Learning General Language Processing Agents

Dani Yogatama
dyogatama@gmail.com

DeepMind, London, United Kingdom

Table of Contents

1. Executive Summary

The broad objective of the proposed research is to create a general language processing agent (model) that is capable of performing multiple language-related tasks (e.g., question answering, translation, summarization) in multiple languages (e.g., English, Chinese, Malay, Tamil) by learning from data of different modalities (e.g., text, images, structured knowledge bases).

State-of-the-art machine learning models work well when optimized for a particular task, but they require many in-domain training examples (i.e., input-output pairs that are often costly to annotate). In order to do well on multiple tasks, we need a separate model for each task which is trained on specific training examples for the task. As a result, being able to do multiple tasks requires maintaining multiple specialized models.

A key reason for this limitation is the inability of existing machine learning models to learn knowledge (i.e., concepts) that is transferable to new tasks outside their training distribution (Yogatama et al., 2019). The central hypothesis of this research is that obtaining such an ability requires significant advances in how we acquire, represent, and store knowledge in artificial systems. In particular, we will focus on developments of novel training methods and model architectures to learn a general language processing agent.

Our training paradigms will place an emphasis on learning from multiple sources (e.g., multiple languages, images, knowledge bases) to allow a model to learn grounded task-agnostic concepts. On the architecture side, building on our latest work on semiparametric language models (Yogatama et al., 2021), we will design a new generation of deep learning models that combines a large neural network with a key-value database to effectively store and reuse previously acquired knowledge.

The proposed research is a scientific endeavor to replicate natural intelligence in artificial systems with many immediate applications (e.g., question answering systems, machine translation services). Success in this research will strengthen Singapore's position as both a scientific leader in artificial intelligence and a pioneer in using computer science to solve real-world problems.

2. Aims and Objectives

Our goal is to make progress toward a general language processing agent (i.e., a language model) that can perform multiple tasks in multiple languages by learning efficiently from data of multiple modalities. We use a broad definition of multiple tasks, which includes: (i) generating text from many domains (e.g., news articles, social media, scientific text), (ii) generating summaries of news articles; and (iii) answering factual questions (e.g., Who is the current prime minister of Singapore?). While this is a long-term endeavor, the intended end results for this specific proposal are:

- To advance the state-of-the-art on language models that learn from data of multiple modalities (multilingual text, images, knowledge bases), as measured by their performance on benchmark question answering datasets (Goyal et al., 2018; Kwiatkoski et al., 2019; Boratko et al., 2020).
- To create the first large-scale language model that continually gets better over time as it sees more data from new tasks, without having to be retrained on older tasks (contrast this with existing models that always need to be retrained on older tasks as

they see new tasks). As a concrete measure of this second objective, our goal is to have a language model that achieves comparable performance to state-of-the-art models on all (previously seen) tasks with significant reduction in training time.

3. Reasons for Wanting to do Research in Singapore

My motivation to do this research in Singapore is twofold. As a researcher, I am passionate about using artificial intelligence (AI) to improve people's lives. Singapore is a multicultural country with world-class research universities and many industry research laboratories. This creates a unique opportunity to shorten the amount of time needed to transfer fundamental scientific advances to real-world applications.

In particular, I plan to collaborate with research groups at the National University of Singapore (e.g., Professor Min Yen Kan, Professor Wee Sun Lee, Professor Harold Soh, and others). I have had initial discussions with them and they are supportive of my plan to conduct this research at NUS. I believe that this research program will complement existing AI research at NUS and provide a collaboration hub for several research groups working on related topics. Access to top-tier students at a school such as NUS is critical to the success of my research. Industry laboratories such as Salesforce Research—whose Asia's branch is headquartered in Singapore, Baidu Research Singapore, and Grab AI are potential partners for future collaborations. My experience working at Baidu Silicon Valley Research Lab (Sunnyvale, California) provides a starting point for this potential collaboration. I believe that my research will also benefit other technological companies in Singapore who are interested in incorporating automation and artificial intelligence to their workflows.

Second, I am enthusiastic about increasing ethnic, gender, and cultural diversity in AI. I believe that progress in artificial intelligence should benefit everyone in the world. At 640 million people, Southeast Asians constitute nearly 10% of the world's population, but we remain extremely underrepresented in artificial intelligence. Data from ICML 2018, a premier machine learning conference, suggests that fewer than 0.5% of ICML attendees are from Southeast Asian institutions (all of them are from Singapore). I am eager to train next-generation machine learning researchers and engineers in Southeast Asia. I have taken concrete steps toward this goal. For example, I have been involved as a main organizer of the Southeast Asia Machine Learning School (SeaMLS), which is an annual five-day event where participants (students, industry practitioners) attended lectures given by world-renowned experts. Beyond Singapore, I have strong connections with many technological companies in the region (e.g., Indonesia, Vietnam) that sponsor SeaMLS. My goal is to build a regional machine learning community in Southeast Asia that could actively contribute to the global AI ecosystem. I believe that conducting my research program in Singapore, will strengthen Singapore's position as the AI center of Southeast Asia and accelerate the creation of a thriving regional collaboration network.

4. Background and Significance

4.1 Background

Advances in representation learning have considerably improved natural language processing models on many tasks. These models (Peters et al., 2018; Devlin et al., 2019;

Yang et al., 2019) are first pretrained on a large corpus in an unsupervised fashion (without any labeled training data). The models are then used for various downstream tasks by retraining (i.e., fine tuning) on labeled (supervised) training examples for the task of interest. The pretraining stage (i.e., representation learning) has been shown to significantly reduce the number of supervised training examples that are needed to adapt a model to perform a downstream task. Since it only relies on unannotated corpora, we can leverage the vast amount of data that is available on the web.

There are three imitations of state-of-the-art models that we seek to address. First, fine tuning requires a large number of supervised training examples (input-output pairs), which are often costly to annotate. Second, the resulting downstream model only works well on a single task it is fine-tuned on. Performing well on many tasks requires maintaining multiple variants of the models, fine-tuned separately for each task. Recent work attempts to design a single model to perform many language generation tasks (McCann et al., 2018; Lewis et al., 2019; Raffel et al., 2020, Brown et al., 2020). However, the performance of such a model is still below its task-specific counterparts. Third, existing model architectures based on transformer (Vaswani et al., 2017) do not work well for out-of-distribution examples (e.g., when processing long articles or presented with new tasks) due to its inability to remember long-term context. Many variants of memory-augmented neural networks have been developed to try to address this problem (Grave et al., 2017, Dai et al., 2019, Kitaev et al., 2020, Khandelwal et al., 2020, *inter alia*). However, each of them has its own drawbacks and the best method to incorporate long-term context still an open research question.

Our approach to overcome these limitations is through innovations in model architectures and training paradigms. In particular, we seek to use advances in multilingual (Conneau and Lample, 2019; Huang et al., 2019; Cao et al., 2020) and multimodal learning (Lu et al., 2019; Tan and Bansal, 2019; Chen et al., 2020; Li et al., 2020) to learn transferable representations and to combine neural networks with key-value databases to remember long-term context. We provide a more detailed description of our proposal in Section 5.2.

This research builds on an ongoing research program that I lead at DeepMind. I joined DeepMind in March 2016 and I have been leading a team to work on general language models since the summer of 2018. My PhD work (2010–2015) on efficient (sparse) language models greatly influences my thinking on this subject.

At DeepMind, my current group consists of two research scientists, one research engineer, and one research intern. My past and present group members include PhD graduates from Carnegie Mellon University, University of Oxford, University of Cambridge, and PhD interns from University of Washington, Carnegie Mellon University, University of Oxford, University of the Basque Country. I regularly publish at top-venues in machine learning and natural language processing. More recently, my group has published conference papers and journal articles on general language models at leading venues (Yogatama et al.; 2017; Yogatama et al.; 2018; Kong et al., 2019; de Masson d'Autume et al., 2019, Kong et al., 2020; Artetxe et al., 2020; Peng et al., 2021; Yogatama et al; 2021), which demonstrate my ability to lead a team to make concrete progress toward this goal. We also have a track record of releasing open-source resources. For example, we released a new multilingual question answering dataset (XQuAD; https://github.com/deepmind/xquad) last year.

DeepMind is a world leader in artificial intelligence research. I strongly believe that my experience leading a team at such a place provides a solid foundation to lead a similar research program in Singapore.

4.2 Significance

The ability to continuously learn and generalize to new problems quickly is a hallmark of general intelligence. Humans are able to accumulate task-agnostic knowledge from multiple modalities to facilitate faster learning of new skills. The goal of the proposed research—which is to build an artificial agent with human-level linguistic ability—is a scientific endeavor to better understand natural intelligence and replicate it in artificial systems with many immediate applications. For example, real-world applications that will benefit from my research include multilingual question answering systems and machine translation services. As many public services move to autonomous systems, the ability to provide information in multiple languages (e.g., the four official languages of Singapore) becomes more and more important.

There are two main hypotheses that drive my research program. First, learning a multilingual model from multiple modalities allows discovery of higher-level concepts, which will help a model to generalize to new tasks more efficiently. In contrast to existing research that either focuses on text-only models or take a rudimentary approach to combine multimodal information (i.e., by input concatenation), we will focus on innovative architectures and training paradigms (Section 5.2). Success in this research will produce language models that can be easily adapted to any downstream task with very few supervised training examples, allowing widespread deployments for many services.

Second, limitations of existing approaches (Section 4.1) are inherently caused by the way we train our models as parametric neural networks. In other words, all of the learned knowledge is encoded in the parameters of a large neural network, making it difficult to interpret and sensitive to changes in environments (i.e., distributions from which examples come from). We take a unique semiparametric approach, where we aim to combine a large neural network—which is responsible for processing an input to produce an output—with a key-value database responsible for storing long-term knowledge. We consider such a modular design to be a necessary structural bias and a more viable alternative to training an ever-larger neural network. Success in this research will result in more efficient (i.e., a smaller number of parameters is needed to achieve comparable performance, which translates to reduction in training and prediction time), more interpretable, and overall better language models.

In addition, this research will train students and postdoctoral researchers who will continue to benefit Singapore as they will be experts in artificial intelligence, which is shaping up to be one of the most important technologies of the next few decades.

5. Research Design and Methods

5.1 Research Design

This project involves myself as the principal investigator, three PhD student researchers, and one postdoctoral researcher at a given time. The proposal would support the three graduate student researchers for five years, one postdoctoral researcher in Years 2 and 3, and another postdoctoral researcher in Years 4 and 5. In particular:

- Student 1 will focus on learning from multimodal data.
- Student 2 will focus on learning multilingual models.
- Student 3 will focus on model architectures for memory networks.
- The first postdoctoral researcher (Years 2 and 3) will create a prototype for a multilingual multimodal model. Their responsibilities will include co-advising Students 1 and 2 to ensure that the progress made can be consolidated to a single model.
- The second postdoctoral researcher (Years 4 and 5) will create a prototype of our general language processing agents. They have the option to co-advise two students above (Students 1 or 2; Students 3).

We estimate 40% of the grant to be allocated to support researchers. Throughout the duration of the project, there will be regular weekly group meetings to discuss overall progress and brainstorm ideas as a group, as well as individual weekly meetings between each researcher and myself.

The funds from the research grant will also be used to buy equipment necessary to carry out this research. The main requirement is a compute cluster that is equipped with latest-generation graphical processing units. NVIDIA DGX A100 is the most suitable infrastructure for training large scale deep learning models. The cost to obtain it is approximately 199,000 USD. We plan to obtain two DGX A100 at Year 1 to make sure that exploratory research can be performed side by side with training more-matured large-scale models. We expect this resource to still be useful beyond the duration of the grant. In addition to the two DGX A100s, we also plan to use on-demand cloud services such as Google Compute Engine (GCE) and Amazon Web Services (AWS) when there is a periodical increase in our compute demands (e.g., near conference deadlines). Overall, I estimate 40% of the grant to be allocated for equipment purchase (including personal workstations for our researchers) and monthly or hourly rental (i.e., GCE and AWS).

The remaining 20% of the grant will be used for travel costs for presenting the work and as a backup for other miscellaneous expenses (e.g., paying crowd annotators to create a new dataset if a new dataset is needed to evaluate our models).

5.2 Research Methods

A perfect generative language model—in theory—should be able to do any language-related task (e.g., by formulating the task as a question and querying the language model to generate answers). Our research will focus on improving language models in several directions. We divide this section into three parts, each outlining a concrete innovation to drive this research forward.

5.2.1 Multimodal Learning

This research builds on our work on representation learning and semiparametric language models (Kong et al., 2020; Yogatama et al., 2021). In this new direction, we aim to make advances on a language processing agent that learns from multimodal data. We plan to use two sources of multimodal data: image databases and knowledge bases.

Existing multimodal (vision and language) models rely on paired image-text datasets (e.g., image captioning datasets) to pretrain on. This limits the availability of pretrained datasets to restricted domains where such paired data is available. We will design a model that can learn from any unaligned pair of corpus (e.g., Wikipedia articles) and image database (e.g., ImageNet). For a given sequence of text from a Wikipedia article, the model will retrieve a relevant set of images from the database. This retrieval mechanism needs to be trained, since otherwise the model does not know which images are relevant. We will rely on existing image captioning and visual question answering datasets to pretrain the retrieval. It will then be further updated during the language model training, where given a set of retrieved images and a text sequence, the model is trained to predict the next word by attending to textual context and retrieved images. Finally, a gating function will be used to combine visual and textual information. Our approach is different from existing approaches that concatenate visual and textual information (no gating function) and are trained on paired image-text datasets (no retrieval mechanism).

In addition to vision and language models, we will apply the same framework to train a language model that retrieves relevant information from a knowledge base (e.g., Google Knowledge Graph). Initially, this will be done independently of the vision and text model above, by replacing the image database where we retrieve relevant signals from by a structured knowledge base.

We plan to evaluate our multimodal representation learning work on benchmark question-answering datasets, including open-domain question answering (e.g., Natural Questions; Kwiatkowski et al., 2019), common-sense question answering (e.g., ProtoQA; Boratko et al., 2020), and visual question answering (e.g., VQA; Goyal et al., 2017).

5.2.2 Multilingual Learning

This research builds on our work on learning multilingual representations (Artetxe et al., 2020). Our goal is to learn multilingual models that work well across many languages.

Existing multilingual models work much better for high resource languages (e.g., English) compared to low resource languages. There are many corpora that cover many possible distributions (i.e., different domains such as news text, social media text, scientific text, and others) in high resource languages. Since existing models are trained with empirical risk minimization (i.e., minimizing the error of examples sampled from a training distribution), it naturally results in models that perform better on high resource languages. However, this can create a problem when the models are used in practice. A fair multilingual model should perform equally well on any of the languages, instead of overfitting to one language and neglecting the rest.

In order to learn a better model, we will rely on Distributionally Robust Optimization (DRO). DRO is an optimization framework that is used to train a machine learning model on a collection of distributions. In DRO, the objective function that we minimize intuitively upweights examples that come from distributions which incur high losses (errors) and downweights examples from distributions with low losses, resulting in a model that performs uniformly well across all distributions in the collection. We note that even though DRO is a natural fit to train a multilingual model, it has not been used in this context before. Success in this project could lead to a widespread adoption of the framework in the multilingual community.

Our multilingual language models will be evaluated based on their generation capability on benchmark multilingual corpora—e.g., multilingual C4 (Xue et al., 2020). If necessary, we also plan to create other multilingual corpora as a part of this project. For example, we can target a more representative set of languages for Singapore by creating a new multilingual corpus that consists of articles in English, Chinese, Malay, and Tamil.

5.2.3 Memory Networks

Machine learning models work well on a dataset given enough training examples, but they often fail when the data distribution shifts (e.g., when presented with very long context or a new dataset)—a phenomenon known as catastrophic forgetting (McCloskey and Cohen, 1989; Ratcliff, 1990). In language models, this problem manifests in several ways: (i) the inability to work with long documents (more than a few thousand words), (ii) the tendency to hallucinate answers when a question answering model is asked fictional questions, and (iii) performance degradation over time.

This research builds on our effort on learning memory-augmented language models (Yogatama et al., 2018; de Masson d’Autume et al., 2019; Nematzadeh et al., 2020; Peng et al., 2021; Yogatama et al., 2021), where we have shown the benefit of augmenting a language model with a long-term memory in the form of a key-value database.

While the long-term memory database mitigates catastrophic forgetting, several key limitations exist. The size of the database grows linearly with the amount of experience (i.e., the number of training examples) the model acquire. We will explore novel techniques to selectively store past examples, discard unused examples, and dynamically merge similar examples into a single database entry (i.e., learning what to remember and forget). Previous work has argued that some examples are “easy” for a neural network to predict correctly, and it spends most of its training time to learn “hard” examples (Toneva et al., 2019). They also show that a significant number of examples can be omitted from the training set without much degradation in performance. We will use insights from forgettable examples to better structure our long-term memory database. In addition, we will continue innovating on the architecture side, particularly on how to combine information retrieved from the long-term memory component with the current input. Our goal is to demonstrate comparable performance of a constant-size memory model to a linear-size memory model on existing tasks where long-term memory has been shown to be beneficial such as book-level language modeling and continual question answering. We will continue taking inspirations from neuroscience and cognitive science to make progress in this area.

6. Milestones and Deliverables

6.1 Milestones

Year 1

- **Multimodal learning.** Initial work on retrieval-augmented vision and language model, focusing on learning retrieval of relevant images for a given sequence of text.
- **Multilingual learning.** Initial work on DRO for multilingual models using up to ten languages to show that DRO is a feasible framework for training multilingual models.
- **Memory networks.** Set up the necessary infrastructure to combine large neural networks with a key-value database, including a replication of our work on adaptive semiparametric language models (Yogatama et al., 2021) as open-source software.

Year 2

- **Multimodal learning.** Integration of the retrieval model developed in Year 1 to a representation learning model. Evaluation on question answering datasets. Given a positive outcome, submit a paper to a top-tier conference.
- **Multilingual learning.** Scaling up DRO training to a hundred languages. Given a positive outcome, submit a paper to a top-tier conference.
- **Memory networks.** Using the infrastructure set up in Year 1, focus on research on constant-size memory architectures.

Year 3

- **Multimodal learning.** Research on incorporating structured knowledge bases to our representation learning model. The focus is on ensuring that relevant information from the knowledge bases can be retrieved for a given sequence of text.
- **Multilingual multimodal learning.** Initial work on incorporating visual information to ground multilingual models to language-independent concepts.
- **Memory networks.** Continue research on architectures, focusing on combining long-term memory with current inputs. Evaluation on language modeling and question answering. Given a positive outcome, submit a paper to a top-tier conference.

Year 4

- **Multimodal learning.** Continue research on incorporating knowledge bases to a representation learning model. The focus in the second year of this project is on designing architectures to make the best use of the retrieved information from the knowledge bases. Given a positive outcome, submit a paper to a top-tier conference.
- **General language processing agents.** Initial work on a language model that combines images, knowledge bases, and multilingual textual information. The focus is on setting up the infrastructure to incorporate progress made during the duration of the project into a single model.

Year 5

- **General language processing agents.** The culmination of our work in Years 1, 2, 3, and 4. Research on general language processing agents, open source the resulting models, and submit a journal article about the work.

6.2 Deliverables

The primary product of this work will be new publications appearing in Years 2, 3, 4, and 5 satisfying Objectives 1 and 2 in Section 2. We target top-tier journals and conferences in natural language processing and machine learning such as Journal of Machine Learning Research, Transactions of the Association for Computational Linguistics, Conference on Neural Information Processing Systems, Conference of the Association for Computational Linguistics, and other similar venues. When appropriate, we will release our code and model parameters as open-source software.

I regularly give high-profile invited talks about my group's work. The last three talks are at New York University (November 2020), University of Cambridge (November 2020), and ICML 2020 Retrospective Workshop (July 2020). I expect this to continue and intend to frequently communicate results from this research to the wider scientific community.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the Cross-Lingual Transferability of Monolingual Representations. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2020.

Michael Boratko, Xiang Lorraine Li, Rajarshi Das, Tim O'Gorman, Dan Le, and Andrew McCallum. ProtoQA: A Question Answering Dataset for Prototypical Commonsense Reasoning. In Proceedings of *Conference on Empirical Methods in Natural Language Processing*. 2020.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dahriwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Grethchen Krueger, Tom Hennighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language Models are Few-Shot Learners. In *Proceedings of the Conference on Neural Information Processing Systems*. 2020.

Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual Alignment of Contextual Word Representations. In *Proceedings of the International Conference on Learning Representations*. 2020.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. In Proceedings of the European Conference on Computer Vision. 2020.

Alexis Conneau and Guillaume Lample. Cross-lingual Language Model Pretraining. In *Proceedings of the Conference on Neural Information Processing Systems*. 2019.

Cyprien de Masson d'Autume, Sebastian Ruder, Lingpeng Kong, and Dani Yogatama. Episodic Memory in Lifelong Language Learning. In *Proceedings of the Conference on Neural Information Processing Systems*. 2019.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models Beyond a Fixed-length Context. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2019.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2019.

Yash Goyal, Tejas Khot, Douglas Summer-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. 2017.

Edouard Grave, Armand Joulin, and Nicolas Usunier. Improving Neural Language Models with a Continuous Cache. In *Proceedings of the International Conference on Learning Representations*. 2017.

Haoyang Huang, Yaobo Liang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, and Ming Zhou. 2019. Unicoder: A Universal Language Encoder by Pretraining with Multiple Cross-lingual Tasks. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. 2019.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through Memorization: Nearest Neighbor Language Models. In *Proceedings of the International Conference on Learning Representations*. 2020.

Nikita Kitaev, Lukasz Kaiser, and Anselm Kevskaya. Reformer: The Efficient Transformer. In *Proceedings of the International Conference on Learning Representations*. 2020.

Lingpeng Kong, Cyprien de Masson d'Autume, Wang Ling, Lei Yu, Zihang Dai, and Dani Yogatama. A Mutual Information Maximization Perspective of Language Representation Learning. In *Proceedings of the International Conference on Learning Representations*. 2020.

Lingpeng Kong, Gabor Melis, Wang Ling, Lei Yu, and Dani Yogatama. Variational Smoothing in Recurrent Neural Network Language Models. In *Proceedings of the International Conference on Learning Representations*. 2019.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*. 2019.

Mike Lewis, Yinhan Liu, Naman Goyal, Marian Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, Luke Zettlemoyer. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 2020.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and Performant Baseline for Vision and Language. *arXiv:1908.03557.* 2019

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision and Language Tasks. In *Proceedings of the Conference on Neural Information Processing Systems*. 2019.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The Natural Language Decathlon: Multitask Learning as Question Answering. *arXiv:1806.08730*. 2018.

Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*. 1989.

Aida Nematzadeh, Sebastian Ruder, and Dani Yogatama. On Memory in Human and Artificial Language Processing Systems. In *Proceedings of ICLR Workshop on Bridging AI and Cognitive Science*. 2020.

Hao Peng, Nikolaos Pappas, Dani Yogatama, Roy Schwartz, Noah A. Smith, and Lingpeng Kong. Random Feature Attention. In *Proceedings of the International Conference on Learning Representations*. 2021.

Colin Raffel, Noam Shazeer, Adam Robert, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J. Liu. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*. 2020.

Roger Ratcliff. Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions. *Psychological Review*. 1990.

Hao Tan and Mohit Bansal. LXMERT: Learning Cross-modality Encoder Representations from Transformers. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*. 2019.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, Geoffrey J. Gordon. An Empirical Study of Example Forgetting during Deep Neural Network Learning. In *Proceedings of the International Conference on Learning Representations*. 2019.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep Contextualized Word Representations. n *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the Conference on Neural Information Processing Systems*. 2017.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffle. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. *arXiv 2010.11934. 2020*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc Le. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In *Proceedings of the Conference on Neural Information Processing Systems*. 2019.

Dani Yogatama, Phil Blunsom, Chris Dyer, Edward Grefenstette, and Wang Ling. Learning to Compose Words into Sentences with Reinforcement Learning. In *Proceedings of the International Conference on Learning Representations*. 2017.

Dani Yogatama, Yishu Miao, Gabor Melis, Wang Ling, Adhiguna Kuncoro, Chris Dyer, and Phil Blunsom. Memory Architectures in Recurrent Neural Network Language Models. In *Proceedings of the International Conference on Learning Representations*. 2018.

Dani Yogatama, Cyprien de Masson d'Autume, and Lingpeng Kong. Adaptive Semiparametric Language Models. *Transactions of the Association for Computational Linguistics*. 2021.

Dani Yogatama, Cyprien de Masson d'Autume, Jerome Connor, Tomas Kocisky, Mike Chrzanowski, Lingpeng Kong, Angeliki Lazaridou, Wang Ling, Lei Yu, Chris Dyer, and Phil Blunsom. Learning and Evaluating General Linguistic Intelligence. *arXiv 1901.11373*. 2019.