November, 2021

Mega AI, an internal investment at Pacific Northwest National Laboratory, aims to develop next-generation artificial intelligence (AI) capabilities unique to the Department of Energy's national lab complex to address research gaps in large-scale multimodal representation learning, multitask inferences, and the need for increased generalizability, rapid adaptivity, and usability of AI technologies. In this newsletter, we highlight recent developments in the research community on next-generation AI technologies focusing on massive-scale model development, deployment and evaluation, data and code availability, model interactions, and new features and capabilities that are relevant to Mega AI's goals and science and security applications.

NEW MODELS AND CAPABILITIES

- **September 29 | Paper**: *MLIM: Vision-and-Language Model Pre-training with Masked Language and Image Modeling.* READ MORE
- **September 11 | Article**: *Stuck in GPT-3's waitlist? Try out the AI21 Jurassic-1*. READ MORE
- **September 1 | Paper**: *TABERT: Pretraining for Joint Understanding of Textual and Tabular Data*. READ MORE
- **August 31 | Paper**: Great arXiv post exploring using Codex to calculate the properties of molecules. READ MORE
- **August 19 | Paper**: Another language model (LM) for program synthesis. READ MORE
- **May 15 | Article**: *GPT-3's free Alternative GPT-Neo is Something to be Excited About.* READ MORE

NEW CODE FOR REPRODUCIBILITY

- **August 24 | GitHub**: New open-source code released from Stanford to train large-scale LMs e.g., GPT-2. READ MORE
- **GitHub**: Unified Toolkit for Deep Learning Based Document Image Analysis. READ MORE

NEW DATASETS FOR TRAINING AND EVALUATION

- **October 5 | Blog**: Google Research open-sourced a Wikipedia-based image-test dataset to train multimodal vision-language systems. READ MORE
- **August 30 | Paper**: *The Colossal Clean Crawled Corpus*. READ MORE
- **August 9 | Paper**: *The Pile, an 800Gb Dataset of Diverse Text for Language Modeling*. READ MORE

NEW REPORTS AND STUDIES

- **August 20 | Paper**: The most comprehensive review of large-scale models and their impact to date from Stanford; 200-page report. READ MORE
- **August 11 | Paper**: The most recent empirical study on studying scaling laws of language models. READ MORE

COMMUNITY DISCUSSION

- **October 8 | Article**: *DeepMind Presents Neural Algorithmic Reasoning: The Art of Fusing Neural Networks With Algorithmic Computation*. READ MORE
- **September 30 | Article**: New IEEE Spectrum article on the computational costs of deep learning. It's getting some traction on Twitter. READ MORE
- **August 30 | Video**: Credibility of science communication. WATCH
- **August 12 | Video**: More situational awareness on how computers can write code. WATCH

- **August 10 | Podcast**: Relevant discussion on parallelism and acceleration to train large-scale LMs. LISTEN
- **Blog**: AI and compute analysis from Open AI. READ MORE
- **Article**: *DeepMind's FIRE PBT: Automated Hyperparameter Tuning with Faster Model Training and Better Final Performance.* READ MORE
- **Article**: *Synced: Google Replaces BERT Self-Attention with Fourier Transform: 92% Accuracy, 7 Times Faster on GPUs*. READ MORE
- **Article**: *Game theory as an engine for large-scale data analysis*. READ MORE
- **Article**: *The Challenges of Applied Machine Learning*. READ MORE
- **Podcast**: "Chaos Orchestra" will explore "how knowledge graphs can be applied over the next decade to boost many areas of Artificial Intelligence and address the most pressing challenges of our times." LISTEN
- **Paper**: *Are Pre-trained Convolutions Better than Pre-trained Transformers?* READ MORE
- **Paper**: *Societal Biases in Language Generation: Progress and Challenges*. READ MORE
- **Paper**: *Tackling Climate Change with Machine Learning.* READ MORE

HIGHLIGHTED TECHNICAL RESOURCES

- **August 17 | Article**: See how transformers help solve compositional natural language processing (NLP) tasks. READ MORE
- **August 16 | Article**: How to train a BERT model from scratch. READ MORE
- **GitHub | Repo**: Annotated Transformers. READ MORE
- **Article**: *Time Series Anomaly Detection with PyCaret*. READ MORE
- **Paper**: *The NLP Cookbook: Modern Recipes for Transformer based Deep Learning Architectures.* READ MORE
- **Pypi Project**: NeuroX Toolkit for Interpretation and Analysis of Deep Neural Networks centered around Probing. READ MORE

TOWARDS AGI DISCUSSION

- **Article**: Interesting read about how creativity is born. READ MORE
- **Article**: *The Map of Artificial Intelligence*. READ MORE
- **GitHub | Repo**: Summary of NLP papers related to fairness collated in the awesome-fairness-papers repository. READ MORE

EVENTS

Recordings of the livestreams from the **Stanford Institute for Human-Centered Artificial Intelligence Workshop on Foundation Models** are available on YouTube.

- Day 1
- Day 2
- Agenda

For more information, contact:

Svitlana Volkova, svitlana.volkova@pnnl.gov
Mega AI Investment Lead

Maria Glenski, maria.glenski@pnnl.gov
Mega AI Deputy