Attention, GPT, ViT, and ChatGPT

Lab Talk

TECHNISCHE UNIVERSITÄT DARMSTADT

Christoph Reich
https://christophreich1996.github.io
creich@nec-labs.com

Content

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Attention Is All You Need
- Introduction
- Attention
- Self- *vs.* Cross-Attention
- Multi-Head-Attention
- Transformer
- Limitations of Attention
- Recent Advances

2. GPT
- Overview
- Architecture & Training
- Scaling Law

3. ViT
- Overview
- Architecture

4. ChatGPT
- Overview
- Training

Content

TECHNISCHE UNIVERSITÄT DARMSTADT

Introduction - Linear Layer

Attention Is All You Need

TECHNISCHE UNIVERSITÄT DARMSTADT

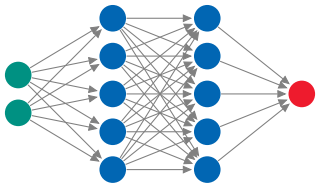


Figure: Feed forward neural network composed of three linear layers.

$$\boldsymbol{f} : \mathbb{R}^2 \rightarrow \mathbb{R}^1,\ \boldsymbol{f}(\boldsymbol{x}) = g(g(g(\boldsymbol{x}^\intercal \boldsymbol{W}_1)\, \boldsymbol{W}_2)\, \boldsymbol{W}_3)\,,\ \boldsymbol{x} \in \mathbb{R}^2,\ \boldsymbol{W}_1 \in \mathbb{R}^{2\times 5},\ \boldsymbol{W}_2 \in \mathbb{R}^{5\times 5},\ \boldsymbol{W}_1 \in \mathbb{R}^{5\times 1} \tag{1}$$

Introduction - Linear Layer

Attention Is All You Need

TECHNISCHE UNIVERSITÄT DARMSTADT

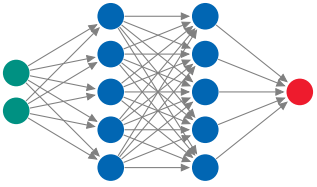


Figure: Feed forward neural network composed of three linear layers.

$$\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^1,\ \boldsymbol{f}(\boldsymbol{x}) = g(g(g(\boldsymbol{x}^\intercal \boldsymbol{W}_1)\,\boldsymbol{W}_2)\,\boldsymbol{W}_3)\,,\ \boldsymbol{x} \in \mathbb{R}^2,\ \boldsymbol{W}_1 \in \mathbb{R}^{2\times 5},\ \boldsymbol{W}_2 \in \mathbb{R}^{5\times 5},\ \boldsymbol{W}_1 \in \mathbb{R}^{5\times 1} \tag{1}$$

Observation

f applies the same weights to all inputs!

Introduction - Attention High-Lever Intuition

Attention Is All You Need

TECHNISCHE UNIVERSITÄT DARMSTADT

Intuition

Attention is just a **dynamic** linear layer.

Introduction - Attention High-Lever Intuition

Attention Is All You Need

TECHNISCHE UNIVERSITÄT DARMSTADT

Intuition

Attention is just a **dynamic** linear layer.

Weights are generated dynamically based on the given input.

$$\boldsymbol{x}^\top \boldsymbol{W}(\boldsymbol{x}) \tag{2}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Intuition

Attention is just a **dynamic** linear layer.

Weights are generated dynamically based on the given input.

$$\boldsymbol{x}^\intercal \boldsymbol{W}(\boldsymbol{x}) \tag{2}$$

Advantages of Attention

- Attention can vastly adapt to different inputs
- In practice more expressive than linear layers
- Attention can adapt to different shapes of $\boldsymbol{x}$
- Attention is a global operation

Scaled Dot-Product Attention I

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

Scaled Dot-Product Attention (standard form of Attention) is defined as: [Bahdanau et al., 2015]

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\intercal}{\sqrt{n}}\right)\boldsymbol{V},\ \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{n \times c}. \tag{3}$$

Assume for now $\boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{V}$.

Scaled Dot-Product Attention I

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

Scaled Dot-Product Attention (standard form of Attention) is defined as: [Bahdanau et al., 2015]

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\intercal}{\sqrt{n}}\right)\boldsymbol{V},\ \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{n \times c}. \tag{3}$$

Assume for now $\boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{V}$.

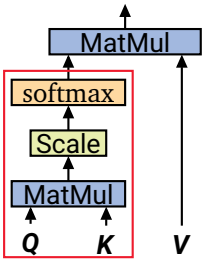


Figure: Compute graph of the Scaled Dot-Product Attention operation.

Scaled Dot-Product Attention II

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

Scaled Dot-Product Attention with linear mappings:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{(\boldsymbol{Q}\boldsymbol{W}_Q)\,(\boldsymbol{K}\boldsymbol{W}_K)^\mathsf{T}}{\sqrt{n}}\right)\boldsymbol{V}\boldsymbol{W}_V,\ \boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{n \times c_i}, \boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{c_i \times c_a}. \quad (4)$$

$\boldsymbol{W}_Q$, $\boldsymbol{W}_K$, $\boldsymbol{W}_V$ are just learnable linear mappings.

Self- *vs.* Cross-Attention

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

In **Self-Attention** $\boldsymbol{X} = \boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{V}$ $(\text{Attention}(\boldsymbol{X}, \boldsymbol{X}, \boldsymbol{X}))$

- Attention is performed between the input itself

Self- *vs.* Cross-Attention

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

In **Self-Attention** $\boldsymbol{X} = \boldsymbol{Q} = \boldsymbol{K} = \boldsymbol{V}$ $(\text{Attention}(\boldsymbol{X}, \boldsymbol{X}, \boldsymbol{X}))$

- Attention is performed between the input itself

In **Cross-Attention** $\boldsymbol{X}_1 = \boldsymbol{Q}, \boldsymbol{X}_2 = \boldsymbol{K} = \boldsymbol{V}$ $(\text{Attention}(\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_2))$

- Attention is performed between two different inputs
- Number of tokens in $\boldsymbol{X}_1 \in \mathbb{R}^{n_1 \times c_i}$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n_2 \times c_i}$ can differ
- Can be interpreted intuitively as a conditioning

Multi-Head-Attention

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

Limitation of Attention, just a single Attention matrix constructed.

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \underbrace{\text{softmax}\left(\frac{(\boldsymbol{Q}\boldsymbol{W}_Q)\,(\boldsymbol{K}\boldsymbol{W}_K)^\intercal}{\sqrt{n}}\right)}_{\in\mathbb{R}^{n\times n}}\boldsymbol{V}\boldsymbol{W}_V,\ \boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbb{R}^{n\times c_i},\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V\in\mathbb{R}^{c_i\times c_a}. \quad (5)$$

Multi-Head-Attention

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

Limitation of Attention, just a single Attention matrix constructed.

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \underbrace{\text{softmax}\left(\frac{(\boldsymbol{Q}\boldsymbol{W}_Q)\,(\boldsymbol{K}\boldsymbol{W}_K)^\intercal}{\sqrt{n}}\right)}_{\in\mathbb{R}^{n\times n}}\boldsymbol{V}\boldsymbol{W}_V,\ \boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbb{R}^{n\times c_i},\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V\in\mathbb{R}^{c_i\times c_a}. \quad (5)$$

Idea

Let's use multiple Attention matrices, each learning different features.

Multi-Head-Attention

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

Limitation of Attention, just a single Attention matrix constructed.

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \underbrace{\text{softmax}\left(\frac{(\boldsymbol{Q}\boldsymbol{W}_Q)\,(\boldsymbol{K}\boldsymbol{W}_K)^\intercal}{\sqrt{n}}\right)}_{\in\mathbb{R}^{n\times n}} \boldsymbol{V}\boldsymbol{W}_V,\ \boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\in\mathbb{R}^{n\times c_i}, \boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V\in\mathbb{R}^{c_i\times c_a}. \quad (5)$$

Idea

Let's use multiple Attention matrices, each learning different features.

$$\text{MultiHeadAttention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{ConCat}(\text{head}_1,\ldots,\text{head}_h)\,\boldsymbol{W}_O,\ \boldsymbol{W}_O\in\mathbb{R}^{c_a\times c_o} \quad (6)$$

$$\text{where head}_i = \text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) \quad (7)$$

Attention utilizes $\boldsymbol{W}_Q,\boldsymbol{W}_K,\boldsymbol{W}_V\in\mathbb{R}^{c_i\times c_a/h}$.

Transformer

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

The Transformer block utilizes multiple advances in deep learning.

- Multi-Head Attention [Vaswani et al., 2017]
- Layer Normalization [Ba et al., 2016]
- Residual skip connections [He et al., 2016]

Transformer

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

The Transformer block utilizes multiple advances in deep learning.

- Multi-Head Attention [Vaswani et al., 2017]
- Layer Normalization [Ba et al., 2016]
- Residual skip connections [He et al., 2016]

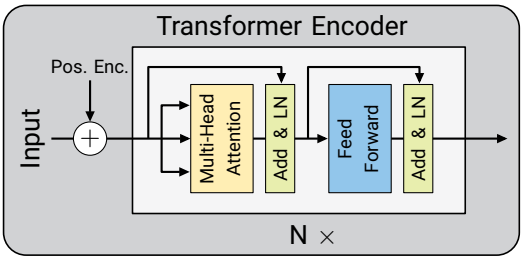


Figure: Transformer encoder composed of N Transformer blocks. Figure adapted from [Rohr et al., 2022].

Limitations of Attention

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

The Attention operation and Transformers entails multiple issues:

- Computationally expensive
  - Computational complexity of Self-Attention $\mathcal{O}(n^2\, d)$
  - Memory complexity Self-Attention $\mathcal{O}(n^2 + nd)$ (in practice or during training with Backprop [Rabe and Staats, 2021])

Limitations of Attention

Attention Is All You Need [Vaswani et al., 2017]

TECHNISCHE UNIVERSITÄT DARMSTADT

The Attention operation and Transformers entails multiple issues:

- Computationally expensive
  - Computational complexity of Self-Attention $\mathcal{O}(n^2 d)$
  - Memory complexity Self-Attention $\mathcal{O}(n^2 + nd)$ (in practice or during training with Backprop [Rabe and Staats, 2021])
- Training requires generally a lot of data
  - Attention operations is very general and does not encode inductive biases
  - Architectures with more inductive biases (*e.g.* CNNs) typically perform better on limited data
  - Large scale pre-training is needed to achieve strong results on small datasets

Attention – Recent Advances

Attention Is All You Need

TECHNISCHE UNIVERSITÄT DARMSTADT

Overcoming the limitations of Attention and Transformers is an active body of research.

- Linearizing the Attention operation
  - Linformer [Wang et al., 2020]
  - Performer [Choromanski et al., 2021]
- Reformulating the Attention operation
  - Efficient Attention [Shen et al., 2021]
  - XCiT [Ali et al., 2021]
- Local & Sparse Attention
  - Reformer [Kitaev et al., 2020]
  - Axial-Attention [Ho et al., 2019]
  - Shifted Window Attention [Liu et al., 2021, Liu et al., 2022]
- Replacing the Attention operation in Transformers
  - FNet [Lee-Thorp et al., 2021]
  - MLP-Mixer [Tolstikhin et al., 2021]
  - Token Pooling [Marin et al., 2023]

Content

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Attention Is All You Need
   - Introduction
   - Attention
   - Self- *vs.* Cross-Attention
   - Multi-Head-Attention
   - Transformer
   - Limitations of Attention
   - Recent Advances
2. GPT
   - Overview
   - Architecture & Training
   - Scaling Law
3. ViT
   - Overview
   - Architecture
4. ChatGPT
   - Overview
   - Training

Overview
GPT

TECHNISCHE UNIVERSITÄT DARMSTADT

Generative Pre-trained Transformer (GPT) is an autoregressive language model.

Generative Pre-trained Transformer (GPT) is an autoregressive language model.

Language models (LM) aim to model the conditional probability of the next word given all previous ones [Bengio et al., 2000].

In particular, large language models (LLM), such as GPT, model the joint probability distribution over words/symbols as the product of conditional probabilities.

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, \ldots, s_{n-1}) \tag{8}$$

TECHNISCHE UNIVERSITÄT DARMSTADT

Generative Pre-trained Transformer (GPT) is an autoregressive language model.

Language models (LM) aim to model the conditional probability of the next word given all previous ones [Bengio et al., 2000].

In particular, large language models (LLM), such as GPT, model the joint probability distribution over words/symbols as the product of conditional probabilities.

$$p(x) = \prod_{i=1}^{n} p(s_n | s_1, \ldots, s_{n-1}) \tag{8}$$

Intuition

GPT takes in past words and predicts a distribution over all words (in dictionary) describing how probable each word is to come next. Sample from this distribution and repeat the prediction.

Architecture & Training

GPT

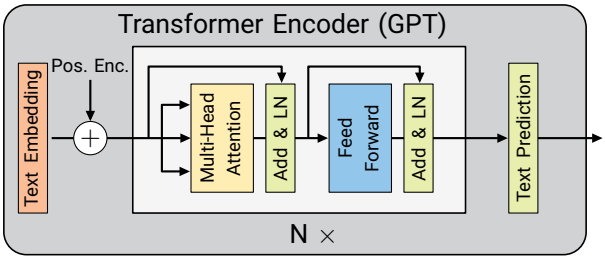


Figure: GPT architecture composed of stacked transformer blocks [Radford et al., 2019, Brown et al., 2020].

GPT-3 stacks 96 transformer blocks, resulting in 175B parameters [Brown et al., 2020].

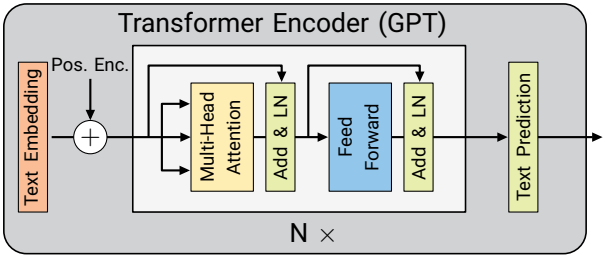


Figure: GPT architecture composed of stacked transformer blocks [Radford et al., 2019, Brown et al., 2020].

GPT-3 stacks 96 transformer blocks, resulting in 175B parameters [Brown et al., 2020].

Unsupervised language pre-training utilizes, the **Common Crawl dataset** (filtered). Dataset size before filtering is 45 TB and after filtering 570 GB. [Brown et al., 2020]

Scaling Law

GPT

TECHNISCHE UNIVERSITÄT DARMSTADT

Scaling up LLMs (model and/or dataset) can drastically improve their performance.

Scaling Law

GPT

TECHNISCHE UNIVERSITÄT DARMSTADT

Scaling up LLMs (model and/or dataset) can drastically improve their performance.

The performance of LLM improves as a power law with respect to the **dataset** size, **model** size, and the amount of **compute** used for training.

Scaling Law

GPT

TECHNISCHE UNIVERSITÄT DARMSTADT

Scaling up LLMs (model and/or dataset) can drastically improve their performance.

The performance of LLM improves as a power law with respect to the **dataset** size, **model** size, and the amount of **compute** used for training.

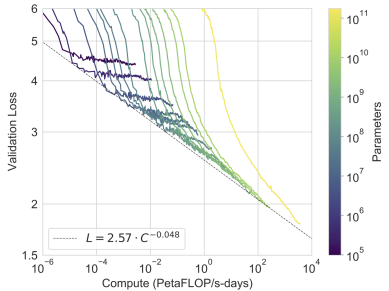


Figure: Scaling the model size and compute of GPT-3. Figure from [Brown et al., 2020].

Content

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Attention Is All You Need
   - Introduction
   - Attention
   - Self- *vs.* Cross-Attention
   - Multi-Head-Attention
   - Transformer
   - Limitations of Attention
   - Recent Advances
2. GPT
   - Overview
   - Architecture & Training
   - Scaling Law
3. **ViT**
   - Overview
   - Architecture
4. ChatGPT
   - Overview
   - Training

How to utilize the Transformer architecture for vision tasks?

How to utilize the Transformer architecture for vision tasks?

Issue

Interpreting every pixel as a token is infeasible due to quadratic complexity of the Attention operation.

How to utilize the Transformer architecture for vision tasks?

Issue

Interpreting every pixel as a token is infeasible due to quadratic complexity of the Attention operation.

Vision Transformer (ViT) idea

Utilize image patches as tokens and not pixels [Dosovitskiy et al., 2021].

Architecture

ViT

TECHNISCHE UNIVERSITÄT DARMSTADT

The Vision Transformer utilizes the standard Transformer encoder architecture.

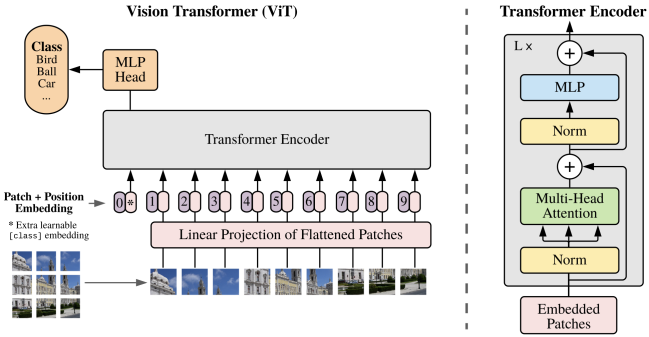


Figure: Vision Transformer architecture. Image taken from [Dosovitskiy et al., 2021].

Content

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Attention Is All You Need
   - Introduction
   - Attention
   - Self- *vs.* Cross-Attention
   - Multi-Head-Attention
   - Transformer
   - Limitations of Attention
   - Recent Advances
2. GPT
   - Overview
   - Architecture & Training
   - Scaling Law
3. ViT
   - Overview
   - Architecture
4. **ChatGPT**
   - Overview
   - Training

Overview

ChatGPT [OpenAI, 2022]

TECHNISCHE UNIVERSITÄT DARMSTADT

Disclaimer

No peer-reviewed article on ChatGPT has yet been published! Currently, only a non-peer-reviewed blog post and a demo have been released by OpenAI [OpenAI, 2022].

Overview

ChatGPT [OpenAI, 2022]

TECHNISCHE UNIVERSITÄT DARMSTADT

Disclaimer

No peer-reviewed article on ChatGPT has yet been published! Currently, only a non-peer-reviewed blog post and a demo have been released by OpenAI [OpenAI, 2022].

ChatGPT is a chatbot based on the GPT model family.

Overview

ChatGPT [OpenAI, 2022]

TECHNISCHE UNIVERSITÄT DARMSTADT

Disclaimer

No peer-reviewed article on ChatGPT has yet been published! Currently, only a non-peer-reviewed blog post and a demo have been released by OpenAI [OpenAI, 2022].

ChatGPT is a chatbot based on the GPT model family.

- ChatGPTs architecture is based on GPT-3.5, a Transformer-based model.
- Presumable ChatGPT entails 175B parameters. Smaller variants are probably also available (6B and 1.3B).
- ChatGPT is using “same” methods as InstructionGPT [Ouyang et al., 2022] (with slight differences).

Training

ChatGPT [OpenAI, 2022]

TECHNISCHE UNIVERSITÄT DARMSTADT

ChatGPT (GPT-3.5) is pre-trained on **text** (similar to GPT-3) and **code** (similar to GitHub Copilot).

To align ChatGPT on instructions (chat-setting) **Reinforcement Learning from Human Feedback** (RLHF) is employed.

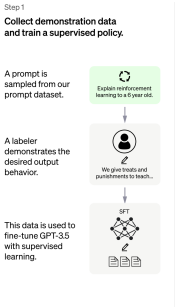
Training

ChatGPT [OpenAI, 2022]

TECHNISCHE UNIVERSITÄT DARMSTADT

ChatGPT (GPT-3.5) is pre-trained on **text** (similar to GPT-3) and **code** (similar to GitHub Copilot).

To align ChatGPT on instructions (chat-setting) **Reinforcement Learning from Human Feedback** (RLHF) is employed.



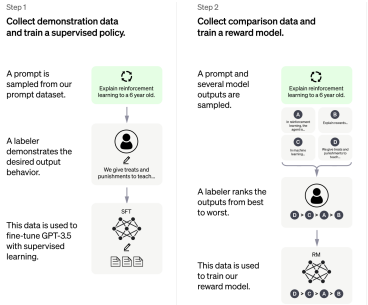
Training

ChatGPT [OpenAI, 2022]

TECHNISCHE UNIVERSITÄT DARMSTADT

ChatGPT (GPT-3.5) is pre-trained on **text** (similar to GPT-3) and **code** (similar to GitHub Copilot).

To align ChatGPT on instructions (chat-setting) **Reinforcement Learning from Human Feedback** (RLHF) is employed.



Training

ChatGPT [OpenAI, 2022]

TECHNISCHE UNIVERSITÄT DARMSTADT

ChatGPT (GPT-3.5) is pre-trained on **text** (similar to GPT-3) and **code** (similar to GitHub Copilot).

To align ChatGPT on instructions (chat-setting) **Reinforcement Learning from Human Feedback** (RLHF) is employed.

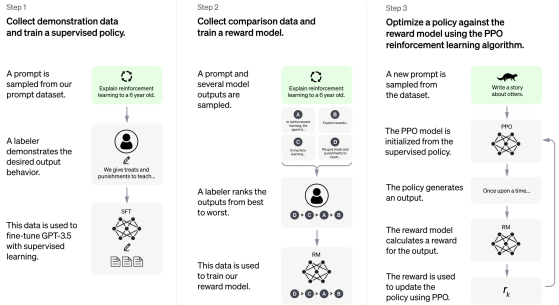


Figure: RLHF-based fine tuning procedure of ChatGPT. Figure taken from [OpenAI, 2022].

Questions?

Slides are available at:

https://christophreich1996.github.io/pdfs/lab_talk_16_1_2023.pdf

References I

TECHNISCHE UNIVERSITÄT DARMSTADT

Ali, A., Touvron, H., Caron, M., Bojanowski, P., Douze, M., Joulin, A., Laptev, I., Neverova, N., Synnaeve, G., Verbeek, J., and Jegou, H. (2021).
XCiT: Cross-Covariance Image Transformers.
In *Advances in Neural Information Processing Systems*, volume 34, pages 20014–20027.

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016).
Layer Normalization .
In *NIPS 2016 Deep Learning Symposium recommendation*.

Bahdanau, D., Cho, K., and Bengio, Y. (2015).
Neural Machine Translation by Jointly Learning to Align and Translate.
In *International Conference on Learning Representations*.

References II

TECHNISCHE UNIVERSITÄT DARMSTADT

Bengio, Y., Ducharme, R., and Vincent, P. (2000).
A Neural Probabilistic Language Model.
In *Advances in Neural Information Processing Systems*, volume 13, pages 24261–24272.

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 *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

References III

TECHNISCHE UNIVERSITÄT DARMSTADT


Choromanski, K. M., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J. Q., Mohiuddin, A., Kaiser, L., Belanger, D. B., Colwell, L. J., and Weller, A. (2021).
Rethinking Attention with Performers.
In *International Conference on Learning Representations*.


Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021).
An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.
In *International Conference on Learning Representations*.


He, K., Zhang, X., Ren, S., and Sun, J. (2016).
Deep Residual Learning for Image Recognition.
In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 770–778.

References IV

Ho, J., Kalchbrenner, N., Weissenborn, D., and Salimans, T. (2019).
Axial Attention in Multidimensional Transformers.
arXiv preprint arXiv:1912.12180.

Kitaev, N., Kaiser, L., and Levskaya, A. (2020).
Reformer: The Efficient Transformer.
In *International Conference on Learning Representations*.

Lee-Thorp, J., Ainslie, J., Eckstein, I., and Ontañón, S. (2021).
Fnet: Mixing tokens with fourier transforms.
In *North American Chapter of the Association for Computational Linguistics*.

References V

TECHNISCHE UNIVERSITÄT DARMSTADT

Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., Wei, Y., Ning, J., Cao, Y., Zhang, Z., Dong, L., et al. (2022).
Swin Transformer V2: Scaling Up Capacity and Resolution.
In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12009–12019.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021).
Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows.
In *IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

Marin, D., Chang, J.-H. R., Ranjan, A., Prabhu, A., Rastegari, M., and Tuzel, O. (2023).
Token Pooling in Vision Transformers for Image Classification.
In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 12–21.

OpenAI (2022).
ChatGPT: Optimizing Language Models for Dialogue.

References VI

TECHNISCHE UNIVERSITÄT DARMSTADT


Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022).
Training language models to follow instructions with human feedback.
arXiv preprint arXiv:2203.02155.


Rabe, M. N. and Staats, C. (2021).
Self-Attention Does Not Need $\mathcal{O}(n^2)$ Memory.
arXiv preprint arXiv:2112.05682.


Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019).
Language Models are Unsupervised Multitask Learners.

References VII

TECHNISCHE UNIVERSITÄT DARMSTADT

Rohr, M., Reich, C., Höhl, A., Lilienthal, T., Dege, T., Plesinger, F., Bulkova, V., Clifford, G. D., Reyna, M. A., and Antink, C. H. (2022).
Exploring Novel Algorithms for Atrial Fibrillation Detection by Driving Graduate Level Education in Medical Machine Learning.
Physiological Measurement.

Shen, Z., Zhang, M., Zhao, H., Yi, S., and Li, H. (2021).
Efficient Attention: Attention With Linear Complexities.
In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3531–3539.

References VIII

Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., Yung, J., Steiner, A., Keysers, D., Uszkoreit, J., Lucic, M., and Dosovitskiy, A. (2021).
MLP-Mixer: An all-MLP Architecture for Vision.
In *Advances in Neural Information Processing Systems*, volume 34, pages 24261–24272.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017).
Attention is All You Need.
In *Advances in Neural Information Processing Systems*, volume 30.

Wang, S., Li, B. Z., Khabsa, M., Fang, H., and Ma, H. (2020).
Linformer: Self-Attention with Linear Complexity.
arXiv preprint arXiv:2006.04768.