# Contribution and Influence of GPT Models

*Yuqun Wu*
*yuqunwu2@illinois.edu*

## 1. Introduction

Natural Language Processing contains diverse tasks, and people used to design networks for specific tasks, to solve them individually. However, General Purpose Text (Radford & Narasimhan, 2018) manage to use Transformer (Vaswani et al. 2017) for pretraining and then finetuning the network for specific downstream tasks. GPT then becomes the state-of-the-art method in most Natural Language Processing tasks and draws attention to pretraining, and transformers. This review tries to explore its contribution and influence among different research fields.

## 2. Main Body

From my point of view, GPT has two major contributions -- the demonstration of the importance of unsupervised pretraining, the introduction of Transformer, and the leading to more general purpose methods.

First, it manages to demonstrate how unsupervised pretraining can achieve the generality of networks, and boost the performance of downstream applications after finetuning. As introduced by the paper, it is time-consuming to acquire adequate data for any specific downstream task training, and therefore unsupervised pretraining is more applicable and efficient. In their paper, they just guide the network to predict the next words based on previous observations, which is proven to be beneficial for the feature embedding of the networks. By doing this, the method, with a fixed network, manages to outperform 9 state-of-the-art methods out of 12 tasks, which are always designed specifically for the tasks. This explores a potential direction, as it would save a lot of time and energy for people to explore different solutions, and choose a method that acquires accuracy and generality at the same time. After that, more and more papers about pretraining have been published, and not limited to NLP. For instance, He et al. (2021) develops a Masked autoencoder using a similar idea to the paper, to pretrain a general network for downstream tasks like semantic

segmentation, object detection, etc, and achieve state-of-the-art performance. Baker et al.(2022) also manage to apply a pretraining technique for reinforcement learning, and their method can even build a diamond cave Pickaxe in 20 mins, which is faster than some players. However, I think the former one makes sense, as the unsupervised learning data is easy to acquire, while the latter one still makes extensive data collection works for the pretraining.

Second, it draws attention to transformers, which use self-attention to handle the correlation of each term in a sentence. This has been proven to be more powerful than traditional Recurrent Neural Network or Convolution Neural Network. The importance of self-attention is also explored. Similar to pretraining, this idea also helps researchers in other domains to develop different solutions for tasks. Dosovitskiy et al.(2020) also explore this method for semantic segmentation of images, which leads the popularity of applying transformers on computer vision tasks. Researchers in Computer Vision begin to rethink whether Convolution Neural Network is always a correct answer for different tasks.

Third, It also shows the possibility to make a general method to solve different types of problems. With this development, people develop GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020). They use more data for training, and also add more downstream tasks, managing to show what a method can achieve in different tasks. For instance, GPT-2 has been a network that can solve many different tasks at the same time, and GPT-3 is not only a prediction model, but also a generative method, which can imitate people to generate a paragraph. This idea also encourages other researchers to design general-purpose methods in other domains. Gupta et al. (2021) also apply similar ideas to develop a general-purpose vision, which solves different Computer Vision tasks at the same time.

## 3. Conclusion

GPT has outperformed most networks of NLP in that period, and introduced efficient training ideas, and powerful networks. At the same time, its influence should also be noticed, as it also enhances the development of different AI research fields to develop lots of decent solutions for different tasks.

# Reference


Baker, B., Akkaya, I., Zhokhov, P., Huizinga, J., Tang, J., Ecoffet, A., Houghton, B., Sampedro, R., & Clune, J. (2022). Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos. *ArXiv, abs/2206.11795*.

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T.J., Child, R., Ramesh, A., Ziegler, D.M., 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., & Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv, abs/2005.14165*.

Gupta, T., Kamath, A., Kembhavi, A., & Hoiem, D. (2021). Towards General Purpose Vision Systems. *ArXiv, abs/2104.00743*.

He, K., Chen, X., Xie, S., Li, Y., Doll'ar, P., & Girshick, R.B. (2022). Masked Autoencoders Are Scalable Vision Learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15979-15988.

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training.

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

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *ArXiv, abs/1706.03762*.