Systematic Generalization In Machine Learning

Usman Anwar

Introduction

Systematic generalization is the most remarkable property of human intelligence. A fascinating example of this generalization is food preparation; from a finite set of a few ingredients, humans create thousands of varied dishes. However, when GPT-3, a large neural network (NN) trained on internet data, including a large number of food recipes, is tasked to innovate and produce a novel food recipe; it completely falters and makes mistakes that sound comical. For example, in one of its recipes, GPT-3 prescribed heating a watermelon and adding egg white to a pot of hot watermelon sugar water to make "scrambled watermelon egg" [Matz, 2021]. On the other hand, when given a familiar list of ingredients that it had seen during training, GPT-3 can reproduce the corresponding recipe effortlessly. This example illustrates the failure of systematic generalization in GPT-3. Unfortunately, it is not just GPT-3 that fails to reason in a systematic way but many popular families of NNs fail in this way [Lake and Baroni, 2018, Bahdanau et al., 2018].

Failings of GPT-3 aside, there is encouraging evidence that by using special kinds of supervisions [Hill et al., 2020, 2019, Akyürek et al., 2020, Stammer et al., 2021] or GOFAI[1] inspired architectures [Russin et al., 2019, Hudson and Manning, 2019, Bergen et al., 2021] or by simply tuning the NN very well [Csordás et al., 2021a], NNs can become significantly better at systematic generalization. However, the broader understanding of when and why NNs generalize systematically remains missing. Within this proposal, I have highlighted a few research directions in this regard that I would like to explore through my doctoral studies. A better understanding of the *ingredients* which bias NNs towards systematic generalization will not only be useful in enabling a sound comparison of existing methods but will also provide insights for developing newer, and potentially better, methods.

What Makes NNs Fail or Succeed at Systematic Generalization?

Traditionally connectionist systems like NNs were thought to be *incapable* of learning systematic properties in the data [Marcus, 1998, Fodor and Pylyshyn, 1988]. This belief has been strongly challenged of late [Hill et al., 2019, Akyürek et al., 2020, Hill et al., 2020, Csordás et al., 2021a]. In particular, the success of [Csordás et al., 2021a] is most surprising and I think deciphering the tricks used therein can provide highly useful insights into when and why systematicity arises within NNs.

1 Why Early Stopping Is At Odds With Systematic Generalization?

Csordás et al. [2021a] have noted that performing early stopping significantly decreases the systematic generalization capabilities of the neural network while **having no effect on iid generalization**.

[1] Good Old Fashioned AI

In line with the popular belief that standard neural networks do not have systematic generalization capabilities, Kim and Linzen [2020] reported negative results about the generalizability of transformers on systematic generalization. However, Csordás et al. [2021a] noted that by simply disabling early stopping in the codebase of [Gulrajani and Lopez-Paz, 2020], drastic gains are achieved and this finding holds across benchmarks. This finding is also supported indirectly by the results given in [Andreas, 2019] which show that while compositionality in the representations of neural networks initially decreases, if training is continued, the trend is eventually reversed and compositionality begins to increase. Understanding this phenomenon better is quite important and I suggest following investigation steps for that.

- In the loss curves reported by [Csordás et al., 2021a] the training loss remains approximately constant for a long time. Previous works have established that different minima found by SGD are connected by a path of non-increasing loss in a phenomenon called mode connectivity [Draxler et al., 2018, Garipov et al., 2018]. **I hypothesise that neural network during further training shifts from one minimum, let's call it minimum A, with poor systematic generalization but good iid generalization to a new minimum, let's call it minimum B, with good systematic generalization as well as iid generalization.** This hypothesis can be verified empirically by studying the out of distribution generalization and system generalization properties of the neural networks at the various points of the loss curve. *If* the hypothesis stands true, then it may be interesting to study how minima A and B differ from each other and what kind of regularization can promote the neural network to reach the minima B and avoid minima A.
- The current neural networks are trained under various regularizations, implicitly in the form of use of SGD based optimizers or data augmentation, or explicitly in form of early stopping and auxiliary objectives (e.g. ℓ_2 regularization). Considering that one of these i.e. early stopping negatively impacts systematic generalization, it is important to empirically study the effect of other regularizations on systematic generalization as well and develop a better understanding of their role in systematic generalization.

2 Modularity as a Prior: Necessary, Sufficient or Just Useful?

Modularity[2] has been shown to be a useful inductive bias for promoting systematic generalization and many GOFAI inspired recent approaches to enhancing systematic generalization of NNs have proposed various explicit modular structures of NNs [Russin et al., 2019, Hudson and Manning, 2019, Bergen et al., 2021]. In fact, on the basis of a an empirical investigation, [Bahdanau et al., 2018] concluded modularity is essential for systematic generalization. However, recent empirical results from [Csordás et al., 2021a] cast doubts on the belief that *modularity is essential for systematic generalization*.

I assert here that it is important to distinguish between explicit modularity and implicit modularity. While results from [Csordás et al., 2021a] provide convincing evidence that explicit modularity is not essential, it does not provide any insights into the role that implicit modularity may have played in their success. **One possible explanation for this can be that the modularity arises implicitly inside NNs that achieve systematic generalization**. Indeed several works support the hypothesis that neural networks can be modular [Csordás et al., 2021b, Filan et al., 2021, Zhang

[2]Modularity here refers to both representational modularity and functional modularity.

et al., 2021] [3]. Further, the proposed explanation can be verified by using tools from [Csordás et al., 2021b] to inspect the NNs from [Csordás et al., 2021a] which achieve systematic generalization. Both the negative and positive results from this experiment will be highly insightful. A negative result showing that NNs from [Csordás et al., 2021a] are not even implicitly modular will provide convincing evidence that modularity is not necessary for systematic generalization. On the other hand, a positive result showing that the neural networks in [Csordás et al., 2021a] are implicitly modular will provide strong evidence that connectionist schemes, trained through gradient descent, are capable of utilizing modularity. In case of positive results, it is important to further investigate the conditions in which implicit modularity emerges. Does not using early stopping help in this? How does overparameterization affect emergent modularity? Is implicit modularity, in some way, a function of neural network architectures? Answering this question will help develop methods to promote emergent modularity in NNs.

3 Data and Supervision: What More is Needed?

In his popular book *Thinking Fast and Slow*, psychologist Daniel Kahneman has distinguished between two forms of thinking; system 1 thinking and system 2 thinking. System 1 thinking is fast, automatic, frequent and unconscious while system 2 thinking is slow, effortful, logical and conscious [Kahneman, 2011]. The terminology has since been appropriated by the machine learning community where system 1 models allude to the current generation of machine learning models which are fast pattern recognition machines and system 2 models is used to refer to hypothetical machine learning models which can reason in more human-like ways and generalize systematically. With system 1 ML models, we have arguably gotten away with very simple supervision signals; for example, GPT-3, despite its humongous size, was only trained on predicting missing tokens from already written human texts. Many computer vision models are initialized with a NN model trained on classification dataset *ImageNet*. However, as we move towards system 2 ML models, where we hunt for models that can scheme systematically and do human-style logical thinking, we will likely have to move beyond simple supervision.

In my opinion, **for system** 2 **models, we will likely have to shift our training signal from *what is right* to *why it is right* or a combination of both**. While explicit explanations are one form of this signal, explaining things by giving negative examples or by giving more examples where the concerned rule is used are also prominent in human-to-human teaching [Nam and McClelland, 2021]. Few of the current works with impressive systematic generalization properties already loosely follow this principle. For example, [Stammer et al., 2021] uses explicit feedback from human to correct its internal representation while [Akyürek et al., 2020, Hill et al., 2019] use intelligent data augmentation techniques to help their models learn the right concept. **Active learning is also likely to be much more useful for system-**2 **models than it has been for system**1 **models in my opinion.** There are several reasons for this. First is that system 2 models need to be much more accurate in their learning. Secondly, they are likely to have to disambiguate between a much larger number of plausible concepts [Hill et al., 2019]. Finally, static datasets will likely have to cover a large number of combinations of composite parts which can quickly become impractical. OpenAI's attempt [Cobbe et al., 2021] at training GPT-3 to solve mathematical problems using verifiers uses active learning to train the verifier and provides support for my

[3] Csordás et al. [2021b] provides mixed evidence in terms of functional modularity. Based on empirical evidence, it makes two conclusions. (1) Disjoint units within NN do specialize in specific functions. (2) NNs do not learn to reuse these units for similar inputs. However, in my opinion, the experiment used in drawing the second conclusion did not provide strong enough evidence for this conclusion.

claims. The training setup used in this work is quite similar to the one used in learning from human feedback [Christiano et al., 2017] in inverse reinforcement learning (IRL) [Ng et al., 2000]. The problem of efficient supervision is at the heart of IRL and if active learning is indeed the answer to inducing systematic generalization, then insights from IRL can be hugely beneficial to minimize the cost of human supervision. if this is indeed the case, I expect my prior experience [Malik et al., 2021] with IRL to be quite beneficial here.

Conclusion

Systematic generalization, and out of distribution generalization [Zhang et al., 2021, Krueger et al., 2021] in general, are criticial challenges for machine learning. I have proposed attacking this problem by focusing on achieving a better understanding of *ingredients* of systematicity within NNs. While developing a complete theory here may be too ambitious, developing an engineering level theory, for inducing systematic generalization in NNs is very much possible and should at least be attempted.

I must also stress here that while I have singled out the problem of understanding the ingredients of systematicity in this proposal, there are many other aspects of systematic generalization that are worth studying as well. For example, measuring systematic generalization itself is a challenging task [Kim and Linzen, 2020, Keysers et al., 2020]. Similarly, systematic generalization of foundation models [Brown et al., 2020, Bommasani et al., 2021] or reinforcement learning [Anonymous, 2022, Kirk et al., 2021] are other rich areas with great research potential. Indeed, these problems are highly interesting and I remain open to research about them and other similar problems.

References

E. Akyürek, A. F. Akyürek, and J. Andreas. Learning to recombine and resample data for compositional generalization. *arXiv preprint arXiv:2010.03706*, 2020.

J. Andreas. Measuring compositionality in representation learning. *arXiv preprint arXiv:1902.07181*, 2019.

Anonymous. Procedural generalization by planning with self-supervised world models. In *Submitted to The Tenth International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=FmBegXJToY`. under review.

D. Bahdanau, S. Murty, M. Noukhovitch, T. H. Nguyen, H. de Vries, and A. Courville. Systematic generalization: what is required and can it be learned? *arXiv preprint arXiv:1811.12889*, 2018.

L. Bergen, T. J. O'Donnell, and D. Bahdanau. Systematic generalization with edge transformers. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.

R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *arXiv preprint arXiv:1706.03741*, 2017.

K. Cobbe, V. Kosaraju, M. Bavarian, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

R. Csordás, K. Irie, and J. Schmidhuber. The devil is in the detail: Simple tricks improve systematic generalization of transformers. *arXiv preprint arXiv:2108.12284*, 2021a.

R. Csordás, S. van Steenkiste, and J. Schmidhuber. Are neural nets modular? inspecting functional modularity through differentiable weight masks. In *International Conference on Learning Representations*, 2021b. URL `https://openreview.net/forum?id=7uVcpu-gMD`.

F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht. Essentially no barriers in neural network energy landscape. In *International conference on machine learning*, pages 1309–1318. PMLR, 2018.

D. Filan, S. Casper, S. Hod, C. Wild, A. Critch, and S. Russell. Clusterability in neural networks. *arXiv preprint arXiv:2103.03386*, 2021.

J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

T. Garipov, P. Izmailov, D. Podoprikhin, D. Vetrov, and A. G. Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8803–8812, 2018.

I. Gulrajani and D. Lopez-Paz. In search of lost domain generalization. *arXiv preprint arXiv:2007.01434*, 2020.

F. Hill, A. Santoro, D. G. Barrett, A. S. Morcos, and T. Lillicrap. Learning to make analogies by contrasting abstract relational structure. *arXiv preprint arXiv:1902.00120*, 2019.

F. Hill, A. Lampinen, R. Schneider, S. Clark, M. Botvinick, J. L. McClelland, and A. Santoro. Environmental drivers of systematicity and generalization in a situated agent. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SklGryBtwr`.

D. A. Hudson and C. D. Manning. Learning by abstraction: The neural state machine. *arXiv preprint arXiv:1907.03950*, 2019.

D. Kahneman. *Thinking, fast and slow*. Macmillan, 2011.

D. Keysers, N. Schärli, N. Scales, H. Buisman, D. Furrer, S. Kashubin, N. Momchev, D. Sinopalnikov, L. Stafiniak, T. Tihon, D. Tsarkov, X. Wang, M. van Zee, and O. Bousquet. Measuring compositional generalization: A comprehensive method on realistic data. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=SygcCnNKwr`.

N. Kim and T. Linzen. Cogs: A compositional generalization challenge based on semantic interpretation. *arXiv preprint arXiv:2010.05465*, 2020.

R. Kirk, A. Zhang, E. Grefenstette, and T. Rocktäschel. A survey of generalisation in deep reinforcement learning. *arXiv preprint arXiv:2111.09794*, 2021.

D. Krueger, E. Caballero, J.-H. Jacobsen, A. Zhang, J. Binas, D. Zhang, R. Le Priol, and A. Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, 2021.

B. Lake and M. Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018.

S. Malik, U. Anwar, A. Aghasi, and A. Ahmed. Inverse constrained reinforcement learning. In *International Conference on Machine Learning*, pages 7390–7399. PMLR, 2021.

G. F. Marcus. Rethinking eliminative connectionism. *Cognitive psychology*, 37(3):243–282, 1998.

R. Matz. Why ai alignment could be hard with modern deep learning. `https://edition.cnn.com/2020/07/28/tech/ai-recipes-sound-human/index.html`, 2021. Accessed: 20 October, 2021.

A. J. Nam and J. L. McClelland. What underlies rapid learning and systematic generalization in humans. *arXiv preprint arXiv:2107.06994*, 2021.

A. Y. Ng, S. J. Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, page 2, 2000.

J. Russin, J. Jo, R. C. O'Reilly, and Y. Bengio. Compositional generalization in a deep seq2seq model by separating syntax and semantics. *arXiv preprint arXiv:1904.09708*, 2019.

W. Stammer, P. Schramowski, and K. Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3619–3629, 2021.

D. Zhang, K. Ahuja, Y. Xu, Y. Wang, and A. Courville. Can subnetwork structure be the key to out-of-distribution generalization? *arXiv preprint arXiv:2106.02890*, 2021.