Software Project: Adapting Large Language Models to Human Feedback (w/ and w/o Reinforcement Learning)

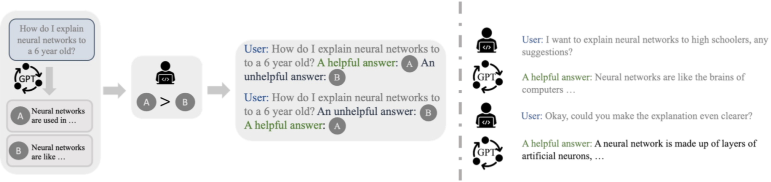
Stefan Riezler

SoSe 2023

Large Language Models (LLMs)

- chatGPT and friends - Revolution not only in societal impact, but also paradigm shift in research
    - Pre-trained large language models learn from few-shot demonstrations, specified via text interactions with the model (GPT-3 [Brown et al., 2020])
    - Additional fine-tuning by reinforcement learning from human feedback (chatGPT [Christiano et al., 2017, Kreutzer et al., 2018, Ouyang et al., 2022])

Reinforcement Learning from Human Feedback via Text Interactions


How do I explain neural networks to a 6 year old?
GPT
A Neural networks are used in …
B Neural networks are like …
A > B

User: How do I explain neural networks to a 6 year old? A helpful answer: A An unhelpful answer: B

User: How do I explain neural networks to a 6 year old? An unhelpful answer: B A helpful answer: A

User: I want to explain neural networks to high schoolers, any suggestions?

A helpful answer: Neural networks are like the brains of computers …

User: Okay, could you make the explanation even clearer?

A helpful answer: A neural network is made up of layers of artificial neurons, …

- Human feedback is input as text sequence, fine-tuning only by few-shot prompting / in-context learning [Liu et al., 2023, Madaan et al., 2023]

Project Idea

- Choose a text generation task
  - Example: Machine translation by prompting LLMs (Bloom[1], mT5[2]) [Vilar et al., 2022]
- Collect feedback on model output
  - Example: Human feedback on machine translations [Kreutzer et al., 2018], or simulated by metrics like COMET[3]
- Fine-tune model on textual feedback to model outputs
  - Option 1: Fine-tune with RLHF [Ouyang et al., 2022][4]
  - Option 2: Fine-tune with Cross-Entropy [Liu et al., 2023][5]
  - Option 3: Fine-tune with few-shot prompting/in-context learning (apply [Madaan et al., 2023] to human feedback)

[1] https://huggingface.co/docs/transformers/model_doc/bloom

[2] https://huggingface.co/docs/transformers/model_doc/mt5

[3] https://huggingface.co/Unbabel/wmt22-comet-da

[4] https://github.com/CarperAI/trlx

[5] https://huggingface.co/blog/peft

Project Goal

- Learn to work with pre-trained large language models
- Provide small amounts of feedback yourself
- Understand how large language models learn from your feedback

References

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).
Language models are few-shot learners.
In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 1877–1901. Curran Associates, Inc.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017).
Deep reinforcement learning from human preferences.
In *Advances in Neural Information Processing Systems (NIPS)*, Long Beach, CA, USA.

Kreutzer, J., Uyheng, J., and Riezler, S. (2018).
Reliability and learnability of human bandit feedback for sequence-to-sequence reinforcement learning.
In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, Melbourne, Australia.

Liu, H., Sferrazza, C., and Abbeel, P. (2023).
Chain of hindsight aligns language models with feedback.

Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegreffe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Welleck, S., Majumder, B. P., Gupta, S., Yazdanbakhsh, A., and Clark, P. (2023).
Self-refine: Iterative refinement with self-feedback.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Gray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022).
Training language models to follow instructions with human feedback.
In *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, USA.

Vilar, D., Freitag, M., Cherry, C., Luo, J., Ratnakar, V., and Foster, G. (2022).
Prompting palm for translation: Assessing strategies and performance.