

# Attention is All You Need

overview of the transformer architecture,  
applications and established improvements

---

Felix Karg

April 3, 2023

Institute for Theoretical Informatics:  
Artificial Intelligence for Materials Science



**Compression**

**Successes**

**Compression**

Successes

# Compression

---

**Quantization**

Distillation

Rank Reduction

Memorizing Transformer

Reflexion

# Quantization

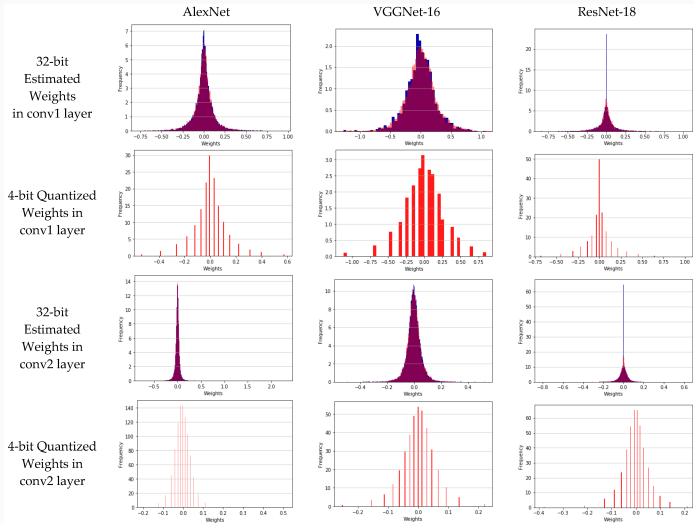


Image Source: [1] SOTA: GPTQ [2]

# Compression

---

Quantization

**Distillation**

Rank Reduction

Memorizing Transformer

Reflexion

# Knowledge Distillation

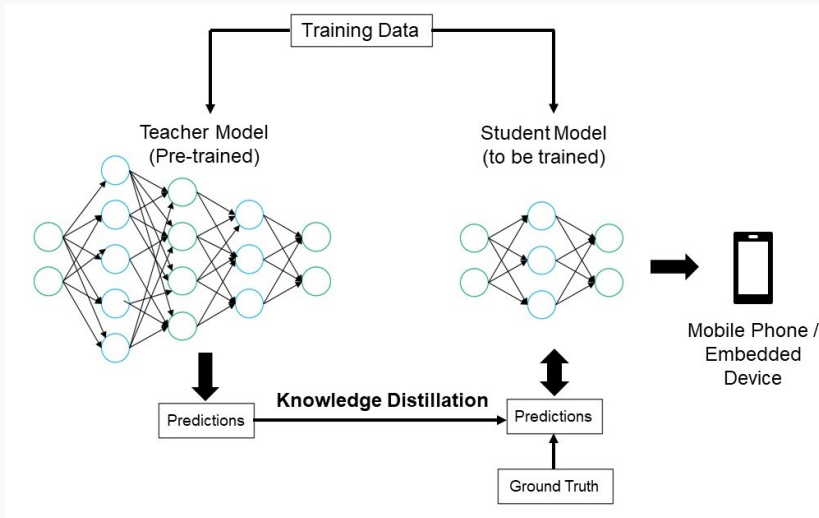


Image Source: [3]

# Patient Knowledge Distillation

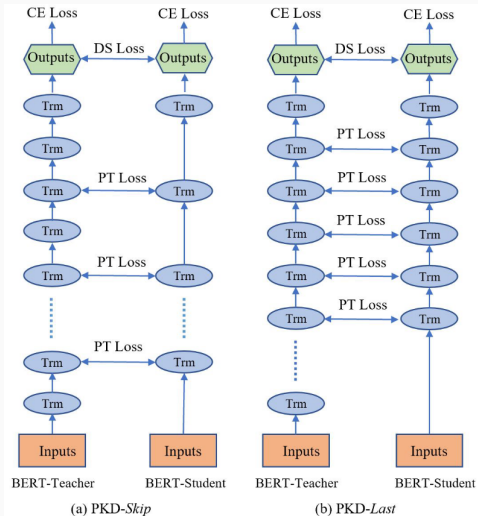


Image Source: [4]



# Compression

---

Quantization

Distillation

**Rank Reduction**

Memorizing Transformer

Reflexion

LoRA [5] (spun up for LLaMa just a month after 'release')  
(TODO: use graphics from LoRA?)

# Compression

---

Quantization

Distillation

Rank Reduction

**Memorizing Transformer**

Reflexion

Memorizing [6] (TODO: use loss curves?)

# Compression

---

Quantization

Distillation

Rank Reduction

Memorizing Transformer

**Reflexion**

Reflexion [7] (TODO: use loss curves and architecture)

Compression

**Successes**

# Successes

---

**BERT**

GPT

CLIP

Latent Diffusion Models

Reinforcement Learning

Physics Simulation

Meta Cognition



generating language embeddings for NLP tasks BERT [8]  
(TODO: show image of architecture)

# Successes

---

BERT

**GPT**

CLIP

Latent Diffusion Models

Reinforcement Learning

Physics Simulation

Meta Cognition

# GPT

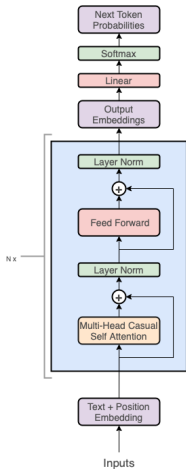


Image Source: [9]

# Successes

---

BERT

GPT

**CLIP**

Latent Diffusion Models

Reinforcement Learning

Physics Simulation

Meta Cognition

Building semantic embeddings shared from pictures and text  
CLIP [10] (TODO: show image of architecture)

# Successes

---

BERT

GPT

CLIP

**Latent Diffusion Models**

Reinforcement Learning

Physics Simulation

Meta Cognition

# Latent Diffusion Models

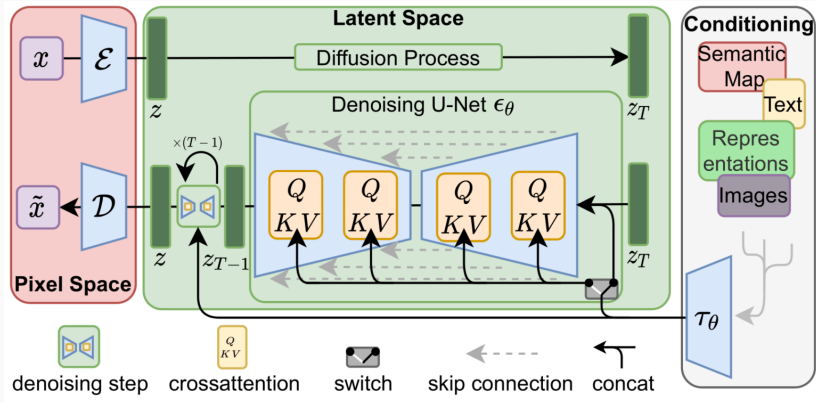


Image Source: [7]

# Successes

---

BERT

GPT

CLIP

Latent Diffusion Models

**Reinforcement Learning**

Physics Simulation

Meta Cognition



# GATO: A Generalist Agent

GATO [11] (TODO: architecture? Loss curves?)

# Successes

---

BERT

GPT

CLIP

Latent Diffusion Models

Reinforcement Learning

**Physics Simulation**

Meta Cognition

Speed up simulation by 150x [12] (TODO: show comparison graph)

# Successes

---

BERT

GPT

CLIP



Latent Diffusion Models

Reinforcement Learning




Physics Simulation

**Meta Cognition**




Reflexion [13] (TODO: include architecture)

-  [1] S. Seo and J. Kim, “Efficient Weights Quantization of Convolutional Neural Networks Using Kernel Density Estimation based Non-uniform Quantizer,” *Applied Sciences*, vol. 9, p. 2559, Jan. 2019.
-  [2] E. Frantar, S. Ashkboos, T. Hoefler, and D. Alistarh, “GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers,” *arXiv:arXiv:2210.17323*, Oct. 2022.


-  [3] J. Gou, B. Yu, S. J. Maybank, and D. Tao, “Knowledge Distillation: A Survey,” *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, June 2021.
-  [4] S. Sun, Y. Cheng, Z. Gan, and J. Liu, “Patient Knowledge Distillation for BERT Model Compression,” *arXiv:arXiv:1908.09355*, Aug. 2019.
-  [5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-Rank Adaptation of Large Language Models,” *arXiv:arXiv:2106.09685*, Oct. 2021.


-  [6] Y. Wu, M. N. Rabe, D. Hutchins, and C. Szegedy, “Memorizing Transformers,” *arXiv:arXiv:2203.08913*, Mar. 2022.
-  [7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, “High-Resolution Image Synthesis with Latent Diffusion Models,” *arXiv:arXiv:2112.10752*, Apr. 2022.
-  [8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.



-  [9] J. Mody, “GPT in 60 Lines of NumPy,” Jan. 2023.
-  [10] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, “Learning Transferable Visual Models From Natural Language Supervision,” *arXiv:arXiv:2103.00020*, Feb. 2021.
-  [11] S. Reed, K. Zolna, E. Parisotto, S. G. Colmenarejo, A. Novikov, G. Barth-Maron, M. Gimenez, Y. Sulsky, J. Kay, J. T. Springenberg, T. Eccles, J. Bruce, A. Razavi, A. Edwards, N. Heess, Y. Chen, R. Hadsell,

O. Vinyals, M. Bordbar, and N. de Freitas, “A Generalist Agent,” *arXiv:arXiv:2205.06175*, Nov. 2022.

 [12] S. Wiewel, M. Becher, and N. Thuerey, “Latent Space Physics: Towards Learning the Temporal Evolution of Fluid Flow,” *Computer Graphics Forum*, vol. 38, no. 2, pp. 71–82, 2019.

 [13] N. Shinn, B. Labash, and A. Gopinath, “Reflexion: An autonomous agent with dynamic memory and self-reflection,” *arXiv:arXiv:2303.11366*, Mar. 2023.