Research

Home Blog

Google Research, 2022 & beyond: Language, vision and generative models

January 18, 2023

Posted by Jeff Dean, Senior Fellow and SVP of Google Research, on behalf of the Google Research community

Research

Today we kick off a series of blog posts about exciting new developments from Google Research. Please keep your eye on this space and look for the title "Google Research, 2022 & Beyond" for more articles in the series.

I've always been interested in computers because of their ability to help people better understand the world around them. Over the last decade, much of the research done at Google has been in pursuit of a similar vision — to help people better understand the world around them *and* get things done. We want to build more capable machines that partner with people to accomplish a huge variety of tasks. All kinds of tasks. Complex, information-seeking tasks. Creative tasks, like creating music, drawing new pictures, or creating videos. Analysis and synthesis tasks, like crafting new documents or emails from a few sentences of guidance, or partnering with people to jointly write software together. We want to solve complex mathematical or scientific problems. Transform modalities, or translate the world's information into any language. Diagnose complex diseases, or understand the physical world. Accomplish complex, multi-step actions in both the virtual software world and the physical world of robotics.

We've demonstrated early versions of some of these capabilities in research artifacts, and we've partnered with many teams across Google to ship some of these capabilities in Google products that touch the lives of billions of users. But the most exciting aspects of this journey still lie ahead!

With this post, I am kicking off a series in which researchers across Google will highlight some exciting progress we've made in 2022 and present our vision for 2023 and beyond. I will begin with a discussion of language, computer vision, multi-modal models, and generative machine learning models. Over the next several weeks, we will discuss novel developments in research topics ranging from responsible AI to algorithms and computer systems to science, health and robotics. Let's get started!

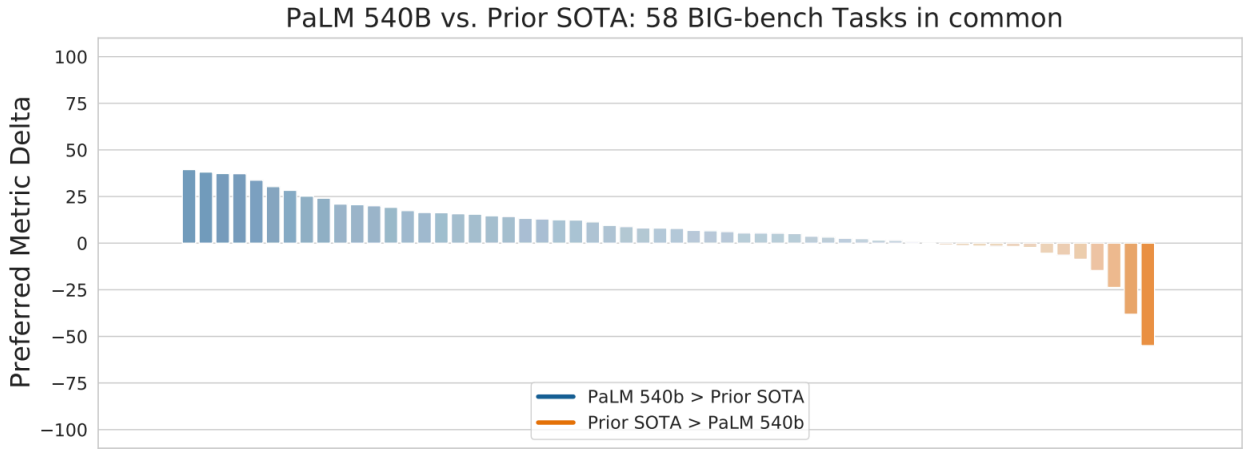
Language Models
Computer Vision
Multimodal Models
Generative Models
Responsible AI
ML & Computer Systems
Efficient Deep Learning
Algorithmic Advances
Robotics
Natural Sciences
Health
Community Engagement

Research

The progress on large and more powerful language models has been one of the most exciting areas of machine learning (ML) research over the last decade. Important advances along the way have included new approaches like sequence-to-sequence learning and our development of the Transformer model, which underlies most of the advances in this space in the last few years. Although language models are trained on surprisingly simple objectives, like predicting the next token in a sequence of text given the preceding tokens, when large models are trained on sufficiently large and diverse corpora of text, the models can generate coherent, contextual, natural-sounding responses, and can be used for a wide range of tasks, such as generating creative content, translating between languages, helping with coding tasks, and answering questions in a helpful and informative way. Our ongoing work on LaMDA explores how these models can be used for safe, grounded, and high-quality dialog to enable contextual multi-turn conversations.

Natural conversations are clearly an important and emergent way for people to interact with computers. Rather than contorting ourselves to interact in ways that best accommodate the limitations of computers, we can instead have natural conversations to accomplish a wide variety of tasks. I'm excited about the progress we've made in making LaMDA useful and factual.

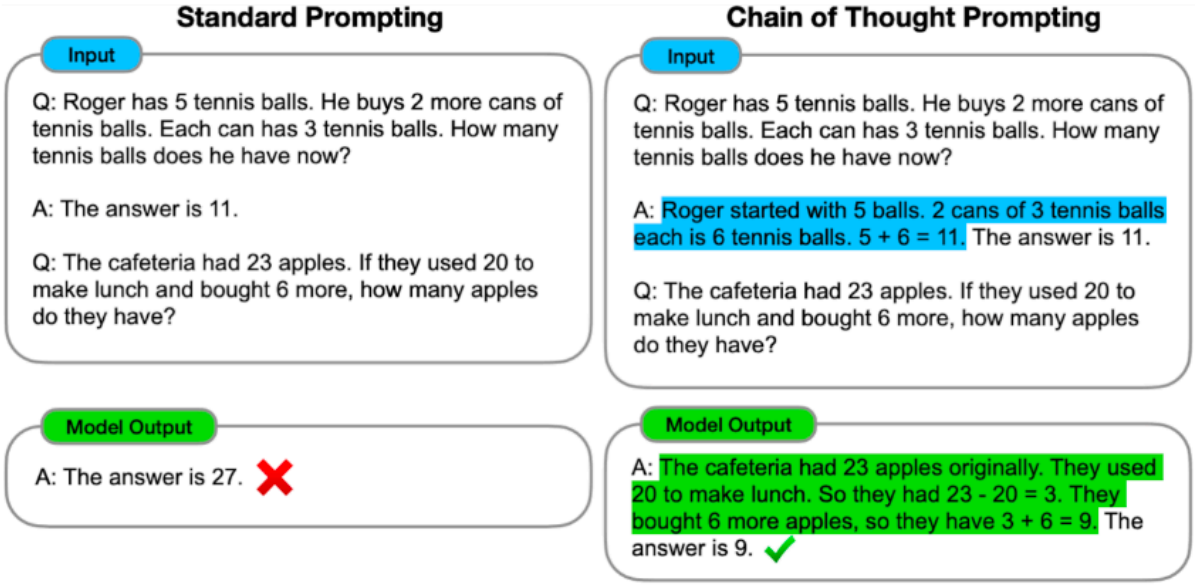
In April, we described our work on PaLM, a large, 540 billion parameter language model built using our Pathways software infrastructure and trained on multiple TPU v4 Pods. The PaLM work demonstrated that, despite being trained solely on the objective of predicting the next token, large-scale language models trained on large amounts of multi-lingual data and source code are capable of improving the state-of-the-art across a wide variety of natural language, translation, and coding tasks, despite never having been trained to specifically perform those tasks. This work provided additional evidence that increasing the scale of the model and training data can significantly improve capabilities.



Performance comparison between the PaLM 540B parameter model and the prior state-of-the-art (SOTA) on 58 tasks from the Big-bench suite. (See paper for details.)

Productivity. Using a variety of code completion suggestions from a 500 million parameter language model for a cohort of 10,000 Google software developers using this model in their IDE, we've seen that 2.6% of all code comes from suggestions generated by the model, reducing coding iteration time for these developers by 6%. We are working on enhanced versions of this and hope to roll it out to even more developers.

One of the broad key challenges in artificial intelligence is to build systems that can perform multi-step reasoning, learning to break down complex problems into smaller tasks and combining solutions to those to address the larger problem. Our recent work on Chain of Thought prompting, whereby the model is encouraged to "show its work" in solving new problems (similar to how your fourth-grade math teacher encouraged you to show the steps involved in solving a problem, rather than just writing down the answer you came up with), helps language models follow a logical chain of thought and generate more structured, organized and accurate responses. Like the fourth-grade math student that shows their work, not only does this make the problem-solving approach much more interpretable, it is also more likely that the correct answer will be found for complex problems that require multiple steps of reasoning.



Models that use standard prompting directly provide the answer to a multi-step reasoning problem. In contrast, chain of thought prompting teaches the model to deconstruct the problem into intermediate reasoning steps, better enabling it to reach the correct final answer.

One of the areas where multi-step reasoning is most clearly beneficial and measurable is in the ability of models to solve complex mathematical reasoning and scientific problems. A key research question is whether ML models can learn to solve complex problems using multi-step reasoning. By taking the general-purpose PaLM

substantial improvements over the state-of-the-art for mathematical reasoning and scientific problems across a wide variety of scientific and mathematical benchmark suites.

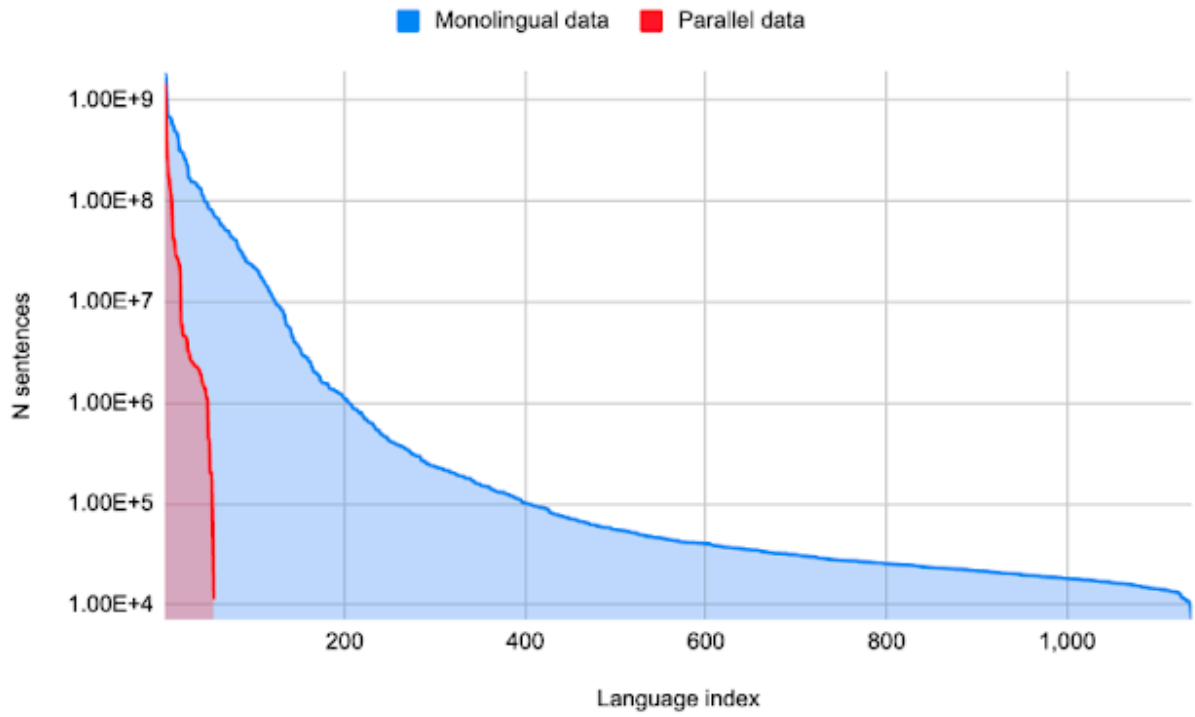
	MATH	MMLU-STEM	OCWCourses	GSM8k
Minerva	*50.3%*	*75%*	*30.8%*	*78.5%*
Published state-of-the-art	6.9%	55%	–	74.4%

Minerva 540B significantly improves state-of-the-art performance on STEM evaluation datasets.

Chain of Thought prompting is one way of better-expressing natural language prompts and examples to a model to improve its ability to tackle new tasks. The similar learned prompt tuning, in which a large language model is fine-tuned on a corpus of problem-domain–specific text, has shown great promise. In "Large Language Models Encode Clinical Knowledge", we demonstrated that learned prompt tuning can adapt a general-purpose language model to the medical domain with relatively few examples and that the resulting model can achieve 67.6% accuracy on US Medical License Exam questions (MedQA), surpassing the prior ML state-of-the-art by over 17%. While still short compared to the abilities of clinicians, comprehension, recall of knowledge and medical reasoning all improve with model scale and instruction prompt tuning, suggesting the potential utility of LLMs in medicine. Continued work can help to create safe, helpful language models for clinical application.

Large language models trained on multiple languages can also help with translation from one language to another, even when they have never been taught to explicitly translate text. Traditional machine translation systems usually rely on parallel (translated) text to learn to translate from one language to another. However, since parallel text exists for a relatively small number of languages, many languages are often not supported in machine translation systems. In "Unlocking Zero-Resource Machine Translation to Support New Languages in Google Translate" and the accompanying papers "Building Machine Translation Systems for the Next Thousand Languages" and "Towards the Next 1000 Languages in Multilingual Machine Translation: Exploring the Synergy Between Supervised and Self-Supervised Learning", we describe a set of techniques that use massively multilingual language

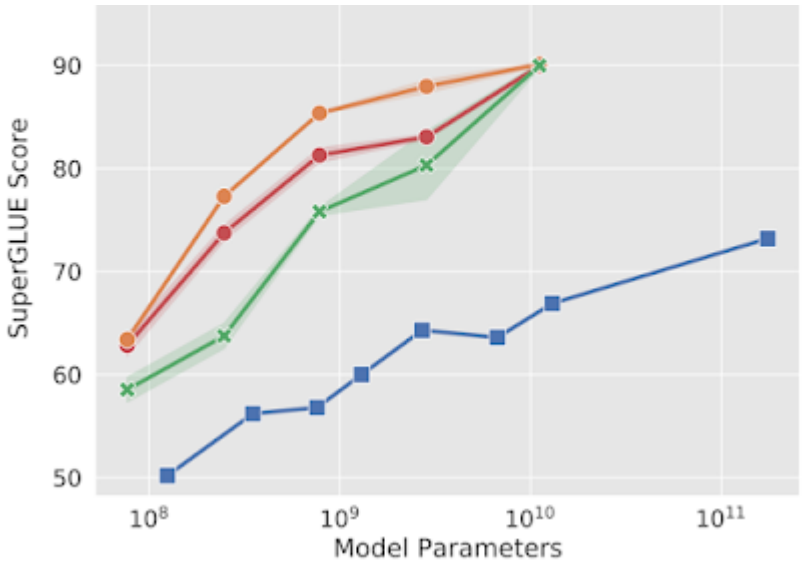
Research



The amount of monolingual data per language versus the amount of parallel (translated) data per language. A small number of languages have large amounts of parallel data, but there is a long tail of languages with only monolingual data.

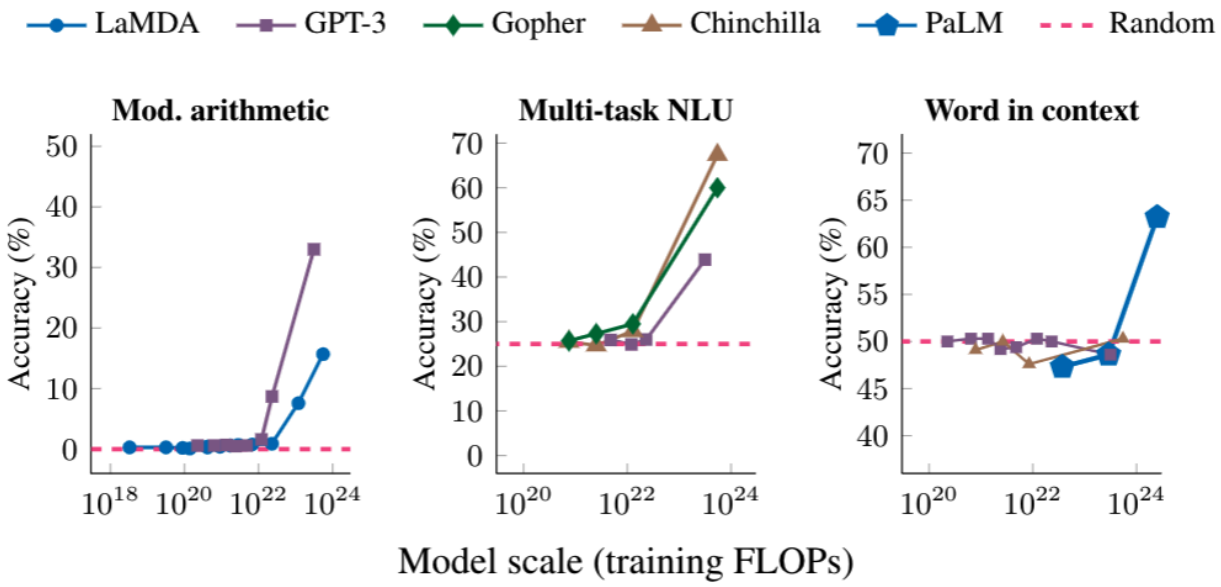
Another approach is represented with learned soft prompts, where instead of constructing new input tokens to represent a prompt, we add a small number of tunable parameters per task that can be learned from a few task examples. This approach generally yields high performance on tasks for which we have learned soft prompts, while allowing the large pre-trained language model to be shared across thousands of different tasks. This is a specific example of the more general technique of task adaptors, which allow a large portion of the parameters to be shared across tasks while still allowing task-specific adaptation and tuning.

Research



As scale increases, prompt tuning, which conditions frozen models using tunable soft prompts, matches the performance of model tuning, despite using 25,000 fewer parameters.

Interestingly, the utility of language models can grow significantly as their sizes increase due to the emergence of new capabilities. "Characterizing Emergent Phenomena in Large Language Models" examines the sometimes surprising characteristic that these models are not able to perform particular complex tasks very effectively until reaching a certain scale. But then, once a critical amount of learning has happened (which varies by task), they suddenly show large jumps in the ability to perform a complex task accurately (as shown below). This raises the question of what new tasks will become feasible when these models are trained further.



*The ability to perform multi-step arithmetic (**left**), succeed on college-level exams (**middle**), and identify the intended meaning of a word in context (**right**) all emerge only for models of sufficiently large scale. The models shown include LaMDA, GPT-3, Gopher, Chinchilla, and PaLM.*

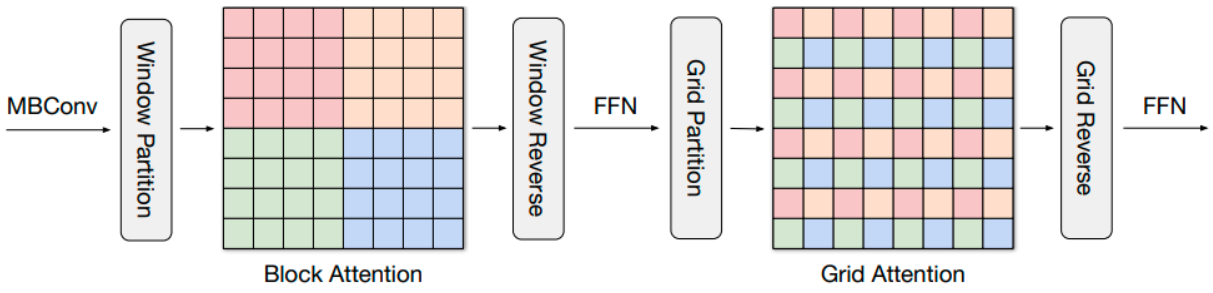
play an increasingly important role in many aspects of our lives.

Top

Computer Vision

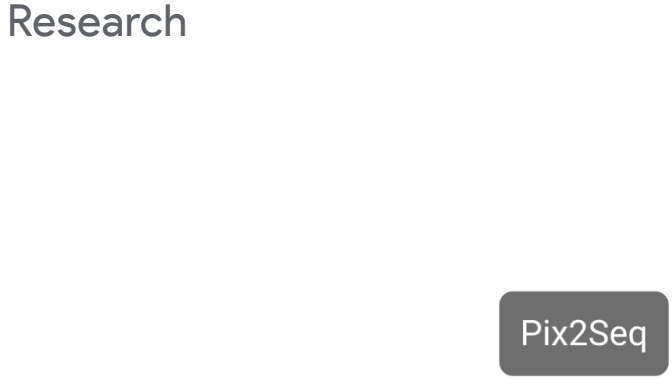
Computer vision continues to evolve and make rapid progress. One trend that started with our work on Vision Transformers in 2020 is to use the Transformer architecture in computer vision models rather than convolutional neural networks. Although the localized feature-building abstraction of convolutions is a strong approach for many computer vision problems, it is not as flexible as the general attention mechanism in transformers, which can utilize both local and non-local information about the image throughout the model. However, the full attention mechanism is challenging to apply to higher resolution images, since it scales quadratically with image size.

In "MaxViT: Multi-Axis Vision Transformer", we explore an approach that combines both local and non-local information at each stage of a vision model, but scales more efficiently than the full attention mechanism present in the original Vision Transformer work. This approach outperforms other state-of-the-art models on the ImageNet-1k classification task and various object detection tasks, but with significantly lower computational costs.



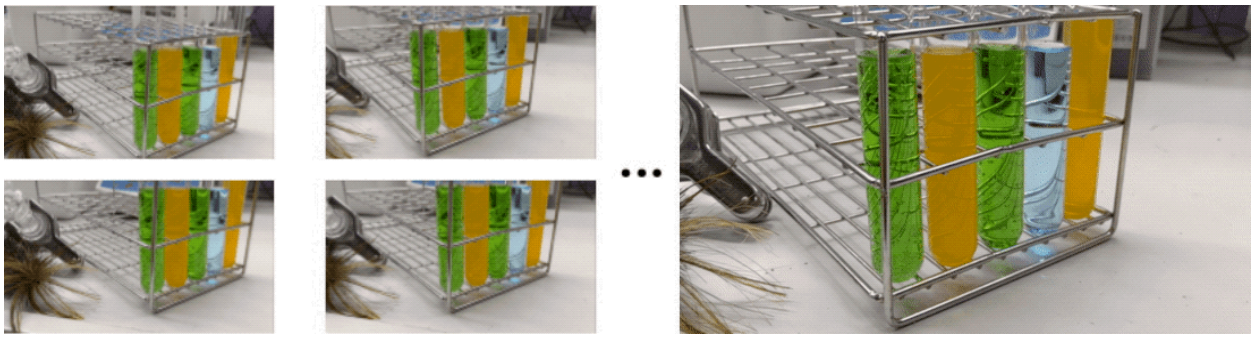
In MaxViT, a multi-axis attention mechanism conducts blocked local and dilated global attention sequentially followed by a FFN, with only a linear complexity. The pixels in the same colors are attended together.

In "Pix2Seq: A Language Modeling Framework for Object Detection", we explore a simple and generic method that tackles object detection from a completely different perspective. Unlike existing approaches that are task-specific, we cast object detection as a language modeling task conditioned on the observed pixel inputs with the model trained to "read out" the locations and other attributes about the objects of interest in the image. Pix2Seq achieves competitive results on the large-scale object detection COCO dataset compared to existing highly-specialized and well-



The Pix2Seq framework for object detection. The neural network perceives an image, and generates a sequence of tokens for each object, which correspond to bounding boxes and class labels.

Another long-standing challenge in computer vision is to better understand the 3-D structure of real-world objects from one or a few 2-D images. We have been trying multiple approaches to make progress in this area. In "Large Motion Frame Interpolation", we demonstrated that short slow-motion videos can be created by interpolating between two pictures that were taken many seconds apart, even when there might have been significant movement in some parts of the scene. In "View Synthesis with Transformers", we show how to combine two new techniques, light field neural rendering (LFNR) and generalizable patch-based neural rendering (GPNR), to synthesize novel views of a scene, a long-standing challenge in computer vision. LFNR is a technique that can accurately reproduce view-dependent effects by using transformers that learn to combine reference pixel colors. While LFNR works well on single scenes, its ability to generalize to novel scenes is limited. GPNR overcomes this by using a sequence of transformers with canonicalized positional encodings that can be trained on a set of scenes to synthesize views of new scenes. Together, these techniques enable high-quality view synthesis of novel scenes from just a couple of images of the scene, as shown below:



By combining LFNR and GPNR, models are able to produce new views of a scene given only a few images of it. These models are particularly effective when handling view-dependent effects

Going even further, in "LOLNerf: Learn from One Look", we explore the ability to learn a high quality representation from just a single 2-D image. By training on many different examples of particular categories of objects (e.g., lots of single images of different cats), we can learn enough about the expected 3-D structure of objects to create a 3-D model from just a single image of a novel category (e.g., just a single image of your cat, as shown in the LOLCats clips below).

Top: *Example cat images from AFHQ.* ***Bottom:*** *A synthesis of novel 3-D views created by LOLNeRF.*

A general thrust of this work is to develop techniques that help computers have a better understanding of the 3-D world — a longstanding dream of computer vision!

Top

Multimodal Models

Most past ML work has focused on models that deal with a single modality of data (e.g., language models, image classification models, or speech recognition models). While there has been plenty of amazing progress in these areas, the future is even more exciting as we look forward to multi-modal models that can flexibly handle many different modalities simultaneously, both as model inputs and as model outputs. We have pushed in this direction in many ways over the past year.

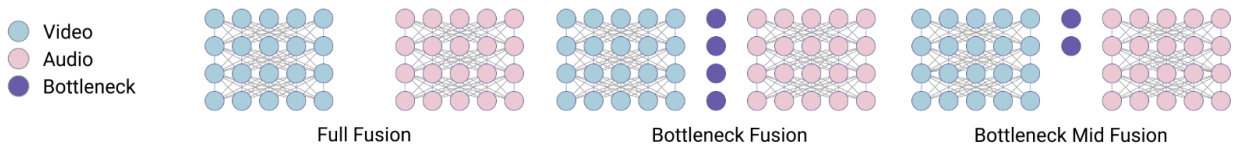
Research

Rather than relying on individual models tailored to specific tasks or domains, the next generation of multi-modal models can handle different modalities simultaneously by activating only the model pathways necessary for a given problem.

There are two key questions when building a multi-modal model that must be addressed to best enable cross-modality features and learning:

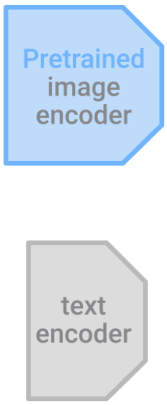
1. How much modality-specific processing should be done before allowing the learned representations to be merged?
2. What is the most effective way to mix the representations?

In our work on "Multi-modal Bottleneck Transformers" and the accompanying "Attention Bottlenecks for Multimodal Fusion" paper, we explore these tradeoffs and find that bringing together modalities after a few layers of modality-specific processing and then mixing the features from different modalities through a bottleneck layer is more effective than other techniques (as illustrated by the Bottleneck Mid Fusion in the figure below). This approach substantially improves accuracy on a variety of video classification tasks by learning to use multiple modalities of data to make classification decisions.



*Sample attention configurations for multi-modal transformer encoders. Red and blue rows of dots represent encoder layers. Typical approaches to fusion of multi-modal transformer encoder features ("full fusion") use pairwise self attention across hidden units in a layer (**left**). Bottleneck fusion (**middle**) restricts attention flow within a layer through tight latent units called attention bottlenecks. Bottleneck mid fusion (**right**) applies bottleneck fusion only to later layers in the model for optimal performance.*

improve image classification accuracy, even on unseen object categories. A modern variant of this general idea is found in Locked-image Tuning (LiT), a method that adds language understanding to an existing pre-trained image model. This approach contrastively trains a text encoder to match image representations from a powerful pre-trained image encoder. This simple method is data and compute efficient, and substantially improves zero-shot image classification performance compared to existing contrastive learning approaches.



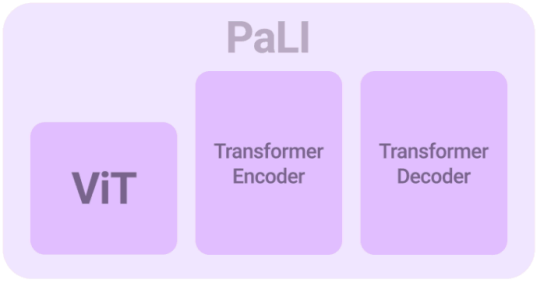
LiT-tuning contrastively trains a text encoder to match a pre-trained image encoder. The text encoder learns to compute representations that align to those from the image encoder.

Another example of the uni-modal utility of multi-modal models is observed when co-training on related modalities, like images and videos. In this case, one can often improve accuracy on video action classification tasks compared to training on video data alone (especially when training data in one modality is limited).

Combining language with other modalities is a natural step for improving how users interact with computers. We have explored this direction in quite a number of ways this year. One of the most exciting is in combining language and vision inputs, either still images or videos. In "PaLI: Scaling Language-Image Learning", we introduced a unified language-image model trained to perform many tasks in over 100 languages. These tasks span vision, language, and multimodal image and language applications, such as visual question answering, image captioning, object detection, image classification, optical character recognition, text reasoning, and others. By combining a vision transformer (ViT) with a text-based transformer encoder, and then a transformer-based decoder to generate textual answers, and training the whole system end-to-end on many different tasks simultaneously, the system achieves state-of-the-art results across many different benchmarks.

For example, PaLI achieves state-of-the-art results on the CrossModal-3600 benchmark, a diverse test of multilingual, multi-modal capabilities with an average

captioning and question answering, will lead to computer systems where you can have a natural conversation about other kinds of sensory inputs, asking questions and getting answers to your needs in a wide variety of languages (*"In Thai, can you say what is above the table in this image?"*, *"How many parakeets do you see sitting on the branches?"*, *"Describe this image in Swahili"*, *"What Hindi text is in this image?"*).



The PaLI model addresses a wide range of tasks in the language-image, language-only and image-only domain using the same API (e.g., visual-question answering, image captioning, scene-text understanding, etc.). The model is trained to support over 100 languages and tuned to perform multilingually for multiple language-image tasks.

In a similar vein, our work on FindIt enables natural language questions about visual images to be answered through a unified, general-purpose and multitask visual grounding model that can flexibly answer different types of grounding and detection queries.

*FindIt is a unified model for referring expression comprehension (**first column**), text-based localization (**second**), and the object detection task (**third**). FindIt can respond accurately when tested on object types and classes not known during training, e.g., "Find the desk" (**fourth**). We show the MattNet results for comparison.*

The area of video question answering (e.g., given a baking video, being able to answer a question like *"What is the second ingredient poured into the bowl?"*)

versions of the same video input (e.g., a high resolution, low frame-rate video and a low resolution, high frame-rate video), are efficiently fused together with the text input to produce a text-based answer by the decoder. Instead of processing the inputs directly, the video-text iterative co-tokenization model learns a reduced number of useful tokens from the fused video-language inputs. This process is done iteratively, allowing the current feature tokenization to affect the selection of tokens at the next iteration, thus refining the selection.

An example input question for the video question answering task "What is the second ingredient poured into the bowl?" which requires deeper understanding of both the visual and text inputs. The video is an example from the 50 Salads dataset, used under the Creative Commons license.

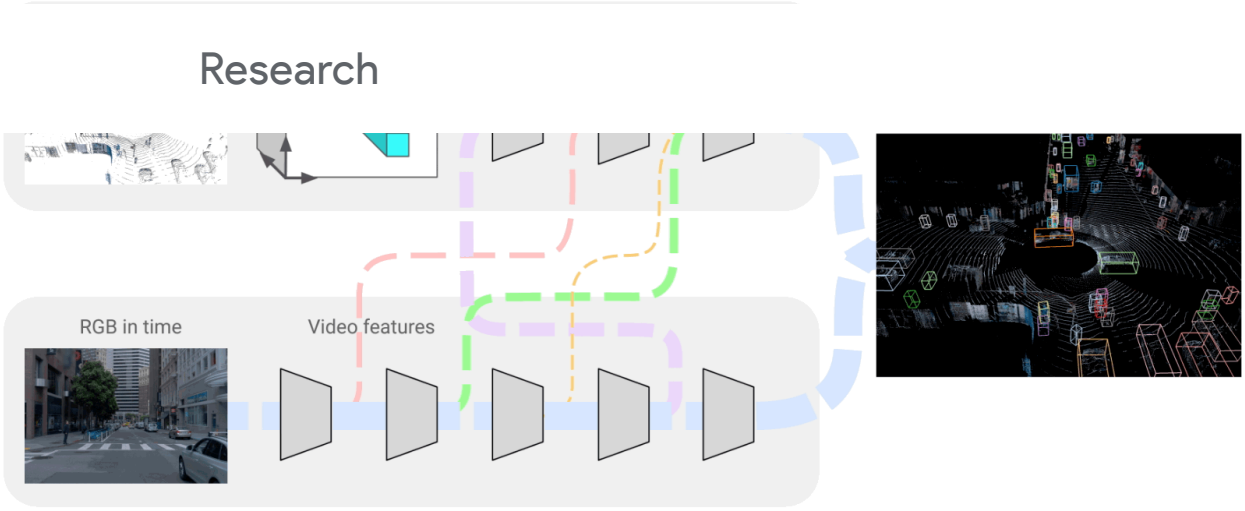
The process of creating high-quality video content often includes several stages, from video capturing to video and audio editing. In some cases, dialogue is re-recorded in a studio (referred to as dialog replacement, post-sync or dubbing) to achieve high quality and replace original audio that might have been recorded in noisy or other suboptimal conditions. However, the dialog replacement process can be difficult and tedious because the newly recorded audio needs to be well synced with the video, often requiring several edits to match the exact timing of mouth movements. In "VDTTS: Visually-Driven Text-To-Speech", we explore a multi-modal model for accomplishing this task more easily. Given desired text and the original video frames of a speaker, the model can generate speech output of the text that matches the video while also recovering aspects of prosody, such as timing or emotion. The system shows substantial improvements on a variety of metrics related to video-sync, speech quality, and speech pitch. Interestingly, the model can produce video-synchronized speech without any explicit constraints or losses in the model training to promote this.

Research

Original *displays the original video clip.* ***VDTTS*** *displays the audio predicted using both the video frames and the text as input.* ***VDTTS video-only*** *displays audio predictions using video frames only.* ***TTS*** *displays audio predictions using text only.* ***Transcript****: "absolutely love dancing I have no dance experience whatsoever but as that".*

In "Look and Talk: Natural Conversations with Google Assistant", we show how an on-device multi-modal model can use both video and audio input to make interacting with Google Assistant much more natural. The model learns to use a number of visual and auditory cues, such as gaze direction, proximity, face matching, voice matching and intent classification, to more accurately determine if a nearby person is actually trying to talk to the Google Assistant device, or merely happens to be talking near the device without the intent of causing the device to take any action. With just the audio or visual features alone, this determination would be much more difficult.

Multi-modal models don't have to be limited to just combining human-oriented modalities like natural language or imagery, and they are increasingly important for real-world autonomous vehicle and robotics applications. In this context, such models can take the raw output of sensors that are unlike any human senses, such as 3-D point cloud data from Lidar units on autonomous vehicles, and can combine this with data from other sensors, like vehicle cameras, to better understand the environment around them and to make better decisions. In "4D-Net for Learning Multi-Modal Alignment for 3D and Image Inputs in Time", the 3-D point cloud data from Lidar is fused with the RGB data from the camera in real-time, with a self-attention mechanism controlling how the features are mixed together and weighted at different layers. The combination of the different modalities and the use of time-oriented features gives substantially improved accuracy in 3-D object recognition over using either modality on its own. More recent work on Lidar-camera fusion introduced learnable alignment and better geometric processing through inverse augmentation to further improve the accuracy of 3-D object recognition.



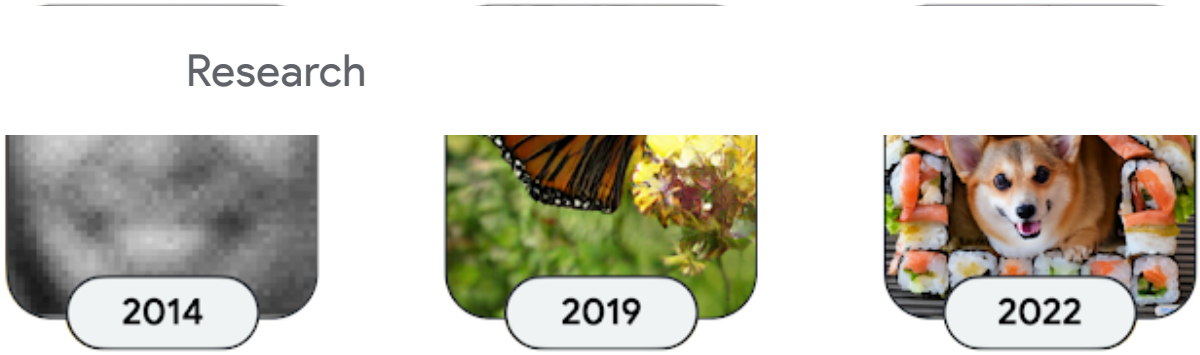
4D-Net effectively combines 3D LiDAR point clouds in time with RGB images, also streamed in time as video, learning the connections between different sensors and their feature representations.

Having single models that understand many different modalities fluidly and contextually and that can generate many different kinds of outputs (e.g., language, images or speech) in that context, is a much more useful, general purpose framing of ML. We're excited about where this will take us because it will enable new exciting applications in many Google products and also advance the fields of health, science, creativity, robotics and more!

Top

Generative Models

The quality and capabilities of generative models for imagery, video, and audio has shown truly stunning and extraordinary advances in 2022. There are a wide variety of approaches for generative models, which must learn to model complex data sets (e.g., natural images). Generative adversarial networks, developed in 2014, set up two models working against each other. One is a *generator*, which tries to generate a realistic looking image (perhaps conditioned on an input to the model, like the category of image to generate), and the other is a *discriminator*, which is given the generated image and a real image and tries to determine which of the two is generated and which is real, hence the *adversarial* aspect. Each model is trying to get better and better at winning the competition against the other, resulting in both models getting better and better at their task, and in the end, the generative model can be used in isolation to generate images.



Advances in generative image model capabilities over the past decade.
Left: *From I. Goodfellow, et al. 2014.* ***Middle:*** *From M. Lucic, et al. 2019.* ***Right:*** *From Imagen.*

Diffusion models, introduced in "Deep Unsupervised Learning using Nonequilibrium Thermodynamics" in 2015, systematically and slowly destroy structure in a data distribution through an iterative forward diffusion process. They then learn a reverse diffusion process that can restore the structure in the data that has been lost, even given high levels of noise. The forward process can be used to generate noisy starting points for the reverse diffusion process conditioned on various useful, controllable inputs to the model, so that the reverse diffusion (generative) process becomes controllable. This means that it is possible to ask the model to "*generate an image of a grapefruit*", a much more useful capability than just "*generate an image*" if what you are after is indeed a sampling of images of grapefruits.

Various forms of autoregressive models have also been applied to the task of image generation. In 2016, "Pixel Recurrent Neural Networks" introduced PixelRNN, a recurrent architecture, and PixelCNN, a similar but more efficient convolutional architecture that was also investigated in "Conditional Image Generation with PixelCNN Decoders". These two architectures helped lay the foundation for pixel-level generation using deep neural networks. They were followed in 2017 by VQ-VAE, proposed in "Neural Discrete Representation Learning", a vector-quantized variational autoencoder. Combining this with PixelCNN yielded high-quality images. Then, in 2018 Image Transformer used the autoregressive Transformer model to generate images.

Until relatively recently, all of these image generation techniques were capable of generating images that are relatively low quality compared to real world images. However, several recent advances have opened the door for much better image generation performance. One is Contrastic Language-Image Pre-training (CLIP), a

representation and yielded good zero-shot performance on datasets like ImageNet.

In addition to CLIP, the toolkit of generative image models has recently grown. Large language model encoders have been shown to effectively condition image generation on long natural language descriptions rather than just a limited number of pre-set categories of images. Significantly larger training datasets of images and accompanying captions (which can be reversed to serve as *text→image* exemplars) have improved overall performance. All of these factors together have given rise to a range of models able to generate high-resolution images with strong adherence even to very detailed and fantastic prompts.

We focus here on two recent advances from teams in Google Research, Imagen and Parti.

Imagen is based on the Diffusion work discussed above. In their 2022 paper "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding", the authors show that a generic large language model (e.g., T5), pre-trained on text-only corpora, is surprisingly effective at encoding text for image synthesis. Somewhat surprisingly, increasing the size of the language model in Imagen boosts both sample fidelity and image-text alignment much more than increasing the size of the image diffusion model. The work offers several advances to Diffusion-based image generation, including a new memory-efficient architecture called Efficient U-Net and Classifier-Free Diffusion Guidance, which improves performance by occasionally "dropping out" conditioning information during training. Classifier-free guidance forces the model to learn to generate from the input data alone, thus helping it avoid problems that arise from over-relying on the conditioning information. "Guidance: a cheat code for diffusion models" provides a nice explanation.

Parti uses an autoregressive Transformer architecture to generate image pixels based on a text input. In "Vector-quantized Image Modeling with Improved VQGAN", released in 2021, an encoder based on Vision Transformer is shown to significantly improve the output of a vector-quantized GAN model, VQGAN. This is extended in "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation", released in 2022, where much better results are obtained by scaling the Transformer encoder-decoder to 20B parameters. Parti also uses classifier-free guidance, described above, to sharpen the generated images. Perhaps not surprising given that it is a language model, Parti is particularly good at picking up on subtle cues in the prompt.

Research

Left: *Imagen generated image from the complex prompt, "A wall in a royal castle. There are two paintings on the wall. The one on the left is a detailed oil painting of the royal raccoon king. The one on the right a detailed oil painting of the royal raccoon queen."*
Right: *Parti generated image from the prompt, "A teddy bear wearing a motorcycle helmet and cape car surfing on a taxi cab in New York City. dslr photo."*

User Control

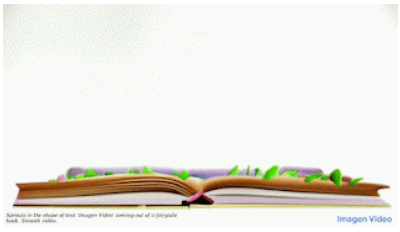
The advances described above make it possible to generate realistic still images based on text descriptions. However, sometimes text alone is not sufficient to enable you to create what you want — e.g., consider "*A dog being chased by a unicorn on the beach*" vs. "***My*** *dog being chased by a unicorn on the beach*". So, we have done subsequent research in providing new ways for users to control the generation process. In "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation", users are able to fine-tune a trained model like Imagen or Parti to generate new images based on a combination of text and user-furnished images (as illustrated below and with more details and examples on the DreamBooth site). This allows users to place images of themselves (or e.g., their pets) into generated images, thus allowing for much more user control. This is exemplified in "Prompt-to-Prompt Image Editing with Cross Attention Control", where users are able to edit images using text prompts like "make the car into a bicycle" and in Imagen Editor, which allows users to iteratively edit images by filling in masked areas using text prompts.

Research

Generative Video

One of the next research challenges we are tackling is to create generative models for video that can produce high resolution, high quality, temporally consistent videos with a high level of controllability. This is a very challenging area because unlike images, where the challenge was to match the desired properties of the image with the generated pixels, with video there is the added dimension of time. Not only must all the pixels in each frame match what should be happening in the video at the moment, they must also be consistent with other frames, both at a very fine-grained level (a few frames away, so that motion looks smooth and natural), but also at a coarse-grained level (if we asked for a two minute video of a plane taking off, circling, and landing, we must make thousands of frames that are consistent with this high-level video objective). This year we've made quite a lot of exciting progress on this lofty goal through two efforts, Imagen Video and Phenaki, each using somewhat different approaches.

Imagen Video generates high resolution videos with Cascaded Diffusion Models (described in more detail in "Imagen Video: High Definition Video Generation from Diffusion Models"). The first step is to take an input text prompt ("*A happy elephant wearing a birthday hat walking under the sea*") and encode it into textual embeddings with a T5 text encoder. A base video diffusion model then generates a very rough sketch 16 frame video at 40×24 resolution and 3 frames per second. This is then followed by multiple temporal super-resolution (TSR) and spatial super-resolution (SSR) models to upsample and generate a final 128 frame video at 1280×768 resolution and 24 frames per second — resulting in 5.3s of high definition video. The resulting videos are high resolution, and are spatially and temporally consistent, but still quite short at ~5 seconds long.



"Phenaki: Variable Length Video Generation From Open Domain Textual Description", released in 2022, introduces a new Transformer-based model for learning video representations, which compresses the video to a small representation of discrete tokens. Text conditioning is achieved by training a bi-directional Transformer model to generate video tokens based on a text description. These generated video tokens are then decoded to create the actual video. Because the model is causal in time, it

Research

Phenaki video generated from the complex prompt, "A photorealistic teddy bear is swimming in the ocean at San Francisco. The teddy bear goes under water. The teddy bear keeps swimming under the water with colorful fishes. A panda bear is swimming under water."

It is possible to combine the Imagen Video and Phenaki models to benefit from both the high-resolution individual frames from Imagen and the long-form videos from Phenaki. The most straightforward way to do this is to use Imagen Video to handle super-resolution of short video segments, while relying on the auto-regressive Phenaki model to generate the long-timescale video information.

Generative Audio

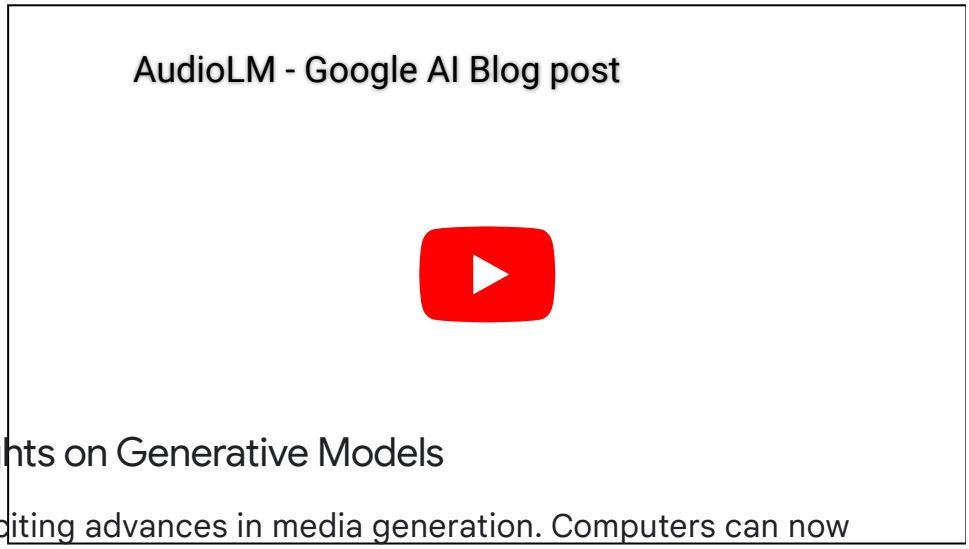
In addition to visual-oriented generative models, we have made significant progress on generative models for audio. In "AudioLM, a Language Modeling Approach to Audio Generation" (and the accompanying paper), we describe how to leverage advances in language modeling to generate audio without being trained on annotated data. Using a language-modeling approach for raw audio data instead of textual data introduces a number of challenges that need to be addressed.

First, the data rate for audio is significantly higher, leading to much longer sequences — while a written sentence can be represented by a few dozen characters, its audio waveform typically contains hundreds of thousands of values. Second, there is a one-to-many relationship between text and audio. This means that the same sentence can be uttered differently by different speakers with different speaking styles, emotional content and other audio background conditions.

To deal with this, we separate the audio generation process into two steps. The first involves a sequence of coarse, semantic tokens that capture both local

modeling long sequences. One part of the model generates a sequence of coarse semantic tokens conditioned on the past sequence of such tokens. We then rely on a portion of the model that can use a sequence of coarse tokens to generate fine-grained audio tokens that are close to the final generated waveform.

When trained on speech, and without any transcript or annotation, AudioLM generates syntactically and semantically plausible speech continuations while also maintaining speaker identity and prosody for unseen speakers. AudioLM can also be used to generate coherent piano music continuations, despite being trained without any symbolic representation of music. You can listen to more samples here.



Concluding Thoughts on Generative Models

2022 has brought exciting advances in media generation. Computers can now interact with natural language and better understand your creative process and what you might want to create. This unlocks exciting new ways for computers to help users create images, video, and audio — in ways that surpass the limits of traditional tools!

This has inspired more research interest in how users can control the generative process. Advances in text-to-image and text-to-video have unlocked language as a powerful way to control generation, while work like Dream Booth has made it possible for users to kickstart the generative process with their own images. 2023 and beyond will surely be marked by advances in the quality and speed of media generation itself. Alongside these advances, we will also see new user experiences, allowing for more creative expression.

It is also worth noting that although these creative tools have tremendous possibilities for helping humans with creative tasks, they introduce a number of concerns — they could potentially generate harmful content of various kinds, or generate fake imagery or audio content that is difficult to distinguish from reality. These are all issues we consider carefully when deciding when and how to deploy these models responsibly.

Research

Responsible AI

AI must be pursued responsibly. Powerful language models can help people with many tasks, but without care they can also generate misinformation or toxic text. Generative models can be used for amazing creative purposes, enabling people to manifest their imagination in new and amazing ways, but they can also be used to create harmful imagery or realistic-looking images of events that never occurred.

These are complex topics to grapple with. Leaders in ML and AI must lead not only in state-of-the-art technologies, but also in state-of-the-art approaches to responsibility and implementation. In 2018, we were one of the first companies to articulate AI Principles that put beneficial use, users, safety, and avoidance of harms above all, and we have pioneered many best practices, like the use of model and data cards. More than words on paper, we apply our AI Principles in practice. You can see our latest AI Principles progress update here, including case studies on text-to-image generation models, techniques for avoiding gender bias in translations, and more inclusive and equitable evaluation skin tones. Similar updates were published in 2021, 2020, and 2019. As we pursue AI both boldly and responsibly, we continue to learn from users, other researchers, affected communities, and our experiences.

Our responsible AI approach includes the following:

- Focus on AI that is useful and benefits users and society.
- Intentionally apply our AI Principles (which are grounded in beneficial uses and avoidance of harm), processes, and governance to guide our work in AI, from research priorities to productization and uses.
- Apply the scientific method to AI R&D with research rigor, peer review, readiness reviews, and responsible approaches to access and externalization.
- Collaborate with multidisciplinary experts, including social scientists, ethicists, and other teams with socio-technical expertise.
- Listen, learn and improve based on feedback from developers, users, governments, and representatives of affected communities.
- Conduct regular reviews of our AI research and application development, including use cases. Provide transparency on what we've learned.
- Stay on top of current and evolving areas of concern and risk (e.g., safety, bias and toxicity) and address, research and innovate to respond to challenges and risks as they emerge.
- Lead on and help shape responsible governance, accountability, and regulation that encourages innovation and maximizes the benefits of AI while mitigating risks.
- Help users and society understand what AI is (and is not) and how to benefit from its potential.

Research

Concluding Thoughts

We're excited by the transformational advances discussed above, many of which we're applying to make Google products more helpful to billions of users — including Search, Assistant, Ads, Cloud, Gmail, Maps, YouTube, Workspace, Android, Pixel, Nest, and Translate. These latest advances are making their way into real user experiences that will dramatically change how we interact with computers.

In the domain of language models, thanks to our invention of the Transformer model and advances like sequence-to-sequence learning, people can have a natural conversation (with a computer!) — and get surprisingly good responses (from a computer!). Thanks to new approaches in computer vision, computers can help people create and interact in 3D, rather than 2D. And thanks to new advances in generative models, computers can help people create images, videos, and audio — in ways they weren't able to before with traditional tools (e.g., a keyboard and mouse). Combined with advances like natural language understanding, computers can understand what you're trying to create — and help you realize surprisingly good results!

Another transformation changing how people interact with computers is the increasing capabilities of multi-modal models. We are working towards being able to create a single model that can understand many different modalities fluidly — understanding what each modality represents in context — and then actually generate different modes in that context. We're excited by progress towards this goal! For example, we introduced a unified language model that can perform vision, language, question answering and object detection tasks in over 100 languages with state-of-the-art results across various benchmarks. In future applications, people can engage more senses to get computers to do what they want — e.g., "Describe this image in Swahili." We've shown that on-device multi-modal models can make interacting with Google Assistant more natural. And we've demonstrated models that can, in various combinations, generate images, video, and audio controlled by natural language, images, and audio. More exciting things to come in this space!

As we innovate, we have a responsibility to users and society to thoughtfully pursue and develop these new technologies in accordance with our AI Principles. It's not enough for us to develop state-of-the-art technologies, but we must also ensure that they are safe before broadly releasing them into the world, and we take this responsibility very seriously.

New advances in AI present an exciting horizon of new ways computers can help people get things done. For Google, many will enhance or transform our

enhance and transform user experiences — helping more people better understand the world around them and get more things done. My own longstanding vision of computers!

Acknowledgements

Thank you to the entire Research Community at Google for their contributions to this work! In addition, I would especially like to thank the many Googlers who provided helpful feedback in the writing of this post and who will be contributing to the other posts in this series, including Martin Abadi, Ryan Babbush, Vivek Bandyopadhyay, Kendra Byrne, Esmeralda Cardenas, Alison Carroll, Zhifeng Chen, Charina Chou, Lucy Colwell, Greg Corrado, Corinna Cortes, Marian Croak, Tulsee Doshi, Toju Duke, Doug Eck, Sepi Hejazi Moghadam, Pritish Kamath, Julian Kelly, Sanjiv Kumar, Ronit Levavi Morad, Pasin Manurangsi, Yossi Matias, Kathy Meier-Hellstern, Vahab Mirrokni, Hartmut Neven, Adam Paszke, David Patterson, Mangpo Phothilimthana, John Platt, Ben Poole, Tom Small, Vadim Smelyanskiy, Vincent Vanhoucke, and Leslie Yeh.

Labels:

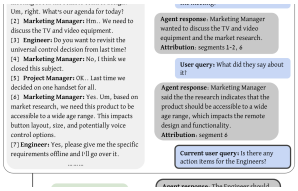
Machine Intelligence

Machine Perception

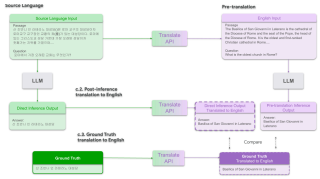
Natural Language Processing

Year in Review

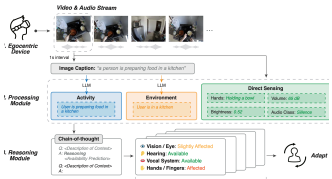
Other posts of interest



JUNE 25, 2024



JUNE 14, 2024



JUNE 14, 2024

Google Research

source-grounded information-seeking dialogs: A use case for meeting transcripts

Machine Intelligence · Natural Language Processing · Open Source Models & Datasets

in multilingual LLM applications

Machine Intelligence · Machine Translation · Natural Language Processing

situational impairments with large language models

Generative AI · Human-Computer Interaction and Visualization · Natural Language Processing

Follow us