与这个问题有关

[1] www.cnn.com

[2] www.dailymail.co.uk

[3] www.cs.cmu.edu/glail/data/race/

的文本段落，然后人工的从这些选择出来的段落中概括提炼出答案，同时对于选出来的段落要标记为 $is\_select = 1$，表示这个段落和答案相关，从而可以训练模型。

因此可以看到这个数据集与前面的数据集很大的不同之处就是答案是人工生成的，不局限于文本段落中固定的一段单词，因此更加的接近现实世界中的人类阅读理解，对模型的推理能力要求也更高。同时如果不能从给出的 10 个段落中推理出答案，那么这个问题就标记为不可回答的问题，同样要保留在数据集中，目的就是让模型能够判别出问题是否具有答案。

2.5 评估方法

对于不同的 MRC 任务有不同的评估指标。对于填空型任务与多项选择型任务都是属于客观题型，用准确率就可以衡量模型的性能。片段选择型任务属于半客观题型，通常用精确匹配 EM（Exact Match）和 F1 值来评估模型，F1 值是精确率和召回率之间的调和平均数。对于自由回答式任务的答案，一般采用单词水平的匹配率作为评分标准，常用标准有 ROUGE [3]。下面详细介绍这几种评估指标如何评估不同的 MRC 任务。

2.5.1 准确率

准确率可以用来评估完形填空和多项选择这两种类型的任务。对于测试集合中的所有问题 $Q = \{Q_1, Q_2, \cdots, Q_m\}$，其中 m 代表问题的个数。如果模型预测出来的 m 个答案中有 n 个是正确的，那么模型的准确率自然是 n/m。精确匹配 EM 评估指标可以看做是准确率的扩展，就片段选择型任务来讲，EM 要求预测出来的所有单词要和标准答案的所有单词要精确匹配下，EM 值才为 1，否则为 0。因此最后模型在测试集上的 EM 值也就是 n/m。

2.5.2 F1

F1 分数是最为普遍使用的一种评估标准，不仅仅局限于 MRC 的各种任务。F1 值是精确率（precision）和召回率（recall）之间的调和平均数。具体计算公式如下：

$$\mathrm{F1} = \frac{2 \times \mathrm{precision} \times \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{2}$$

精确率是指模型预测的答案中有多大比例的单词是标准答案中的单词。召回率是指标准答案中的单词有多大比例在预测答案中出现。

2.5.3 ROUGE

ROUGE（Recall-Oriented Understudy for Gisting Evaluation）最初是用来评估生成文本摘要的一种方法，因为可以 ROUGE 的计算机制来评估 MRC 领域中自由答案型任务。ROUGE 的评分有多种，在 MRC 领域较为常用的是 ROUGE-L，ROUGE-L 用来计算标准答案和预测答案的最长公共子序列（Longest Common Subsequence，LCS），ROUGE-L 的计算公式如下：

$$R_{LCS} = \frac{LCS(X,Y)}{m} \tag{3}$$

$$P_{LCS} = \frac{LCS(X,Y)}{n} \tag{4}$$

$$F_{LCS} = \frac{(1+\beta)^2 R_{LCS} P_{LCS}}{R_{LCS} + \beta^2 P_{LCS}} \tag{5}$$

其中 $LCS(X,Y)$ 表示标准答案与预测答案之间的最长公共子序列，m 和 n 分别代表标准答案和预测答案中单词的个数，β 是 ROUGE-L 的的参数，用来控制精确率和召回率的重要程度。

3 神经机器阅读理解模型

随着大规模机器阅读理解数据集如 CNN&Daily Mail [11]，SQuAD 1.1 [15] 等的发布以及深度学习技术的发展，神经机器阅读理解模型的性能显著的超过传统的基于规则和特征的模型，随着 NLP 领域预训练模型的发展，基于预训练模型来做 MRC 任务的模型性能再一次的提升。用于 MRC 任务的深度学习模型的整体框架主要包括如下几个层：词嵌入层、语义编码层、语义交互层、答案预测层。

1）词嵌入层：如何将文本有效的表示成计算机可以处理的形式同时可以有效地利用单词之间的语义一直的 NLP 领域的重点问题。分布式表示是将单词用一个低维度的稠密向量表示，即将单词嵌入到一个

填空型数据集

CNN&Daily Mail	文章： 22222222
	问题： 3333
	答案： 4444

低维稠密空间中，因此这种表示方式也叫词嵌入。从早期的 one-hot 形式编码到词袋模型再到分布式表示技术最后到基于上下文的词嵌入技术，每一种技术的出现都证明了一个好的文本表示方法可以极大地提升模型的性能。

2）语义编码层：这一层的目的是在词嵌入层的基础上通过对词嵌入层的输入文本做特征提取，进一步获得句子层面的语义信息。常用的特征提取器有基于 RNN 的变体如 LSTM [1] 和 GRU [?] 等。但由于梯度消失问题不能解决长距离依赖问题，使得其特征提取能力始终受限。Transformer [34] 通过利用自注意力机制取代 RNN 那种序列式的计算方式，通过自注意力机制不仅可以做到单词之间的全局交互同时其并行计算使得模型训练时间大幅减少。实验证明在大规模数据集上 transformer 的特征提取能力要强于基于 RNN 的编码器，目前几乎所有的 NLP 预训练模型都是利用 transformer 作为特征提取器。

3）语义交互层：在预测答案时需要将问题的语义信息与文章的语义信息关联，这样模型在预测答案时才能知道文章中哪一部分是问题的答案。通常利用注意力机制实现这一目标，注意力机制就是让模型关注到重点的部分，不同的注意力计算方式很大程度上的影响模型性能，后面将详细介绍基于注意力机制的模型以及它们不同的计算方式。

4）答案预测层：这是整个模型架构的最后一层，用来输出预测的答案。如前面所提到的 MRC 任务大致分成四类，因此这一层的设计需要考虑到答案形式。对于填空型任务，答案的输出是文章中的一个单词。对于多项选择任务，答案的输出是从多个候选答案中选择出正确的选项。对于片段选择型任务，答案的输出是文章中某段连续的文本。对于自由答案型任务，答案的输出不限固定的文本，而是根据文章中的单词生成的答案。此外还有不可回答的问题，此时模型的输出还要考虑到问题是否可以回答。

下面介绍神经机器阅读理解模型中基于注意力机制的模型和基于预训练模型的模型。

3.1 小结

本章是论文的核心部分，首先介绍了神经机器阅读理解模型的结构，涉及的技术如注意力机制以及相关的模型。然后介绍了目前流行的几种预训练模型，最后对于不同的 MRC 任务分别介绍了各自输出层的设计。

在预训练模型的基础上利用具体任务的数据微调模型，即只需要稍微添加简单的输出层即可达到很好的效果。但是预训练模型只是给了更好的初始化参数，如果想要进一步提升模型的性能还需要在其基础上根据具体的任务设计一个更好的模型。如文献 [?] 利用 BERT 作为编码器同时采用文献 [44] 提出的 co-matching 方法提出了 DCMN 模型，在 RACE [27] 数据集上达到了很高的准确率。文献 [?] 在原有的 DCMN 模型的基础上提出了 DCMN+ 模型，主要是增加了句子选择模块和答案交互模块，再一次提升了模型的性能。文献 [?] 提出一种回顾式阅读器（Retropective Reader，Retro-Reader）模型，以

ALBERT 作为编码器。Retro-Reader 分为两个模块，略读模块和精读模块，主要目的就是模仿人类阅读习惯：略读模块先阅读文章和问题然后给出初步判断问题是否可以回答，精读模块用于鉴定可回答性，在可回答的前提下给出答案，在 SQuAD 2.0 [28] 数据集上显著优于其它模型。

因此可以看到如何利用预训练模型结合具体任务改进模型的结构是至关重要的。

4 总结与展望

本文从机器阅读理解任务的定义出发，首先介绍了不同任务下的数据集以及相应的评估标准。然后对神经机器阅读理解模型进行了分析与研究，主要涉及模型的整体框架以及所用到的方法如注意力机制、推理结构等。同时也总结了目前一些主流的预训练模型，分析了它们之间的差异，列举了一些在预训练模型的基础上改进的模型。通过各个模型的实验对比结果可以看到基于预训练模型的模型性能要显著的优于传统的仅仅基于注意力机制的模型。

机器阅读理解赋予了计算机阅读理解文本的能力，在搜索、对话、医疗以及教育领域都有着广阔的应用空间。但是目前机器阅读理解仍然存在一些难题，1）基于注意力机制的匹配模型大多是浅层的语义匹配模型，基于多跳结构的推理模式还过于单一，而阅读理解是需要深层次的推理过程才能更好的理解文章，让模型具有较强的推理能力至关重要。

2）目前机器阅读理解领域的数据集大多是通用领域方向的，而设计专业领域数据集也尤为重要，更重要的是这些适用于通用领域数据集的模型未必在专业领域有一样的性能。

3）生成答案的技术还需要进一步提升，回顾目前机器阅读理解领域的数据集以及相应的模型，大多集中于片段选择式问答且模型准确度很高甚至超过人类水平。而对于自由答案型这种需要生成答案的模型效果很差，有些模型直接将生成式问题转为抽取式问题，主要原因在于生成答案模块对模型要求变得更高。

4）目前机器阅读理解主要集中于非结构化的文本领域，而还有许多其它结构，不同模态的数据如表格、视频、音频、图片等，多模态阅读理解模型也是未来的发展方向之一。

参考文献

[1] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8): 1735–1780, 1997.

[2] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pages 311–318. Association for Computational Linguistics, 2002.

[3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. Text Summariza- tion Branches Out, 2004.

[4] Goodfellow, I.J., Warde-Farley, D., Mirza, M., Courville, A., Bengio, Y.: Maxout networks. arXiv preprint arXiv:1302.4389 (2013)

[5] Mikolov T, Sutskever I, Chen K, et al.Distributed represen- tations of words and phrases and their compositionality[C]// Advances in Neural Information Processing Systems, 2013：3111-3119.

[6] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word//w representation. In Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543, 2014.

[7] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. ICLR, 2015.

[8] J. Weston, S. Chopra, and A. Bordes. Memory networks. In International Conference on Learning Representations (ICLR), 2015.

[9] Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. arXiv preprint arXiv:1505.00387 (2015)

[10] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. End-to-end memory networks. In Advances in neural information processing systems, pages 2440–2448, 2015.

[11] Hermann, K.M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., Blunsom, P.: Teaching machines to read and comprehend. In: Advances in Neural Information Processing Systems, pp. 1693–1701 (2015)

[12] Danqi Chen, Jason Bolton, and Christopher D Manning. A thorough examination of the cnn/daily mail reading comprehension task. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 2358–2367, 2016.

[13] Rudolf Kadlec, Martin Schmid, Ondřej Bajgar, and Jan Kleindienst. Text understanding with the attention sum reader network. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), volume 1, pages 908–918, 2016.

[14] Alessandro Sordoni, Philip Bachman, Adam Trischler, and Yoshua Bengio. Iterative alternat- ing neural attention for machine reading. arXiv preprint arXiv:1606.02245, 2016.

[15] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 2016.

[16] Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843 (2016)

[17] Wang, S., Jiang, J.: Machine comprehension using match-lstm and answer pointer. arXiv preprint arXiv:1608.07905 (2016)

[18] Xiong, C., Zhong, V., Socher, R.: Dynamic coattention networks for question answering. arXiv preprint arXiv:1611.01604 (2016)

[19] Seo, M., Kembhavi, A., Farhadi, A., Hajishirzi, H.: Bidirectional attention flow for machine comprehension. arXiv preprint arXiv:1611.01603 (2016)

[20] Wang, W., Yang, N., Wei, F., Chang, B., Zhou, M.: Gated self-matching networks for reading comprehension and question answering. In: Proceedings of the 55th Annual Meet- ing of the Association for Computational Linguistics (Volume 1: Long Papers), vol. 1, pp. 189–198 (2017)

[21] Shen, Y., Huang, P.S., Gao, J., Chen, W.: Reasonet: Learning to stop reading in machine comprehension. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1047–1055. ACM (2017)

[22] Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1832–1846, 2017.

[23] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In Asso- ciation for Computational Linguistics (ACL), pages 1870–1879. Vancouver, Canada.

[24] Hsin-Yuan Huang, Chenguang Zhu, Yelong Shen, Weizhu Chen. FusionNet: Fusing via

Fully-Aware Attention with Application to Machine Comprehension. In Proceedings of the Sixth International Conference on Learning Representations (ICLR), 2018.

[25] Shuohang Wang and Jing Jiang. Learning natural language inference with LSTM. In Proceedings of the Conference on the North American Chapter of the Association for Computational Linguistics, 2016.

[26] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In Proceedings of the Con- ference on Advances in Neural Information Processing Systems, 2015.

[27] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 785–794, 2017.

[28] Rajpurkar, P., Jia, R., Liang, P.: Know what you don' t know: Unanswerable questions for squad. arXiv preprint arXiv:1806.03822 (2018)

[29] Hill, F., Bordes, A., Chopra, S., Weston, J.: The goldilocks principle: Reading children' s books with explicit memory representations. arXiv preprint arXiv:1511.02301 (2015)

[30] Yu, A.W., Dohan, D., Luong, M.T., Zhao, R., Chen, K., Norouzi, M., Le, Q.V.: Qanet: Combining local convolution with global self-attention for reading comprehension. arXiv preprint arXiv:1804.09541 (2018)

[31] Siva Reddy, Danqi Chen, and Christopher D Manning. Coqa: A conversational question answering challenge. arXiv preprint arXiv:1808.07042, 2018.

[32] Nguyen, T., Rosenberg, M., Song, X., Gao, J., Tiwary, S., Majumder, R., Deng, L.: Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268 (2016)

[33] Kočiskỳ, T., Schwarz, J., Blunsom, P., Dyer, C., Hermann, K.M., Melis, G., Grefenstette, E.: The narrativeqa reading comprehension challenge. Transactions of the Association of Computational Linguistics 6, 317–328 (2018)

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Neural Information Processing Systems, 2017b.

[35] François Chollet. Xception: Deep learning with depthwise separable convolutions. abs/1610.02357, 2016.

[36] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. arXiv preprint arXiv:1802.05365 (2018)

[37] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning. Tech. rep., Technical report, OpenAI (2018)

[38] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

[39] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. arXiv preprint arXiv:1906.08237, 2019.

[40] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized

BERT pre- training approach. arXiv preprint arXiv:1907.11692, 2019.

[41] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In ICLR.

[42] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. arXiv preprint arXiv:1606.00061, 2016.

[43] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language un- derstanding and generation. In NeurIPS, pages 13042–13054, 2019.

[44] Shuohang Wang, Mo Yu, Jing Jiang, and Shiyu Chang. 2018. A Co-Matching Model for Multi-choice Reading Comprehension. In Proceedings of the 56th Annual Meeting of the Association for Computa- tional Linguistics (Volume 2: Short Papers), pages 746–751. Association for Computational Linguis- tics.

[45] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summa- rization with pointer-generator networks. In Annual Meeting of the Association for Compu- tational Linguistics (ACL), pages 1073–1083. Vancouver, Canada.

[46] Chuanqi Tan, Furu Wei, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. S-net: From answer extraction to answer synthesis for machine reading comprehension. In Thirty-Second AAAI Conference on Artificial Intelligence, 2018.