# Image GPT

## A large transformer model trained on language can generate coherent text, the same exact model trained on pixel sequences can generate coherent image completions and samples.

* **Unsupervised and self-supervised learning, or learning without human-labeled data, is a longstanding challenge of machine learning.**
* **Recently, it has seen incredible success in language, as transformermodels like BERT, GPT-2,RoBERTa, T5, and other variant have achieved top performance on a wide array of language tasks.**
* **However, the same broad class of models has not been successful in producing strong features for image classification.**
* **Transformer models like BERT and GPT-2 are domain agnostic, meaning that they can be directly applied to 1-D sequences of any form.**
* **GPT-2 on images unrolled into long sequences of pixels, which is iGPT, appears to understand 2-D image characteristics such as object appearance and category.**
* **This is evidenced by the diverse range of coherent image samples it generates, even without the guidance of human provided labels.**
* **Features from the model achieve state-of-the-art performance on a number of classification datasets and near state-of-the-art unsupervised accuracy.**
* **When faced with a new domain where the correct model priors are unknown, a large GPT-2 can learn excellent features without the need for domain-specificarchitectural design choices.**

**FROM LANGUAGE GPT TO IMAGE GPT**

* **In language, unsupervised learning algorithms that rely on word prediction (like GPT-2 and BERT) have been extremely successful, achieving top performance on a wide array of language tasks.**
* **One possible reason for this success is that instances of downstream language tasks appear naturally in text: questions are often followed by answers (which could help with question-answering) and passages are often followed by summaries (which could help with summarization).**
* **In contrast, sequences of pixels do not clearly contain labels for the images they belong to.**
* **A transformer is trained to maximize the likelihood, and thus is mode covering, which automatically ensures the diversity of its samples.**
* **Even without this explicit supervision, there is still a reason why GPT-2 on images might work: a sufficiently large transformer trained on next pixel prediction might eventually learn to generate diverse samples with clearly recognizable objects.**
* **Once it learns to do so, an idea known as “Analysis by Synthesis” suggests that the model will also know about object categories.**

**TOWARDS GENERAL UNSUPERVISED LEARNING**

* **Generative sequence modeling is a universal unsupervised learning algorithm: since all data types can be represented as sequences of bytes, a transformer can be directly applied to any data type without additional engineering.**
* **As a consequence of its generality, our method requires significantly more compute to achieve competitive performance in the unsupervised setting. Indeed, contrastive methodsare still the most computationally efficient methods for producing high quality features from images. However, in showing that an unsupervised transformer model is competitive with the best unsupervised convolutional nets, we provide evidence that it is possible to trade off hand coded domain knowledge for compute. In new domains, where there isn’t much knowledge to hand code, scaling compute seems an appropriate technique to test.**

**LIMITATIONS**

* **iGPT is capable of learning powerful image features, there are still significant limitations. Because the generic sequence transformer is used for GPT-2 in language, this V100-days while a similarly performing MoCo model can be trained in roughly 70 V100-days.**
* **Relatedly, it is modeled low resolution inputs using a transformer, while most self-supervised results use convolutional-based encoders which can easily consume inputs at high resolution. A new architecture, such as a domain-agnostic multiscale transformer, might be needed to scale further.**

**Given these limitations, this primarily serves as a proof-of-concept demonstration of the ability of large transformer-based language models to learn excellent unsupervised representations in novel domains, without the need for hardcoded domain knowledge. However, the significant resource cost to train these models and the greater accuracy of convolutional neural-network based methods precludes these representations from practical real-world applications in the vision domain.**

* **Finally, generative models can exhibit biases that are a consequence of the data they’ve been trained on. Many of these biases are useful, like assuming that a combination of brown and green pixels represents a branch covered in leaves, then using this bias to continue the image. But some of these biases will be harmful, when considered through a lens of fairness and representation. For instance, if the model develops a visual notion of a scientist that skews male, then it might consistently complete images of scientists with male-presenting people, rather than a mix of genders. We expect that developers will need to pay increasing attention to the data that they feed into their systems and to better understand how it relates to biases in trained models.**