2021 24th International Conference on Computer and Information Technology (ICCIT), 18-20 December, 2021.

Multitask Learning as Question Answering with BERT

Shishir Roy[1] shishir29@student.sust.edu
Nayeem Ehtesham[1] theeyeman18@gmail.com
Md. Saiful Islam[1] saiful-cse@sust.edu
Sabir Ismail[1] sabir-cse@sust.edu

[1]Department of Computer Science and Engineering
Shahjalal University of Science and Technology, Sylhet, Bangladesh

***Abstract*—Question Answering demands a deep understanding of semantic relations among question, answer, and context. Multi-Task Learning (MTL) and Meta Learning with deep neural networks have recently shown impressive performance in many Natural Language Processing (NLP) tasks, particularly when there is inadequate data for training. But a little work has been done for a general NLP architecture that spans over many NLP tasks. In this paper, we present a model that can generalize to ten different NLP tasks. We demonstrate that multi-pointer-generator decoder and pre-trained language model is key to success and suppress all previous state-of-the-art baselines by 74 decaScore which is more than 12% absolute improvement over all of the datasets.**

***Index Terms*—Question Answering, BERT, Multitask Learning, Meta Learning, Transfer Learning**

I. Introduction

In recent years, the advancement of deep learning has addressed factoid question answering systems. Large-scale Knowledge Base (KB) includes Freebase [1] or DBPedia [2] crafted to store the knowledge of the world's facts in a structural database, which is used for factoid open-domain Question Answering (QA). Neural semantic parsing approach [3],[4], [5],[6],[7] for QA is becoming increasingly relevant throughout recent years since it does not rely on handcrafted features and convenient to relocating across domains. In traditional approaches, answers are usually enclosed in the limited KB [5, 8] and large-scale KBs are difficult to maintain. The fact that the answers can be readily derived from the context is widely assumed for a lot of QA models [9],[10],[11]. For question-answering in real world scenarios on an extensive scale things get challenging.

For learning universal language representations, the pre-training language model has proven to be conducive which benefits from the huge amount of unlabeled data. Transformers [12] have gradually become a key component for variety state-of-the-art natural language processing tasks [13],[14],[15]. These include question-answering [16], machine translation [17], semantic role labeling [18], and natural language inference [19, 20]. Embeddings from Language Model (ELMo) [14], Generative Pre-trained Transformer (GPT) [21] and Bidirectional Encoder Representations from Transformer (BERT) [16] are three of the stand out models according to a recent study by Gao, Jianfeng, Michel Galley, and Lihong Li [22]. BERT achieved state-of-the art results on many natural language processing tasks. BERT is a multi-layer bidirectional transformer. To put it into perspective, a cutting edge model that can work on a number of downstream Natural Language Understanding (NLU) tasks can be obtained by fine-tuning BERT [16] with extra task-specific layers using training data that is also task-specific.

For every individual NLP task, there are established models but very few models that can simultaneously be optimized for many different NLP tasks. McCann, Bryan, et al.[23] introduced Multi Task Question Answering Network (MQAN) which is optimized for many different NLP tasks. But this does not work well for all tasks. In this paper, we introduce a model augmenting BERT with MQAN that generalize a single architecture to simultaneously tackle ten different NLP tasks: question answering, document summarization, machine translation, sentiment analysis, semantic parsing, goal oriented dialogue, natural language inference, pronoun resolution, relation extraction, semantic role labeling. An augmented pointer generator is generalized into a hierarchy, resulting in a multi-pointer generator [24] that coalesces the question context pair for our use. In traditional approaches, the model takes the underlying tasks explicitly. However, our model uses the questions and meta learning approaches to determine the underlying tasks. This enables the single model to multitask efficiently, making the model more adaptable to transfer learning and meta learning to generalize across different new related tasks. The model achieves 571.7 decaScore after applying anti-curriculum strategy followed by 74 decaScore improvements by adding BERT layer. We argue that our proposed model achieves state-of-the-art performance because of a better understanding of underlying semantic relations in question, context. The main focus of this paper is to tackle ten NLP different tasks while elevating multiple layers of representations and interaction among question, context, and answer.

II. Related Work

Transfer Learning in NLP: The ideal scenario of deep learning approaches is that there are abundant labeled training data, which have the same the distribution as the test data. However, collecting sufficient balanced data is often time-consuming, expensive, or even unrealistic for many tasks.

978-1-6654-9435-9/21/$31.00 © 2021 IEEE

So lack of a sufficiently large set of data can hamper the curve of a regularly supervised learning paradigm. This likely mishap is dealt with a promising deep learning methodology called transfer learning, which focuses on transferring the knowledge across domains. This concept makes use of outsourced data from some nearby domain in close proximity. For a few groups of tasks, the distantly supervised technique is particularly suited while supervised pretraining leverages existing tasks and datasets. In certain situations, extant tasks are chosen that might be suitable for similar downstream tasks. Gu, Jiatao, et al. [25] trained a machine translation model on a high-resource language pair and then transfer the knowledge of this instructed model to a low-resource language pair. Yang, Jie, Yue Zhang, and Fei Dong [26] pre-trained a vigorous POS tagging model and deployed the model into word segmentation. Pre-training a model [27],[28],[29] on standard large open domain question answering dataset like Stanford Question Answering Dataset (SQuAD) [30] and transferring it to a more specialized QA domain. Contemporary universal supervised pretraining tasks such as predicting definitions in a natural language inference [31], translation[32], and image captioning [33] employ different supervised tasks with large datasets to learn generalize representations. With the recent major breakthroughs in computation, it has now become feasible to pretrain multidisciplinary deep neural language models.

Meta Learning in NLP: Meta-learning provides a framework where a machine learning model gains experience over several learning episodes - often encompassing a set of similar closely related tasks and uses this experience to enhance its potential learning experience. This 'learning-to-learn' [34] paradigm has resulted in a number of benefits such as data and computation efficiency. Many successful applications have been demonstrated in areas such as unsupervised learning [35], data efficiency [36], few-shot image recognition [37, 38], hyperparameter optimization [39], reinforcement learning (RL) [40], and Neural Architecture Search (NAS) [41],[42],[43]. Recently proposed meta-optimization strategy [37, 44] that learns model parameters taking into consideration progressively effective learning of new tasks which has already shown its efficacy in various NLP tasks i.e semantic parsing [45] and low-resource machine translation [46].

Multitask Learning in NLP: For Multi Task Learning, the initial inspiration was taken from studying how human brains work. In other words, how human intelligence evolves over the period. The basic idea is to learn new things using past experiences of learning [47, 48]. For example, a person can tell whether some person is a male or a female without breaking a sweat or can approximate his or her age without much trouble. The same thing can be applied here, to learn a number of things simultaneously so that the tasks can learn from each other through sharing. There is a rising interest in the use of multitask learning to represent learning in Deep Neural Network (DNN) [47],[49],[50],[51],[52],[53]. The reason for this is that task-specific labeled data is not always available, which is a problem since supervised learning of DNN requires a large quantity of task-specific labeled data. Also, the reduction of overfitting for a specific task which results in a universal learned representation across tasks helps multitask learning big time.

III. Our approach

A. Problem Overview:

We frame all the tasks as question answering [23, 54] like Table I, all inputs have a tuple made of a context, question and answer. Usually, NLP tasks have inputs *x* and outputs *y*, and the underlying task *t* is explicitly provided. Rather than using a single representation *t* for any specific task, our model uses natural language questions and meta-learning for detecting the underlying task which allows the model to generalize to new tasks through different but related task descriptions.

Following the architecture introduced by Devlin, Jacob, et al. [16], we present a question *Q* with *m* tokens, a context *C* with *l* tokens, and an answer *A* with *n* tokens. Each of these are represented by a matrix where *i*-th row of the matrix corresponds to a *d* dimensional embedding for the *i*-th token of the sequence.

$$Q \in \mathbb{R}^{m\times d} \qquad C \in \mathbb{R}^{l\times d} \qquad A \in \mathbb{R}^{n\times d}$$

Encoder takes above three matrices as input and uses a deep stack of neural networks to produce $C_{fin} \in \mathbb{R}^{l\times d}$, $Q_{fin} \in \mathbb{R}^{l\times d}$ while capturing local and global interdependencies of both context and question. Decoder takes C_{fin} , Q_{fin} as input and pass through another neural stack to generate the answer for the corresponding question.

B. Encoder

A linear layer projects context and question matrices into a common d-dimensional space.

$$CW_1 = C_{proj} \in \mathbb{R}^{l\times d} \quad QW_1 = Q_{proj} \in \mathbb{R}^{m\times d} \tag{1}$$

These projected representations work as input to a shared Bidirectional Long Short-Term Memory Network (BiLSTM) [55],[56].

$$\left.\begin{aligned} BiLSTM(C_{proj}) = C_{out} \in \mathbb{R}^{l\times d} \\ BiLSTM(Q_{proj}) = Q_{out} \in \mathbb{R}^{m\times d} \end{aligned}\right\}$$

After that, we applied dual coattention to compute weighted information from one sequence that is relevant to a single token. Next, we used Scaled Dot-Product Attention [12] to give attention to a particular portion in each sequence. We calculated dot products of the query P with all keys Q - values R, divided by $\sqrt{d}$ where *d* is the dimension and applied softmax function to get weightage on values.

$$\text{Attention}(\tilde{P}, \tilde{Q}, \tilde{R}) = \text{softmax}(\frac{\tilde{P}\tilde{Q}}{\sqrt{d}})\tilde{R} \tag{2}$$

Task	Dateset	Question	Context	Answer
Question Answering	SQuAD	Which company produces the iPod?	The iPod is a line of portable media players and multi-purpose pocket computers designed and marketed by Apple Inc.	Apple
Document Summarization	CNN/DM	What is the summary?	An American woman died aboard a cruise ship that docked at Rio de Janeiro on Tuesday, the same ship on which 86 passengers previously fell ill, according to the state-run Brazilian news agency, Agencia Brasil....	The elderly woman suffered from diabetes and hypertension
Machine Translation	IWSLT	What is the translation from English to German?	The train was late.	Der Zug war spät.
Sentiment Analysis	SST	Is this sentence positive or negative?	The actors are fantastic.	positive
Semantic Parsing	WikiSQL	What is the translation from English to SQL?	The table has column names... Who was drafted with the 3rd pick of the 1st round?	SELECT Player from mytable WHERE Rnd = 1 AND Pick = 3
Goal Oriented Dialogue	WOZ	What is the change in dialogue state?	Meghna serves moderately priced Indian food and is in the West part of town.	price range: moderate
Natural Language Inference	MNLI	Hypothesis: People formed a line at the end of Pennsylvania Avenue. Entailment, neutral, or contradiction?	Premise: At the other end of Pennsylvania Avenue, people began to line up for a White House tour.	Entailment
Pronoun Resolution	MWSC	Who was upset? Jim/Kevin?	Jim yelled at Kevin because he was so upset	Jim
Relation Extraction	QA-ZRE	Who is Angela Merkel married to?	Angela Merkel's second and current husband is quantum chemist and professor Joachim Sauer, who has largely...	Joachim Sauer
Semantic Role Labeling	QA-SRL	Who finished something?	UCD finished the 2006 championship as Dublin champions , by beating St Vincents in the final.	UCD

TABLE I: Overview of framing ten tasks as Question Answering with one example from each task

Multi-head attention entitles the model to jointly process information from different representation subspaces.

$$\left.\begin{aligned}\text{MultiHead}(P,Q,R) &= [head_1, head_2 \cdot \cdot\cdot; head_h]W_o \\ head_j &= \text{Attention}(XW_j^X, YW_j^Y, ZW_j^Z)\end{aligned}\right\}$$

Finally to aggregate all of the information acquired across different timestamps, we applied another BiLSTM. The final representation of both context and question $C_{fin} \in \mathbb{R}^{l\times d}$, $Q_{fin} \in \mathbb{R}^{m\times d}$ are given to the decoder to generate answer.

C. Decoder:

The decoder go about projecting answer embeddings onto a d-dimensional space:

$$AW_1 = A_{proj} \in \mathbb{R}^{n\times d} \tag{3}$$

Next, we apply a deep stack of Multi-head Decoder Attention [23], recurrent context state, context and question attention α_t^C, α_t^Q that allows the decoder to focus on the encoded information relevant to timestamp t. The new tokens that are not in question or context must be generated in our model. We, therefore, give our model access to further v vocabulary tokens. In the context, question, and external vocabulary we get distributions over tokens as follows.

$$\sum_{i:c_i = w_t} (\alpha_t^C)_i = p_c(w_t) \in \mathbb{R}^n$$

$$\sum_{i:q_i = w_t} (\alpha_t^Q)_i = p_q(w_t) \in \mathbb{R}^m$$

IV. Result and Analysis

A. Implementation Details:

As discussed earlier in section III.A, we have used the PyTorch implementation of BERT [1]. For the input of LSTMs, decoder layers, and layers following coattention, we used 0.2 dropout. We trained the models using Adamax [57] optimizer with $(\beta 1, \beta 2, \epsilon) = (0.9, 0.98, 10^{-9})$ and a warmup schedule which expands the learning rate from 0 to 2.5×10^{-3} linearly over iterations. The complete dataset of our model consists of 10 publicly available competitive datasets for each task mentioned in table I.

[1]https://github.com/huggingface/transformers

B. Experiment:

We evaluate our model with a number of baselines combining sequence-to-sequence model [58],[59],[60] with pointer networks [61],[62],[63],[64, 65], advance attention mechanism [66] curriculum learning [67] and decaNLP [23]. Table II shows the efficacy of our model. We first experimented with pointer-generator sequence-to-sequence model [24]. We concatenated the context and question into one input sequence for seq2seq . But seq2seq model does not perform well in SQuAD dataset because it loses some information for concatenation. To coalesce information from both context and question, we extended seq2seq model with self-attentive (w/ SAtt) encoder and decoder layers [12] which improves performance on SQuAD, WikiSQL, QA-SRL.

Dataset	w/SAtt	+CAtt	+QPtr	+ACurr	+BERT
SQuAD	66.8	71.8	70.8	74.3	81.4
IWSLT	13.6	9.0	16.1	13.7	27.8
CNN/DM	14.0	15.7	23.9	24.6	29.4
MNLI	69.0	70.4	70.5	69.2	74.3
SST	84.7	86.5	86.2	86.4	87.1
QA-SRL	75.1	76.1	75.8	77.6	84.9
QA-ZRE	31.7	28.5	28.0	34.7	47.9
WOZ	82.8	75.1	80.6	84.1	82.1
WikiSQL	64.8	62.9	62.0	58.7	77.6
MWSC	43.9	37.8	48.8	48.4	53.2
decaScore	546.4	533.8	562.7	571.7	**645.7**

TABLE II: Performance Comparison of Different baselines (MultiTask Training)

Next, we try to explore the question and context into two individual sequences and extend seq2seq model with a coattention mechanism (+CAtt). Although this model performs better than the previous model for QA-SRL and SQuAD, but the performance significantly falls down for MNLI and MWSC datasets. The answer can be explicitly extracted from the input sequence for these two specific datasets. As the pointer generator mechanism can directly copy from the

input therefore previous models had concatenated context and questions they performed better. To get rid of this, we added the question pointer (+QPtr) layer to the previous model, which boosts performance on both MWSC and MNLI comparing with previous baselines. This model reaches 62.0% database execution accuracy for the WikiSQL dataset which surpasses all previous baselines.

For training in multitask setting, we have adopted anti-curriculum strategy. Lastly, we have added a pre-trained language model BERT that significantly outperforms all previous baselines on all datasets in multi task setting. This proposed model is efficacious because it increases the model's capacity to hold semantic information more about questions and context as well.

C. Analysis:

For generating an answer, our model selects one of the three available options: pointing to the question, generating from the vocabulary, and pointing to the context. We did not train the model in which approach to process, it learns to switch among options. All the tokens needed to answer the question of QA-SRL, WikiSQL and SQuAD can be directly found in context. So the model mostly copies from the context for generating answers of these datasets. As documents summary consist mostly of the tokens present in context with few tokens outside the context, for giving answer our model copies tokens from context for CNN/DM dataset.

The model advocates generating the answer from the vocabulary for the questions of IWSLT and WOZ dataset because German words and dialogue state fields can rarely be found in the context. For MNLI, SST, and MWSC datasets, the model prefers the question pointer because the question contains the tokens for acceptable classes. Our model learns to generalize beyond the task specific domains while learning the representations that make learning completely new tasks easier.

V. Conclusion:

In this paper, we introduced a new approach that can simultaneously optimize for ten different NLP tasks by unifying them as question answering. Without any task-specific modules, we trained our model on ten tasks together and our experimental result demonstrates the effectiveness of our proposed model. The addition of all of these potential techniques opens up a significantly broader domain of general architecture designs for natural language processing tasks. This motivates the application of pre-trained models, multitask learning, and meta learning to tackle disparate but related difficult tasks.

References

[1] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, 2008, pp. 1247–1250.

[2] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

[3] L. Dong and M. Lapata, "Coarse-to-fine decoding for neural semantic parsing," *arXiv preprint arXiv:1805.04793*, 2018.

[4] D. Guo, D. Tang, N. Duan, M. Zhou, and J. Yin, "Dialog-to-action: Conversational question answering over a large-scale knowledge base," in *Advances in Neural Information Processing Systems*, 2018, pp. 2942–2951.

[5] R. Jia and P. Liang, "Data recombination for neural semantic parsing," *arXiv preprint arXiv:1606.03622*, 2016.

[6] S. Reddy, M. Lapata, and M. Steedman, "Large-scale semantic parsing without question-answer pairs," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 377–392, 2014.

[7] L. Dong and M. Lapata, "Language to logical form with neural attention," *arXiv preprint arXiv:1601.01280*, 2016.

[8] C. Xiao, M. Dymetman, and C. Gardent, "Sequence-based structured prediction for semantic parsing," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2016, pp. 1341–1350.

[9] S. Wang and J. Jiang, "Machine comprehension using match-lstm and answer pointer," *arXiv preprint arXiv:1608.07905*, 2016.

[10] M. Seo, A. Kembhavi, A. Farhadi, and H. Hajishirzi, "Bidirectional attention flow for machine comprehension iclr," 2017.

[11] C. Xiong, V. Zhong, and R. Socher, "Dcn+: Mixed objective and deep residual coattention for question answering," *arXiv preprint arXiv:1711.00106*, 2017.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[13] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018.

[14] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," *arXiv preprint arXiv:1802.05365*, 2018.

[15] A. M. Dai and Q. V. Le, "Semi-supervised sequence learning," in *Advances in neural information processing systems*, 2015, pp. 3079–3087.

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

[17] M. Ott, S. Edunov, D. Grangier, and M. Auli, "Scaling neural machine translation," *arXiv preprint arXiv:1806.00187*, 2018.

[18] E. Strubell, P. Verga, D. Andor, D. Weiss, and A. McCallum, "Linguistically-informed self-attention for semantic role labeling," *arXiv preprint arXiv:1804.08199*, 2018.

[19] A. Williams, N. Nangia, and S. R. Bowman, "A broad-coverage challenge corpus for sentence understanding through inference," *arXiv preprint arXiv:1704.05426*, 2017.

[20] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning, "A large annotated corpus for learning natural language inference," *arXiv preprint arXiv:1508.05326*, 2015.

[21] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf*, 2018.

[22] J. Gao, M. Galley, L. Li *et al.*, "Neural approaches to conversational ai," *Foundations and Trends® in Information Retrieval*, vol. 13, no. 2-3, pp. 127–298, 2019.

[23] B. McCann, N. S. Keskar, C. Xiong, and R. Socher, "The natural language decathlon: Multitask learning as question answering," *arXiv preprint arXiv:1806.08730*, 2018.

[24] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[25] B. Zoph, D. Yuret, J. May, and K. Knight, "Transfer learning for low-resource neural machine translation," *arXiv preprint arXiv:1604.02201*, 2016.

[26] J. Yang, Y. Zhang, and F. Dong, "Neural word segmentation with rich pretraining," *arXiv preprint arXiv:1704.08960*, 2017.

[27] D. Golub, P.-S. Huang, X. He, and L. Deng, "Two-stage synthesis networks for transfer learning in machine comprehension," *arXiv preprint arXiv:1706.09789*, 2017.

[28] S. Min, M. Seo, and H. Hajishirzi, "Question answering through transfer learning from large fine-grained supervision data," *arXiv preprint arXiv:1702.02171*, 2017.

[29] G. Wiese, D. Weissenborn, and M. Neves, "Neural domain adaptation for biomedical question answering," *arXiv preprint arXiv:1706.03610*, 2017.

[30] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "Squad: 100,000+ questions for machine comprehension of text," *arXiv preprint arXiv:1606.05250*, 2016.

[31] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[32] B. McCann, J. Bradbury, C. Xiong, and R. Socher, "Learned in translation: Contextualized word vectors," in *Advances in Neural Information Processing Systems*, 2017, pp. 6294–6305.

[33] D. Kiela, A. Conneau, A. Jabri, and M. Nickel, "Learning visually grounded sentence representations," *arXiv preprint arXiv:1707.06320*, 2017.

[34] S. Thrun and L. Pratt, "Learning to learn: Introduction and overview," in *Learning to learn*. Springer, 1998, pp. 3–17.

[35] L. Metz, N. Maheswaranathan, B. Cheung, and J. Sohl-Dickstein, "Meta-learning update rules for unsupervised representation learning," *arXiv preprint arXiv:1804.00222*, 2018.

[36] R. Houthooft, Y. Chen, P. Isola, B. Stadie, F. Wolski, O. J. Ho, and P. Abbeel, "Evolved policy gradients," in *Advances in Neural Information Processing Systems*, 2018, pp. 5400–5409.

[37] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1126–1135.

[38] J. Snell, K. Swersky, and R. Zemel, "Prototypical networks for few-shot learning," in *Advances in neural information processing systems*, 2017, pp. 4077–4087.

[39] L. Franceschi, P. Frasconi, S. Salzo, R. Grazzi, and M. Pontil, "Bilevel programming for hyperparameter optimization and meta-learning," *arXiv preprint arXiv:1806.04910*, 2018.

[40] F. Alet, M. F. Schneider, T. Lozano-Perez, and L. P. Kaelbling, "Meta-learning curiosity algorithms," *arXiv preprint arXiv:2003.05325*, 2020.

[41] H. Liu, K. Simonyan, and Y. Yang, "Darts: Differentiable architecture search," *arXiv preprint arXiv:1806.09055*, 2018.

[42] E. Real, A. Aggarwal, Y. Huang, and Q. V. Le, "Regularized evolution for image classifier architecture search," in *Proceedings of the aaai conference on artificial intelligence*, vol. 33, 2019, pp. 4780–4789.

[43] B. Zoph and Q. V. Le, "Neural architecture search with reinforcement learning," *arXiv preprint arXiv:1611.01578*, 2016.

[44] A. Nichol, J. Achiam, and J. Schulman, "On first-order meta-learning algorithms," *arXiv preprint arXiv:1803.02999*, 2018.

[45] P.-S. Huang, C. Wang, R. Singh, W.-t. Yih, and X. He, "Natural language to structured query generation via meta-learning," *arXiv preprint arXiv:1803.02400*, 2018.

[46] J. Gu, Y. Wang, Y. Chen, K. Cho, and V. O. Li, "Meta-learning for low-resource neural machine translation," *arXiv preprint arXiv:1808.08437*, 2018.

[47] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. Aug, pp. 2493–2537, 2011.

[48] Y. Zhang and Q. Yang, "A survey on multi-task learning," *arXiv preprint arXiv:1707.08114*, 2017.

[49] X. Liu, J. Gao, X. He, L. Deng, K. Duh, and Y.-Y. Wang, "Representation learning using multi-task deep neural networks for semantic classification and information retrieval," 2015.

[50] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," *arXiv preprint arXiv:1511.06114*, 2015.

[51] Y. Xu, X. Liu, Y. Shen, J. Liu, and J. Gao, "Multi-task learning with sample re-weighting for machine reading comprehension," *arXiv preprint arXiv:1809.06963*, 2018.
[52] H. Guo, R. Pasunuru, and M. Bansal, "Soft layer-specific multi-task summarization with entailment and question generation," *arXiv preprint arXiv:1805.11004*, 2018.
[53] S. Ruder, J. Bingel, I. Augenstein, and A. Søgaard, "Latent multi-task architecture learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 4822–4829.
[54] A. Kumar, O. Irsoy, P. Ondruska, M. Iyyer, J. Bradbury, I. Gulrajani, V. Zhong, R. Paulus, and R. Socher, "Ask me anything: Dynamic memory networks for natural language processing," in *International conference on machine learning*, 2016, pp. 1378–1387.
[55] A. Graves and J. Schmidhuber, "Framewise phoneme classification with bidirectional lstm and other neural network architectures," *Neural networks*, vol. 18, no. 5-6, pp. 602–610, 2005.
[56] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[58] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.
[59] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.
[60] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
[61] S. J. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture architecture," Feb. 18 2020, uS Patent 10,565,493.
[62] C. Gulcehre, S. Ahn, R. Nallapati, B. Zhou, and Y. Bengio, "Pointing the unknown words," *arXiv preprint arXiv:1603.08148*, 2016.
[63] J. Gu, Z. Lu, H. Li, and V. O. Li, "Incorporating copying mechanism in sequence-to-sequence learning," *arXiv preprint arXiv:1603.06393*, 2016.
[64] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.
[65] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," in *NIPS*, 2015, pp. 2692–2700. [Online]. Available: https://arxiv.org/pdf/1506.03134.pdf
[66] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," *arXiv preprint arXiv:1611.01604*, 2016.
[67] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning (icml)," *Google Scholar Google Scholar Digital Library Digital Library*, 2009.