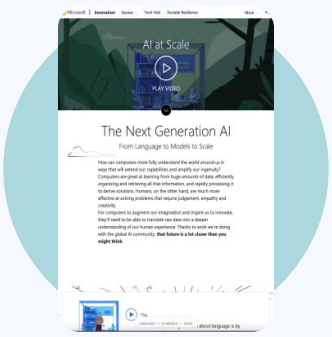

ARTICLE

The Next Generation AI

From Language to Models to Scale

Microsoft © 2021

Read the article

getAbstract Rating

★ 9 | Visionary | Overview | Analytical

What You Will Learn

- How Microsoft Turing sets new language modeling standards.
- Why deep learning accelerates AI language understanding.
- How AI models are customized for specific domains.
- Why collaboration fuels AI innovation and empowerment.

Supported Skills

Understand Large Language Models | AI Transformation | Understand AI

About the Author

Microsoft Corporation is an American multinational technology corporation that produces computer software, consumer electronics, personal computers, and related services.

Recommendation

Microsoft's "ALT Story" on AI at Scale provides a multimedia overview, with links to press releases and technical documents. It gives the history of AI milestones in the last five years, such as object and speech recognition and language translation. It explains how a new class of powerful large-scale language models such as Microsoft's Turing model is driving breakthroughs and accelerating AI innovation in its own products. Microsoft seeks to give customers access to state-of-the-art large-scale AI models, training optimization tools and supercomputing resources to accelerate their own innovation, what it calls "AI at Scale."

Take-Aways

- Microsoft seeks to enable the next generation of AI, starting with human language.
- As one form of machine learning, deep learning is developing rapidly.
- At 17 billion parameters, Microsoft Turing sets a new standard for language modelling tasks.
- Large-scale pre-trained models can be customized for specific domains.
- Making this breakthrough innovation available broadly will fuel AI innovation for customers.

Summary

Microsoft seeks to enable the next generation of AI, starting with human language.

Language is a fundamental aspect of what makes us human. Creating computers that are capable of augmenting human ingenuity requires a deep understanding of human language. But language is complex. As humans, it takes years to learn about the relationships between words and the importance of context. Meanwhile, language itself continues to evolve as new words appear and fresh nuances emerge across culture and societies.

> *"For computers to augment our imagination and inspire us to innovate, they'll need to be able to translate raw data into a deeper understanding of our human experience."*

Just like having access to diverse perspectives enriches our understanding of the world, training an AI model with the largest and most diverse data allows it to gain a richer understanding of language. So, Microsoft set out to enable the next generation of AI experiences by creating the largest, most powerful language model it could imagine.

As one form of machine learning, deep learning is developing rapidly.

The recent explosion of data provided by the internet, coupled with large amounts of fast computers facilitated by the cloud, has enabled the accelerated progress of deep learning. This subset of machine learning uses artificial neural networks to train a machine to perform a task.

> *"Through a sophisticated optimization process, the neural networks learn the strength of each of the connections between neurons by being trained with example data."*

Think of these artificial neural networks as layers of neurons interconnected to each other. Through a sophisticated optimization process, the neural networks learn the strength of each of the connections between neurons by being trained with example data. Through this process, the network is continuously measuring the accuracy of its predictions to update itself. This ultimately leads to a robust AI model of connected neurons. Each unique connection forms a parameter.

At 17 billion parameters, Microsoft Turing sets a new standard for language modeling tasks.

The Microsoft Turing model has established new benchmarks for a variety of practical language modeling tasks, such as summarization, contextual prediction and question answering. Its deep foundational grasp of human language enables it to derive the intended meaning and accurately respond to questions from a document or as real-time conversational agent.

At 17 billion parameters, Microsoft Turing sets a new standard. It establishes a foundation for ever-increasing parameters that researchers anticipate will continue into the trillions.

Large-scale AI models of this magnitude require massive infrastructure. Microsoft built this large-scale infrastructure on the Azure AI platform composed of huge clusters of thousands of graphical processing units (GPUs) for AI acceleration, interconnected with the latest high-bandwidth networks inside each server and across all servers. But raw infrastructure is not enough.

> *"We knew that to eventually scale to trillions of parameters, we'd need to solve some fundamental hardware and software limitations."*

No matter how powerful the infrastructure is, training a really large AI model requires partitioning the model into many layers and distributing these layers across multiple GPUs. Additionally, the huge amount of training data must be split into batches and trained in parallel across the cluster to produce the final model. Engineers knew that to eventually scale to trillions of parameters, they'd need to solve some fundamental hardware and software limitations.

Large-scale pre-trained models can be customized for specific domains.

The evolution of a foundational AI language model to one that exhibits expertise in a specific domain demands a specialized level of model adaptation. Microsoft starts with training a small number of very large AI models to be reused broadly across the company.

Microsoft trains these models with large amounts of unlabeled data, publicly available on the web, curated by Bing. This provides a continuously growing stream of high-quality, multi-topic data about the world to iteratively train richer semantic representations, in diverse human languages, into the same large model. Engineers start with all the foundational grammar and context skills the model has already mastered through self-supervised learning.

Then they elevate its understanding of specific, unique domains and tasks through a process called transfer learning. As the language model learns from domain data (for example, productivity domain from Office) it can do so within the context of each organization, in a privacy-compliant manner.

Eventually, every individual could be served by a personalized language model that could be fine-tuned by them in a private and secure way to serve their specific needs. By combining transfer learning with innovations in processing efficiency, Microsoft is continually elevating the intelligence of its AI language model in ways that will empower people to achieve more.

> *"By creating a common AI foundation that every team can work from, we can efficiently deliver high value, AI-driven user experiences at a massive scale, and with all the privacy and Responsible AI considerations our modern society requires."*

Microsoft uses these fine-tuned, domain-specific models across a number of its products, such as Bing, SharePoint, OneDrive, Outlook, and Dynamics 365.

Making this breakthrough innovation available broadly will fuel AI innovation for customers.

By working collaboratively and continuously – both internally and with the global AI community – Microsoft continues to introduce new AI capabilities into its products, its platform offerings to customers, its partnership endeavors, and the many Microsoft initiatives that support its mission of human empowerment.

> *"Together, we will build increasingly powerful models that will teach AI to more effectively understand our world and help amplify our own ingenuity."*

These collective AI efforts are resulting in breakthroughs destined to help the world solve some of its hardest challenges. Microsoft's aim is to build increasingly powerful models that will teach AI to more effectively understand the world and help amplify human ingenuity.

Did you like this summary?

Read the article

https://getab.li/44144

This document is restricted to personal use only.

getAbstract maintains complete editorial responsibility for all parts of this abstract. All rights reserved. No part of this abstract may be reproduced or transmitted or used for training any machine learning system in any form or by any means – electronic, photocopying or otherwise – without prior written permission of getAbstract AG (Switzerland).