Meta-Learning Improves Lifelong Relation Extraction

Abiola Obamuyide
Department of Computer Science
University of Sheffield
avobamuyide1@sheffield.ac.uk

Andreas Vlachos
Dept. of Computer Science and Technology
University of Cambridge
andreas.vlachos@cst.cam.ac.uk

Abstract

Most existing relation extraction models assume a fixed set of relations and are unable to adapt to exploit newly available supervision data to extract new relations. In order to alleviate such problems, there is the need to develop approaches that make relation extraction models capable of continuous adaptation and learning. We investigate and present results for such an approach, based on a combination of ideas from lifelong learning and optimization-based meta-learning. We evaluate the proposed approach on two recent lifelong relation extraction benchmarks, and demonstrate that it markedly outperforms current state-of-the-art approaches.

1 Introduction

The majority of existing supervised relation extraction models can only extract a fixed set of relations which has been specified at training time. They are unable to detect an evolving set of novel relations observed after training without substantial retraining, which can be computationally expensive and may lead to catastrophic forgetting of previously learned relations. Zero-shot relation extraction approaches (Rocktäschel et al., 2015; Demeester et al., 2016; Levy et al., 2017; Obamuyide and Vlachos, 2018) can extract unseen relations, but at lower performance levels, and are unable to continually exploit newly available supervision to improve performance without considerable retraining. These limitations also extend to approaches to extracting relations in other limited supervision settings, for instance in the one-shot setting (Obamuyide and Vlachos, 2017). It is therefore desirable for relation extraction models to have the capability to learn continuously without catastrophic forgetting of previously learned relations. This would enable them exploit newly available supervision to both identify novel relations and improve performance without substantial retraining.

Recently, Wang et al. (2019) introduced an embedding alignment approach to enable continual learning for relation extraction models. They consider a setting with streaming tasks, where each task consists of a number of distinct relations, and proposed to align the representation of relation instances in the embedding space to enable continual learning of new relations without forgetting knowledge from past relations. While they obtained promising results, a key weakness of the approach is that the use of an alignment model introduces additional parameters to already over-parameterized relation extraction models, which may in turn lead to an increase in the quantity of supervision required for training. In addition, the approach can only align embeddings between observed relations, and does not have any explicit objective that encourages the model to transfer and exploit knowledge gathered from previously observed relations to facilitate the efficient learning of yet to be observed relations.

In this work, we extend the work of Wang et al. (2019) by exploiting ideas from both lifelong learning and meta-learning. We propose to consider lifelong relation extraction as a meta-learning challenge, to which the machinery of current optimization-based meta-learning algorithms can be applied. Unlike the use of a separate alignment model as proposed in Wang et al. (2019), the proposed approach does not introduce additional parameters. In addition, the proposed approach is more data efficient since it explicitly optimizes for the transfer of knowledge from past relations, while avoiding the catastrophic forgetting of previously learned relations. Empirically, we evaluate on lifelong versions of the datasets by Bordes et al. (2015) and Han et al. (2018) and demonstrate con-

Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019), pages 224–229
Florence, Italy, August 2, 2019.

siderable performance improvements over prior state-of-the-art approaches.

2 Background

Lifelong Learning In the lifelong learning setting, also referred to as continual learning (Ring, 1994; Thrun, 1996; Zhao and Schmidhuber, 1996), a model f_θ is presented with a sequence of tasks $\{\mathcal{T}_t\}_{t=1,2,3..,T}$, one task per round, and the goal is to learn model parameters $\{\theta_t\}_{t=1,2,3,..,T}$ with the best performance on the observed tasks. Each task $\mathcal{T}$ can be a conventional supervised task with its own distinct train ($\mathcal{T}^{train}$), development ($\mathcal{T}^{dev}$) and test ($\mathcal{T}^{test}$) splits. At each round t, the model is allowed to exploit knowledge gained from the previous $t-1$ tasks to enhance performance on the current task. In addition, the model is also allowed to have a small-sized buffer memory B, which can be used to store a limited amount of data from previously observed tasks. A prominent line of work in lifelong learning research is developing approaches that enable models learn new tasks without forgetting knowledge from previous tasks, i.e. avoiding catastrophic forgetting of old tasks (McCloskey and Cohen, 1989; Ratcliff, 1990; McClelland et al., 1995; French, 1999). Approaches proposed to address this problem include memory-based approaches (Lopez-Paz and Ranzato, 2017; Rebuffi et al., 2017; Chaudhry et al., 2019); parameter consolidation approaches (Kirkpatrick et al., 2017; Zenke et al., 2017); and dynamic model architecture approaches (Xiao et al., 2014; Rusu et al., 2016; Fernando et al., 2017).

Meta-Learning Meta-learning, or learning to learn (Schmidhuber, 1987; Naik and Mammone, 1992; Thrun and Pratt, 1998), aims to develop algorithms that learn a generic knowledge of how to solve tasks from a given distribution of tasks, by generalizing from solving related tasks from that distribution. Given tasks $\mathcal{T}$ sampled from a distribution of tasks $p(\mathcal{T})$, and a learner model $f(\mathbf{x};\theta)$ parameterized by θ, gradient-based meta-learning methods, such as *MAML* (Finn et al., 2017), learn a prior initialization of the parameters of the model which, at meta-test time, can be quickly adapted to achieve good performance on a new task using a few steps of gradient descent. During adaptation to the new task, the model parameters θ are updated to task-specific parameters θ' with good performance on the task. Formally, the meta-learning algorithms optimize for the meta-objective:

$$\min_\theta \mathbb{E}_{\mathcal{T}\sim p(\mathcal{T})}\left[\mathcal{L}_\mathcal{T}\left(\theta'\right)\right] = \min_\theta \mathbb{E}_{\mathcal{T}\sim p(\mathcal{T})}\left[\mathcal{L}_\mathcal{T}\left(\mathcal{U}\left(\mathcal{D}_\mathcal{T};\theta\right)\right)\right] \quad (1)$$

where $\mathcal{L}_\mathcal{T}$ is the loss and $\mathcal{D}_\mathcal{T}$ is training data from task $\mathcal{T}$, and $\mathcal{U}$ is a fixed gradient descent learning rule, such as vanilla SGD. While these algorithms were proposed and evaluated in the context of few-shot learning, here we demonstrate their effectiveness when utilized in the lifelong learning setting for relation extraction, following similar intuition as recent work by Finn et al. (2019).

3 Meta-Learning for Lifelong Relation Extraction

It can be inferred from the previous section that a lot of lifelong learning research has focused on approaches to avoid catastrophic forgetting (i.e. negative backward transfer of knowledge) while recent meta-learning studies have focused on effective approaches for positive forward transfer of knowledge (for few-shot tasks). Given the complementary strengths of the approaches from the two learning settings, we propose to embed meta-learning into the lifelong learning process for relation extraction.

While we can utilize the MAML algorithm to directly optimize the meta-objective in Equation 1 for our purpose, doing so requires the computation of second-order derivatives, which can be computationally expensive. Nichol et al. (2018) proposed *REPTILE*, a first-order alternative to MAML, which uses only first-order derivatives. Similar to MAML, REPTILE works by repeatedly sampling tasks, training on those tasks and moving the initialization towards the adapted weights on those tasks. Here we adopt the REPTILE algorithm for meta-learning. Our algorithm for lifelong relation extraction is illustrated in Algorithm 1.

We start by randomly initializing the parameters of the relation extraction model (the *learner*) (line 1). Then, as new tasks arrive, we augment their training set with randomly sampled task exemplars from the buffer memory *B* (lines 2-9). We then sample a batch of relations from the augmented training set (line 10). Then for each sampled relation $\mathcal{R}_i$, we sample a batch of supervision instances $\mathcal{D}^{train}_{\mathcal{R}_i}$ from its training set (line 11-12).

We then obtain the adapted model parameters θ_t^i on the relation by first computing the gradient of the training loss on the sampled relation instances (line 13) and backpropagating the gradients with a gradient-based optimization algorithm (such as *SGD* or *Adagrad* (Duchi et al., 2011)) (line 14). At the end of the learning iteration, the adapted parameters on all sampled relations in the batch are averaged, and an update is made on the task parameters θ_t (line 16). This is done until convergence on the current task, after which exemplars of the current task are added to the buffer memory (line 18). Task exemplars are obtained by first clustering all training instances of the current task into 50 clusters using K-Means, then selecting an instance from each cluster with a representation closest to the cluster prototype. Finally, the model parameters are updated to the current task's adapted parameters (line 19).

Algorithm 1 Meta-Learning for Lifelong Relation Extraction (*MLLRE*)

Require: Stream of incoming tasks $\mathcal{T}_1, \mathcal{T}_2, \mathcal{T}_3, ...$
Require: Relation extraction function f_θ
Require: Optimization algorithm (e.g. *SGD*)
Require: Step size ϵ, learning rate α
Require: Buffer memory B

1: Randomly initialize θ
2: **while** there are still tasks **do**
3: Retrieve next task $\mathcal{T}_t$ from stream
4: Initialize $\theta_t \leftarrow \theta$
5: **repeat**
6: **if** B is not empty **then**
7: Retrieve exemplars $\mathcal{E}$ of random task from B
8: Update task training set $\mathcal{D}_t^{train} = \mathcal{D}_t^{train} \cup \mathcal{E}$
9: **end if**
10: Sample random relations $\{\mathcal{R}_i\}_{i=1}^N$ from $\mathcal{D}_t^{train}$
11: **for** each $\mathcal{R}_i$ **do**
12: Sample train instances $\mathcal{D}_{\mathcal{R}_i}^{train}$ of $\mathcal{R}_i$
13: Evaluate $\nabla_{\theta_t}\mathcal{L}_{\mathcal{R}_i}(f_{\theta_t})$ using $\mathcal{D}_{\mathcal{R}_i}^{train}$
14: Compute adapted parameters: $\theta_t^i = SGD(\theta_t, \nabla_{\theta_t}\mathcal{L}_{\mathcal{R}_i}(f_{\theta_t}), \alpha)$
15: **end for**
16: Update task parameters:
$$\theta_t = \theta_t - \epsilon\frac{1}{N}\sum_{i=1}^{N}(\theta_t^i - \theta_t)$$
17: **until** *Convergence*
18: Add exemplars of $\mathcal{T}_t$ to B
19: Update $\theta \leftarrow \theta_t$
20: **end while**

4 Relation Classification Model

In principle the learner model f_θ could be any gradient-optimized relation extraction model. In order to use the same number of parameters and ensure fair comparison to Wang et al. (2019), we adopt as the relation extraction model f_θ the Hierarchical Residual BiLSTM (*HR-BiLSTM*) model of Yu et al. (2017), which is the same model used by Wang et al. (2019) for their experiments. The *HR-BILSTM* is a relation classifier which accepts as input a sentence and a candidate relation, then utilizes two Bidirectional Long Short-Term Memory (Hochreiter and Schmidhuber, 1997; Graves and Schmidhuber, 2005) (BiLSTM) units with shared parameters to process the Glove (Pennington et al., 2014) embeddings of words in the sentence and relation names, then selects the relation with the maximum cosine similarity to the sentence as its response.

Method	*FewRel*		*SimpleQuestions*	
	$ACC_{w.}$	$ACC_{a.}$	$ACC_{w.}$	$ACC_{a.}$
Origin	0.189	0.208	0.632	0.569
GEM	0.492	0.598	0.841	0.796
AGEM	0.361	0.425	0.776	0.722
EWC	0.271	0.302	0.672	0.590
EA-EMR (Full)	0.566	0.673	0.878	0.824
EA-EMR (w/o Sel.)	0.564	0.674	0.857	0.812
EA-EMR (w/o Align.)	0.526	0.632	0.869	0.820
EMR	0.510	0.620	0.852	0.808
MLLRE	**0.602**	**0.741**	**0.880**	**0.842**

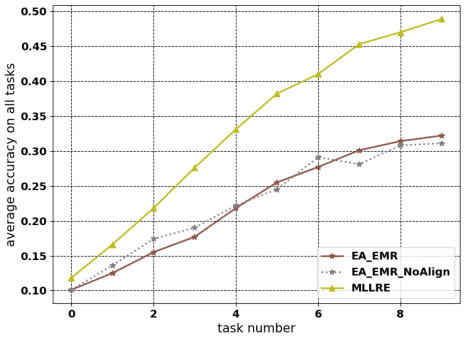
Table 1: Accuracy on the test set of all tasks ACC_{whole} (denoted $ACC_{w.}$) and average accuracy on the test set of only observed tasks ACC_{avg} (denoted $ACC_{a.}$) on the *Lifelong FewRel* and *Lifelong SimpleQuestions* datasets. Best results are in bold. Except for *MLLRE*, results for other models are obtained from Wang et al. (2019).

Hyperparameters Apart from the hyperparameters specific to meta-learning (such as the step size ϵ), all other hyperparameters we use for the learner model are the same as used by Wang et al. (2019). We also use the same buffer memory size (50) for each task. Note that the meta-learning algorithm uses SGD as the update rule ($\mathcal{U}$), and does not add any additional trainable parameters to the learner model.

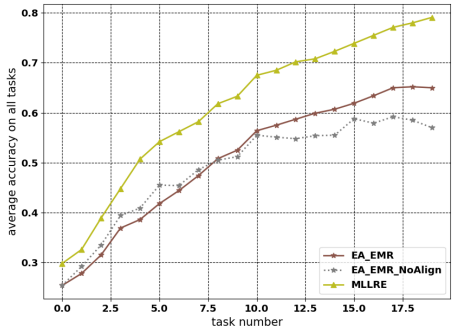
5 Experiments

5.1 Setup

We conduct experiments in two settings. In the full supervision setting, we provide all models with all supervision available in the training set of each task. In the second, we limit the amount of supervision for each task to measure how the models are able to cope with limited supervision. Each experiment is run five (5) times and we report the

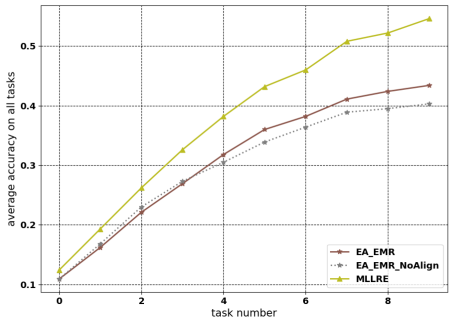


(a)

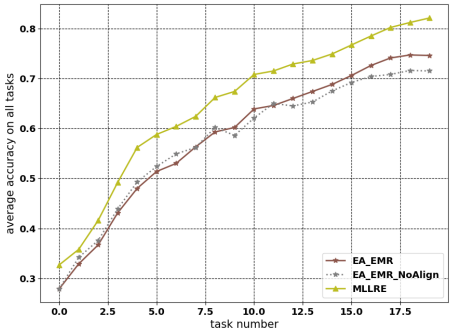


(b)

Figure 1: Results obtained using 100 training instances for each task on (a) *Lifelong FewRel* and (b) *Lifelong SimpleQuestions* datasets.



(a)



(b)

Figure 2: Results obtained using 200 training instances for each task on (a) *Lifelong FewRel* and (b) *Lifelong SimpleQuestions* datasets.

average result.

5.2 Datasets

We conduct experiments on *Lifelong FewRel* and *Lifelong SimpleQuestions* datasets, both introduced in Wang et al. (2019). *Lifelong FewRel* is derived from the *FewRel* (Han et al., 2018) dataset, by partitioning its 80 relations into 10 distinct clusters made up of 8 relations each, with each cluster serving as a task where a sentence must be labeled with the correct relation. The 8 relations in each cluster were obtained by clustering the averaged Glove word embeddings of the relation names in the *FewRel* dataset. Each instance of the dataset contains a sentence, the relation it expresses and a set of randomly sampled negative relations. *Lifelong SimpleQuestions* was similarly obtained from the *SimpleQuestions* (Bordes et al., 2015) dataset, and is made up of 20 clusters of relations, with each cluster serving as a task.

5.3 Evaluation Metrics

We report two measures, ACC_{whole} and ACC_{avg}, both introduced in Wang et al. (2019). ACC_{whole} measures accuracy on the test set of all tasks and gives a balanced measure of model performance on both observed (seen) and unobserved (unseen) tasks, and is the primary metric we report for all experiments. We also report ACC_{avg}, which measures the average accuracy on the test set of only observed (seen) tasks.

5.4 Results and Discussion

Full Supervision Results Table 1 gives both the ACC_{whole} and ACC_{avg} results of our approach compared to other approaches including Episodic Memory Replay (EMR) and its various embedding-aligned variants *EA-EMR* as proposed in Wang et al. (2019). Across all metrics, our approach outperforms the previous approaches, demonstrating its effectiveness in this setting. This result is likely because our approach is able to efficiently learn new relations by exploiting knowledge from previously observed relations.

Limited Supervision Results The aim of our limited supervision experiments is to compare the use of an alignment module as proposed by Wang et al. (2019) to using our approach when only limited supervision is available for all tasks. We compare three approaches, Full *EA-EMR* (which uses their alignment module), its variant without the alignment module (*EA-EMR_NoAlign*) and our approach (*MLLRE*). Figures 1(a) and 1(b) show results obtained using 100 supervision instances for each task on *Lifelong FewRel* and *Lifelong SimpleQuestions*. Figures 2(a) and 2(b) show the corresponding plots using 200 supervision instances for each task. From the figures, we observe that the use of a separate alignment model results in only minor gains when supervision for the tasks is limited, whereas the use of our approach leads to wide gains on both datasets.

In summary, because our approach explicitly encourages the model to learn to share and transfer knowledge between relations (by means of the meta-learning objective), the model is able to learn to exploit common structures across relations in different tasks to efficiently learn new relations over time. This leads to the performance improvements obtained by our approach.

6 Conclusion

We investigated the effectiveness of utilizing a gradient-based meta-learning algorithm within a lifelong learning setting to enable relation extraction models that are able to learn continually. We show the effectiveness of this approach, both when provided full supervision for new tasks and when provided limited supervision for new tasks, and demonstrated that the proposed approach outperformed current state-of-the-art approaches.

Acknowledgements

The authors acknowledge support from the EU H2020 SUMMA project (grant agreement number 688139).

References

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *CoRR*, abs/1506.02075.

Arslan Chaudhry, MarcAurelio Ranzato, Marcus Rohrbach, and Mohamed Elhoseiny. 2019. Efficient lifelong learning with a-gem. In *International Conference on Learning Representations*.

Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2016. Lifted Rule Injection for Relation Embeddings. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1389–1399.

John Duchi, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Chrisantha Fernando, Dylan Banarse, Charles Blundell, Yori Zwols, David Ha, Andrei A Rusu, Alexander Pritzel, and Daan Wierstra. 2017. Pathnet: Evolution channels gradient descent in super neural networks. *arXiv preprint arXiv:1701.08734*.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. 2017. Model-agnostic meta-learning for fast adaptation of deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1126–1135, International Convention Centre, Sydney, Australia.

Chelsea Finn, Aravind Rajeswaran, Sham Kakade, and Sergey Levine. 2019. Online meta-learning. In *Proceedings of the 36th International Conference on Machine Learning*.

Robert M French. 1999. Catastrophic forgetting in connectionist networks. *Trends in cognitive sciences*, 3(4):128–135.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM networks. In *Proceedings of the International Joint Conference on Neural Networks*, volume 4, pages 2047–2052.

Xu Han, Hao Zhu, Pengfei Yu, Ziyun Wang, Maosong Sun, Yuan Yao, and Zhiyuan Liu. 2018. FewRel : A Large-Scale Supervised Few-Shot Relation Classification Dataset with State-of-the-Art Evaluation. In *Emnlp*, pages 4803–4809, Brussels, Belgium. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and Others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Omer Levy, Minjoon Seo, Eunsol Choi, and Luke Zettlemoyer. 2017. Zero-shot relation extraction via reading comprehension. In *Proceedings of the 21st Conference on Computational Natural Language*

Learning (CoNLL 2017), pages 333–342, Vancouver, Canada. Association for Computational Linguistics.

David Lopez-Paz and Marc'Aurelio Ranzato. 2017. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*.

James L McClelland, Bruce L McNaughton, and Randall C O'reilly. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3):419.

Michael McCloskey and Neal J Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, volume 24, pages 109–165.

Devang K Naik and R J Mammone. 1992. Meta-neural networks that learn by learning. In *Proceedings of the International Joint Conference on Neural Networks*, volume 1, pages 437–442.

Alex Nichol, Joshua Achiam, and John Schulman. 2018. On first-order meta-learning algorithms. *CoRR*, abs/1803.02999.

Abiola Obamuyide and Andreas Vlachos. 2017. Contextual pattern embeddings for one-shot relation extraction. In *Proceedings of the NeurIPS 2017 Workshop on Automated Knowledge Base Construction (AKBC)*.

Abiola Obamuyide and Andreas Vlachos. 2018. Zero-shot relation classification as textual entailment. In *Proceedings of the EMNLP 2018 Workshop on Fact Extraction and VERification (FEVER)*, pages 72–78, Brussels, Belgium. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Roger Ratcliff. 1990. Connectionist models of recognition memory: constraints imposed by learning and forgetting functions. *Psychological review*, 97(2):285.

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H Lampert. 2017. icarl: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2001–2010.

Mark Bishop Ring. 1994. *Continual learning in reinforcement environments*. Ph.D. thesis, University of Texas at Austin Austin, Texas 78712.

Tim Rocktäschel, Sameer Singh, and Sebastian Riedel. 2015. Injecting Logical Background Knowledge into Embeddings for Relation Extraction. *North American Association for Computational Linguistics*, pages 1119–1129.

Andrei A Rusu, Neil C Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. 2016. Progressive neural networks. *arXiv preprint arXiv:1606.04671*.

Jurgen Schmidhuber. 1987. Evolutionary principles in self-referential learning. *On learning how to learn: The meta-meta-... hook.) Diploma thesis, Institut f. Informatik, Tech. Univ. Munich*.

Sebastian Thrun. 1996. Is learning the n-th thing any easier than learning the first? In *Advances in neural information processing systems*, pages 640–646.

Sebastian Thrun and Lorien Pratt. 1998. Learning to Learn: Introduction and Overview. In *Learning to Learn*, pages 3–17. Springer US, Boston, MA.

Hong Wang, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. 2019. Sentence embedding alignment for lifelong relation extraction. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 796–806, Minneapolis, Minnesota. Association for Computational Linguistics.

Tianjun Xiao, Jiaxing Zhang, Kuiyuan Yang, Yuxin Peng, and Zheng Zhang. 2014. Error-driven incremental learning in deep convolutional neural network for large-scale image classification. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 177–186.

Mo Yu, Wenpeng Yin, Kazi Saidul Hasan, Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2017. Improved neural relation detection for knowledge base question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–581, Vancouver, Canada. Association for Computational Linguistics.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual Learning Through Synaptic Intelligence. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 3987–3995, International Convention Centre, Sydney, Australia.

Jieyu Zhao and Jurgen Schmidhuber. 1996. Incremental self-improvement for life-time multi-agent reinforcement learning. In *From Animals to Animats 4: Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior, Cambridge, MA*, pages 516–525.