Less is more: An Empirical Analysis of Model Compression for Dialogue

Anonymous Authors[1]

Abstract

Large language models have achieved near-human performance across wide Natural Language Generation tasks such as Question Answering and Open-Domain Conversation. These large models take up large memory footprints and also inference time. Compressed models with fewer parameters are easily deployable on FPGAs and low-end devices with limited storage memory and processing power. In this work, we carry out an empirical evaluation of three model compression techniques on conversational agents specifically pre-trained on large language transformer networks. Using OpenAI GPT-2 transformer network, we evaluate and compare the performance of open-domain dialogue models before and after undergoing compression. When trained and tested on the DailiyDialog corpus, compressed models exhibit performances achieving state-of-the-art results on the corpus while maintaining human likeness.

1. Introduction

Conversational systems or chatbots have found their way into our everyday lives due to their wide range of uses from technical support services (Low et al., 2015; Qi et al., 2021) to entertainment (Zhou et al., 2020) and personal assistants (Google, 2019; Amazon, 2019). The study of chatbots constitutes an interesting field in NLP. Task-oriented and Open-Domain chatbots are the two main variants. Often based around knowledge structures, known as "frames" representing intents in input statements, task-oriented or goal-based systems extract information from input statements into predefined slots to guide response; hence their responses are basically tabular. (Jurafsky & Martin, 2019). Open-domain chatbots in contrast, are data driven which rules out limitation to the kind of topic they can be engaged with by an interlocutor.

Conversational AI, like many other areas of NLP has benefited from the representation capabilities of large language models. Recent advances in dialogue modeling have resulted in the development of chatbots with responses that are close to human.
These models with very high number of parameters can not be democratized since they consume very large memory footprints and also exhibit high inference latencies to be deployable in embedded devices with low storage and processing capacities. Finetuning such models for customized applications is usually not only impractical due to computational resource requirements but also results in high carbon emissions that affects global climate change (Amodei & Hernandez, 2018).

In this paper, we take a step towards circumventing these problems in the dialogue domain by utilizing model compression to improve runtime efficiency in chatbots. We also evaluate model performance after compression. A common approach to model compression is knowledge distillation (KD) which entails the active transfer of representation power of a large or an ensemble network to a smaller one. This is achieved by feeding back the outputs of the large network as *soft targets* to the smaller network. This method has shown a lot of promise both in Image recognition (Zhang et al., 2018b) and Language understanding (Liu et al., 2019). Recent advances in model interpretability has proven that large models with several layers and units contain redundant members (Bylinskii et al.; Springenberg et al., 2015). Pruning is a systematic procedure by which sparsity is induced in a large model with numerous interconnected units by optimizing a parameterized mesh of numbers (or mask) with the sole aim of eliminating redundant network connections or units (Gomez et al., 2019). A common alternative form of sparsification is L_0 norm regularization, essentially penalizing model weights of a deep neural network for being other than zero. Representing model weights with lower precision values is yet another common technique of shrinking the storage memory occupied by large models.
Network quantization and weight sharing are two model shrinkage methods that achieve compression by reducing the effective number of bits required to represent each weight. Weight sharing limits memory storage by configuring multiple connections to share the same set of weights while

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

quantization essentially is representing model weights with low-precision values.

The code for our experiments is publicly available. [1]

2. Related Work

Neural response generation has gained a lot of interesting traction in recent years with the advent of large transformer networks (Adiwardana et al., 2020; Roller et al., 2020; Xu et al., 2020). Traditional systems involved building sub-components like a natural language understanding (NLU) component for detecting intent and extracting associated information, a dialogue state tracker that maintains consistency, a response selection policy selects the next action based on the current the chat state and a natural language generator (NLG) that translates actions to responses. (Gao et al., 2018). These modular techniques have been overtaken by large scale end-to-end methods in recent decades.
Knowledge-based (KB) systems are an interesting example; they are essentially semantic parsers that query a large scale structured database of information (Auer et al., 2007; Berant et al., 2013) at every response step. These systems are notoriously computationally expensive due to the search complexity involved in selecting optimal responses from a myriad of candidate logic relation paths. Also paraphrasing the same queries usually throws KB systems off easily. (Gao et al., 2018)

Early end-to-end approaches towards dialogue modeling applied generative recurrent methods like seq2seq (Vinyals & Le, 2015; Serban et al., 2016; Sordoni et al., 2015). Attempts have also been made to formalize dialogue as hierarchical partially observable Markov decision processes (POMDPs)(Young et al., 2013; Fang et al., 2018; Zhou et al., 2020).

It has been observed that practical behaviours of chit-chat agents have certain shortcomings that make them easily fail the Turing test. They often exhibit the bad quality of producing bland responses like "I don't know". (Li et al., 2016) posited that this is due to the conditional objective being asymmetrical with dialog history and response parameters. They suggested the formulating the generative objective in terms of maximum mutual information (MMI) as it solves this problem in theory. Incoherency is a common problem associated with social chatbots as they lose track of the state of the conversation after few dialogue turns. (Li & Jurafsky, 2016) proposed incorporating a predefined profile into the modeling process such that response do not stray from the chatbot's persona. Recent works (Zhang et al., 2018a; Liu et al., 2020) have shown that this improves consistency in generated responses at different stages of dialogue. (Li et al., 2020; Welleck et al., 2019) also tackled inconsistency by using unlikelihood training as a contrastive learning technique to prevent the model from generating out-of-distribution responses. (See et al., 2019) showed that the high level attribute of a conversation can be controlled through conditional training (Fan et al., 2018; Kikuchi et al., 2016; Peng et al., 2018) and weighted decoding (Ghazvininejad et al., 2017).

Research in model compression (LeCun et al., 1989; Hinton et al., 2015; Hooker et al., 2019) have demonstrated that it is possible to achieve unbelievably high level of compression with very minimal degradation in the representation capacity of deep neural networks. By inducing sparsity in ResNet-50 (He et al., 2016) and the vanilla transformer (Vaswani et al., 2017) networks, (Gale et al., 2019) were able to achieve up to 50% compression ratio in the Image-Net(Deng et al., 2009) dataset and WMT '14 (Bojar et al., 2014) machine translation task while maintaining a good performance at the same time.
(Louizos et al., 2018) demonstrated that a significant speedup in training time can be achieved by incorporating L_0 norm regularization in LeNet (Lecun et al., 1998) and wide ResNet (Zagoruyko & Komodakis, 2016) networks with minimal or no loss in test performance.
(Jacob et al., 2018) achieved remarkable trade-offs between model size and performance on four image tasks by representing weights with low-precision values.
In this work, knowledge distillation, pruning and L_0 norm regularization are our major focus of model compression.

3. Method

3.1. Knowledge Distillation

Knowledge Distillation (KD) aims at reproducing the holistic representation abilities of large models with high inference times into smaller models through active learning. The smaller model, often called "Student" tries to emulate the decisions of the larger model or "Teacher".
The distillation objective: $L_{ce} = \sum_j p_j * log(q_j)$;
p_j and q_j being the *Teacher* and *Student* probability distribution estimates respectively, is to match the predictions made by the *Student* with that of the *Teacher*. Much of the generalization abilities of the *Teacher* is associated with the high entropy in the class probabilities it produces during inference.
A slightly modified version of the softmax output is usually fed into the Student model as soft-targets. (Hinton et al., 2015), used a slightly modified version of the Softmax output of the large model. The softmax temperature *T* is an essential parameter the student optimizes during learning, training is usually done by adjusting the softmax tempera-

[1] `web@link.com`

Table 1. Conversational statistics of the DailyDialog Corpus

DailyDialog Corpus	
Total # conversations	13,118
Average # turns per dialogue	8
Average # tokens per dialogue	115
Maximum # turns per dialogue	14
Minimum # turns per dialogue	5

ture T till the *Student*'s predictions matches the soft targets.

$$p_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}$$

Distilling a generative dialogue model optimizes an overall objective which is the linear combination of the distillation loss L_{ce} and language modeling loss L_{lm}.
Following (Sanh et al., 2019), we added an additional objective term *cosine embedding loss (L_{cos})* that aligns the direction of the student and teacher hidden states vectors during distillation.

3.2. Pruning

By evaluating and eliminating redundant connections in a deep neural network, an optimal sub-network containing fewer number of parameters can be obtained (Frankle & Carbin, 2019). Given a large neural network with weights W_{ij} in an n dimensional euclidean space, a pruning strategy is used to determine importance scores $A_{i,j}$, whose values are then sorted according to the pruning strategy. Sparsity is then achieved by masking out low ranked weights that are less than a threshold ϵ value that reflects the level of sparsity.
In this work, we implemented automated gradual pruning by increase sparsity following a cubic sparsity schedule (Zhu & Gupta, 2018; Sanh et al., 2020), from an initial sparsity value ϕ_i to a final level ϕ_f over a span of n pruning steps starting at training step t_0 and with pruning frequency Δt:

$$\phi_t = \phi_f + (\phi_i - \phi_f)\left(1 - \left(\frac{t - t_0}{n\Delta t}\right)\right)$$

$$\text{for } \ t \in \{t_0, t_0 + \Delta t, \ldots, t_0 + n\Delta t\}$$

4. Experimentation and Results

In this section, we discuss our experiments of knowledge distillation, unstructured pruning, l_0 regularization on a GPT-2 based dialogue model. We evaluate model performance with automatic and human metrics.

4.1. Dataset

Experiments were conducted on the DailyDialog (Yanran et al., 2017) corpus. It consists of 13K dyadic multi-turn conversations. Unlike many open-domain conversation corpora, DailyDialog is made up of a high-quality (clean) dialogue dataset that reflects a wide blend of emotions and intentions which are often present in daily human conversations. Table 1 shows basic statistical information in the conversations contained in the corpus. We normalize the currency symbols in conversations to text and truncated sequences longer than 128 tokens.

4.2. Input Representation

We represent each conversation as a sequence of turns for conditional generative training. We delimit each turn by separator tokens (specifically "<|endoftext|>" in GPT-2). We observe that using arbitrary delimiters such as <EOT> produced bad results, only made the models memorize and generate lots of them during inference.
A typical conversation with a total of n turns is re-framed as:

$$Turn_k\text{<SEP>}Turn_{k+1}\text{<SEP>}\ldots Turn_{k+n}\text{<SEP>}$$

where k is the turn index.

4.3. Training Details

We carry out experiments on two Nvidia Tesla K80 and one P100 machines. Basing dilaogue models on pretrained transformer architectures often used as a good starting point (Rashkin et al., 2019; Wolf et al., 2019). In this work, we leverage DialoGPT pretrained model architecture for conditional training. Finetuning was done over 26000 steps on the DailyDialog dataset. We used Adam with learning rate of 5e-5, β1= 0.9, β2 = 0.999 and noam learning rate schedule. We used beam search decoding with a size of 4 and sampling over candidate response units or symbols.

Table 2. Number of non-zero parameters and Inference times for different Compression methods. Combining L_0 regularization effectively influenced the convergence speed

Method	**Params**	**Latency (s)**
Base	345M	3.61
KD	345M	6.34
Pruning (70%)	247M	4.18
Pruning (50%)	181M	2.09
Pruning (30%)	100M	1.73
Reg. + Pruning (50%)	179M	1.42

Table 3. Performance scores of models with different compression schemes before and after compression.

Method	F1	Bleu		Rouge	Meteor
		Bleu4	Bleu2		
Base	7.13	3.01	6.82	18.35	2.41
KD	8.33	0.6	6.01	11.03	0.65
Pruning (70%)	4.19	1.4	4.21	9.18	0.74
Pruning (50%)	3.22	0.5	3.11	5.13	0.65
Pruning (30%)	2.06	0.9	2.05	3.7	1.01
Reg. + Pruning (50%)	2.69	1.79	1.13	2.69	1.01

Table 4. Performance scores on a four-point likert scale. Small repetition scores reflect how less repetitive a model's responses are and vice versa

Method	Engagingness	Specificity	Fluency	Repetition	Empathy
Base	2.91	3.76	3.71	0.41	1.85
KD	1.85	1.22	3.26	2.41	0.31
Pruning (70%)	2.73	3.53	3.11	1.23	1.79
Pruning (50%)	1.72	1.48	3.33	1.17	1.17
Pruning(30%)	1.23	1.19	3.01	1.62	1.19
Reg. + Pruning (50%)	3.63	2.98	1.12	1.34	1.26

4.4. Results

We report performance based on Automatic metrics and also human metrics using a four-point Likert scale (Venkatesh et al., 2017). Results are summarized in table 3 and 4. We used huggingface transformers datasets package (Lhoest et al., 2021) for automatic evaluation on Bleu (Papineni et al., 2002), Rouge (Lin, 2004) and Meteor scores (Banerjee & Lavie, 2005).

5. Conclusion and Future Work

Chatbots trained on large model architectures can be an interesting feature in embedded devices by carefully compressing them into computationally efficient sub-models. Our paper focuses mainly on memory optimization of chatbots and we did not consider advanced dialogue modeling techniques such as response retrieval, re-ranking beam search outputs much of which have been proven to increase diversity and the human-likeness of responses across many benchmarks. Model compression was not carried out only on regular causal language models, we posit that better bots trained on sophisticated models like Electra, evolved transformer can be achieved. The lowest compression ratio our methods achieved was 30%; we would like to consider exploring high sparsity regimes by augmenting pruning with integer quantization.

References

Adiwardana, D., Luong, M., So, D. R., Hall, J., Fiedel, N., Thoppilan, R., Yang, Z., Kulshreshtha, A., Nemade, G., Lu, Y., and Le, Q. V. Towards a human-like open-domain chatbot. *arXiv e-prints*, abs/2001.09977, 2020.

Amazon. Alexa skills. 2019. URL https://developer.amazon.com/alexa-skills-kit/tutorials.

Amodei, D. and Hernandez, D. AI and Compute. *OpenAI blog*, 2018. URL https://blog.openai.com/openai-five/.

Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. Dbpedia: A nucleus for a web of open data. In *Semantic web*, pp. 722—-735. Springer, 2007.

Banerjee, S. and Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pp. 65–72, Ann Arbor, Michigan, 2005. Association for Computational Linguistics.

Berant, J., Chou, A., Frostig, R., and Liang, P. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods* *in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, oct 2013. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D13-1160.

Bojar, O., Buck, C., Federmann, C., Haddow, B., Koehn, P., Leveling, J., Monz, C., Pecina, P., Post, M., Saint-Amand, H., Soricut, R., Specia, L., and Tamchyna, A. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/W14-3302. URL https://www.aclweb.org/anthology/W14-3302.

Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., Oliva, A., and Torralba, A. Mit Saliency Benchmark. http://saliency.mit.edu/.

Deng, J., Dong, W., Socher, R., Li-Jia, L., Li, K., and Fei-Fei, L. Imagenet: A Large-scale Hierarchical Image Database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Fan, A., Grangier, D., and Auli, M. Controllable Abstractive Summarization. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pp. 45–54, Melbourne, Australia, jul 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-2706. URL https://www.aclweb.org/anthology/W18-2706.

Fang, H., Cheng, H., Sap, M., Clark, E., Holtzman, A., Choi, Y., Smith, N. A., and Ostendorf, M. Sounding board – a user-centric and content-driven social chatbot. In *Proceedings of North American Chapter Association for Computational Linguistics (NAACL)*, pp. 1–6. NACL, June 2018.

Frankle, J. and Carbin, M. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. In *7th International Conference on Learning Representations, ICLR*, LA, USA, 2019.

Gale, T., Elsen, E., and Hooker, S. The state of sparsity in deep neural networks. *arXiv e-prints*, abs/1902.09574, 2019. URL https://arxiv.org/abs/1902.09574.

Gao, J., Galley, M., and Li, L. Neural approaches to conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 2–7, Melbourne, Australia, jul 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-5002. URL https://www.aclweb.org/anthology/P18-5002.

Ghazvininejad, M., Shi, X., Priyadarshi, J., and Kevin, K. Hafez: an interactive poetry generation system. In *Proceedings of ACL 2017, System Demonstrations*, pp. 43–48, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P17-4008.

Gomez, A., Zhang, I., Kamalakara, S., Madaan, D., Swersky, K., Gal, Y., and Hinton, G. Learning sparse networks using targeted dropout. 2019.

Google. Actions on google. 2019. URL https://developers.google.com/actions/overview.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 770–778. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.90. URL https://doi.org/10.1109/CVPR.2016.90.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv e-prints*, arXiv:1503.02531, 2015.

Hooker, S., Courville, A., Dauphin, Y., and Frome, A. Selective brain damage: Measuring the disparate impact of model pruning. *arXiv*, abs/1911.05248, 2019. URL https://arxiv.org/abs/1911.05248.

Jacob, B., Kligys, S., Chen, B., Zhu, M., Tang, M., Howard, A., Adam, H., and Kalenichenko, D. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Jurafsky, D. and Martin, J. H. *Speech and Language Processing*. 3rd edition, 2019. URL https://web.stanford.edu/jurafsky/slp3.

Kikuchi, Y., Neubig, G., Sasano, R., Takamura, H., and Okumura, M. Controlling output length in neural encoder-decoders. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1328–1338, Austin, Texas, nov 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1140. URL https://www.aclweb.org/anthology/D16-1140.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

LeCun, Y., Denker, J. S., and Solla, S. A. Optimal brain damage. In *Proceedings of the Neural Information Processing Systems*, pp. 98—605. Morgan , 1989.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. Gradient-based learning applied to document recognition. volume 86, pp. 2278–2324, 1998. doi: 10.1109/5.726791. URL https://ieeexplore.ieee.org/document/726791.

Lhoest, Q., von Platen, P., Wolf, T., del Moral, A. V., Jernite, Y., Patil, S., Drame, M., Chaumond, J., Plu, J., Tunstall, L., Davison, J., Brandeis, S., Scao, T. L., Sanh, V., Xu, K. C., Patry, N., McMillan-Major, A., Schmid, P., Gugger, S., Delangue, C., Matussière, T., Debut, L., Bekman, S., and Lagunas, F. huggingface/datasets, 2021. URL https://doi.org/10.5281/zenodo.4946100.

Li, J. and Jurafsky, D. Neural net models for open domain discourse coherence. *arXiv e-prints*, arXiv:1606.01545, 2016.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. A diversity-promoting objective function for neural conversation models. *arXiv e-prints*, arXiv:1510.03055, 2016.

Li, M., Roller, S., Kulikov, I., Welleck, S., Boureau, Y., Cho, K., and Weston, J. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4715–4728, Online, jul 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.428. URL https://www.aclweb.org/anthology/2020.acl-main.428.

Lin, C. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, 2004. Association for Computational Linguistics.

Liu, Q., Chen, Y., Chen, B., Lou, J., Chen, Z., Zhou, B., and Zhan, D. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 2020.

Liu, X., He, P., Chen, W., and Gao, J. Multi-Task Deep Neural Networks for Natural Language Understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4487–4496, Florence, Italy, 2019. Association for Computational Linguistics.

Louizos, C., Welling, M., and Kingma, D. P. Learning sparse neural networks through L_0 regularization. In *International Conference on Learning Representations*, 2018.

Low, R., Pow, N., Serban, I. V., and Pineau, J. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. ACL*, pp. 285–294, 2015.

Papineni, K., Roukos, S., Ward, T., and Zhu, W. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, USA, 2002. Association for Computational Linguistics.

Peng, N., Ghazvininejad, M., May, J., and Knight, K. Towards controllable story generation. In *Proceedings of the First Workshop on Storytelling*, New Orleans, Louisiana, jun 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-1505. URL https://www.aclweb.org/anthology/W18-1505.

Qi, H., Pan, L., Sood, A., Shah, A., Kunc, L., and Potdar, S. Benchmarking intent detection for task-oriented dialog systems. In *North American Chapter of the Association for Computational Linguistics NACL*, 2021.

Rashkin, H., Smith, E. M., Li, M., and Boureau, Y.-L. Towards empathetic open-domain conversation models: a new benchmark and dataset. In *ACL*, 2019.

Roller, S., Dinan, E., Goyal, N., Ju, D., Williamson, M., Liu, Y., Xu, J., Ott, M., Shuster, K., Smith, E. M., Boureau, Y., and Weston, J. Recipes for building an open-domain chatbot. *arXiv e-prints*, arXiv/2004.13637, 2020.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings of Neural Information Processing Systems (NeurIPS 2019)*, VBC, Canada, 2019.

Sanh, V., Wolf, T., and Rush, A. M. Movement pruning: Adaptive sparsity by fine-tuning. In *Proceedings of Neural Information Processing Systems (NeurIPS 2020)*, 2020.

See, A., Roller, S., Kiela, D., and Weston, J. What makes a good conversation? how controllable attributes affect human judgments. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, ACL*, pp. 1702–1723, 2019.

Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., and Bengio, Y. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2016. URL https://ojs.aaai.org/index.php/AAAI/article/view/10983.

Sordoni, A., Galley, M., Auli, M., Brockett, C., Ji, Y., Mitchell, M., Nie, J., Gao, J., and Dolan, B. A neural network approach to context-sensitive generation of conversational responses. *arXiv e-prints*, arXiv:1506.06714, 2015.

Springenberg, J. T., Dosovitskiy, A., Brox, T., and Riedmiller, M. Striving for Simplicity: The All Convolutional Net. In *3rd International Conference on Learning Representations, ICLR*, 2015. URL http://lmb.informatik.uni-freiburg.de/Publications/2015/DB15a.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. Attention is all you need. In *Proceedings of Neural Information Processing Systems (NeurIPS 2019)*, 2017. URL https://arxiv.org/pdf/1706.03762.pdf.

Venkatesh, A., Khatri, C., Ram, A., Guo, F., Gabriel, R., Nagar, A., Prasad, R., Cheng, M., Hedayatnia, B., Metallinou, A., Goel, R., Yang, S., and Raju, A. On Evaluating and Comparing Conversational Agents. *ArXiv*, abs/1801.03625, 2017.

Vinyals, O. and Le, Q. V. A neural conversational model. *arXiv e-prints*, abs/1506.05869, 2015.

Welleck, S., Kulikov, I., Roller, S., Dinan, E., Cho, K., and Weston, J. Neural text generation with unlikelihood training. volume abs/1908.04319, 2019.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv*, abs/1901.08149, 2019.

Xu, J., Ju, D., Li, M., Boureau, Y., Weston, J., and Dinan, E. Recipes for safety in open-domain chatbots. *arXiv e-prints*, arXiv:2010.07079, 2020.

Yanran, L., Hui, S., Xiaoyu, S., Wenjie, L., Ziqiang, C., and Shuzi, N. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995, Taipei, Taiwan, 2017. Asian Federation of Natural Language Processing.

Young, S., Gasic, M., Thomson, B., and William, J. D. Pomdp-based statistical spoken dialog systems: A review. In *Proceedings of the IEEE, 101(5)*, pp. 1160–1179. IEEE, 2013. doi: 10.1109/JPROC.2012.2225812.

Zagoruyko, S. and Komodakis, N. Wide residual networks. *arXiv e-prints*, abs/1605.07146, 2016.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pp. 2204–2213, Melbourne, Australia, 2018a.

Zhang, X., Zhou, X., Lin, M., and Sun, J. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, 2018b.

Zhou, L., Gao, J., Li, D., and Shum, H. The design and implementation of XiaoIce, an empathetic social chatbot. volume 46, pp. 53–93, mar 2020. doi: 10.1162/coli_a_00368. URL https://www.aclweb.org/anthology/2020.cl-1.2.

Zhu, M. and Gupta, S. To prune, or not to prune: Exploring the efficacy of pruning for model compression. In *6th International Conference on Learning Representations, ICLR*, Vancouver, Canada, 2018.