CHAPTER 2:

Technical Performance

Artificial Intelligence
Index Report 2021

CHAPTER 2:

Chapter Preview

ACCESS THE PUBLIC DATA

Overview

This chapter highlights the technical progress in various subfields of AI, including computer vision, language, speech, concept learning, and theorem proving. It uses a combination of quantitative measurements, such as common benchmarks and prize challenges, and qualitative insights from academic papers to showcase the developments in state-of-the-art AI technologies.

While technological advances allow AI systems to be deployed more widely and easily than ever, concerns about the use of AI are also growing, particularly when it comes to issues such as algorithmic bias. The emergence of new AI capabilities such as being able to synthesize images and videos also poses ethical challenges.

CHAPTER HIGHLIGHTS

- **Generative everything:** AI systems can now compose text, audio, and images to a sufficiently high standard that humans have a hard time telling the difference between synthetic and non-synthetic outputs for some constrained applications of the technology. That promises to generate a tremendous range of downstream applications of AI for both socially useful and less useful purposes. It is also causing researchers to invest in technologies for detecting generative models; the DeepFake Detection Challenge data indicates how well computers can distinguish between different outputs.

- **The industrialization of computer vision:** Computer vision has seen immense progress in the past decade, primarily due to the use of machine learning techniques (specifically deep learning). New data shows that computer vision is industrializing: Performance is starting to flatten on some of the largest benchmarks, suggesting that the community needs to develop and agree on harder ones that further test performance. Meanwhile, companies are investing increasingly large amounts of computational resources to train computer vision systems at a faster rate than ever before. Meanwhile, technologies for use in deployed systems—like object-detection frameworks for analysis of still frames from videos—are maturing rapidly, indicating further AI deployment.

- **Natural Language Processing (NLP) outruns its evaluation metrics:** Rapid progress in NLP has yielded AI systems with significantly improved language capabilities that have started to have a meaningful economic impact on the world. Google and Microsoft have both deployed the BERT language model into their search engines, while other large language models have been developed by companies ranging from Microsoft to OpenAI. Progress in NLP has been so swift that technical advances have started to outpace the benchmarks to test for them. This can be seen in the rapid emergence of systems that obtain human level performance on SuperGLUE, an NLP evaluation suite developed in response to earlier NLP progress overshooting the capabilities being assessed by GLUE.

- **New analyses on reasoning:** Most measures of technical problems show for each time point the performance of the best system at that time on a fixed benchmark. New analyses developed for the AI Index offer metrics that allow for an evolving benchmark, and for the attribution to individual systems of credit for a share of the overall performance of a group of systems over time. These are applied to two symbolic reasoning problems, Automated Theorem Proving and Satisfiability of Boolean formulas.

- **Machine learning is changing the game in healthcare and biology:** The landscape of the healthcare and biology industries has evolved substantially with the adoption of machine learning. DeepMind’s AlphaFold applied deep learning technique to make a significant breakthrough in the decades-long biology challenge of protein folding. Scientists use ML models to learn representations of chemical molecules for more effective chemical synthesis planning. PostEra, an AI startup used ML-based techniques to accelerate COVID-related drug discovery during the pandemic.

Computer Vision

Introduced in the 1960s, the field of computer vision has seen significant progress and in recent years has started to reach human levels of performance on some restricted visual tasks. Common computer vision tasks include object recognition, pose estimation, and semantic segmentation. The maturation of computer vision technology has unlocked a range of applications: self-driving cars, medical image analysis, consumer applications (e.g., Google Photos), security applications (e.g., surveillance, satellite imagery analysis), industrial applications (e.g., detecting defective parts in manufacturing and assembly), and others.

2.1 COMPUTER VISION—IMAGE

IMAGE CLASSIFICATION

In the 2010s, the field of image recognition and classification began to switch from classical AI techniques to ones based on machine learning and, specifically, deep learning. Since then, image recognition has shifted from being an expensive, domain-specific technology to being one that is more affordable and applicable to more areas—primarily due to advancements in the underlying technology (algorithms, compute hardware, and the utilization of larger datasets).

ImageNet

Created by computer scientists from Stanford University and Princeton University in 2009, ImageNet is a dataset of over 14 million images across 200 classes that expands and improves the data available for researchers to train AI algorithms. In 2012, researchers from the University of Toronto used techniques based on deep learning to set a new state of the art in the ImageNet Large Scale Visual Recognition Challenge.

Since then, deep learning techniques have ruled the competition leaderboards—several widely used techniques have debuted in ImageNet competition entries. In 2015, a team from Microsoft Research said it had surpassed human-level performance on the image classification task[1] via the use of "residual networks"—an innovation that subsequently proliferated into other AI systems. Even after the end of the competition in 2017, researchers continue to use the ImageNet dataset to test and develop computer vision applications.

The image classification task of the ImageNet Challenge asks machines to assign a class label to an image based on the main object in the image. The following graphs explore the evolution of the top-performing ImageNet systems over time, as well as how algorithmic and infrastructure advances have allowed researchers to increase the efficiency of training image recognition systems and reduce the absolute time it takes to train high-performing ones.

Image recognition has shifted from being an expensive, domain-specific technology to being one that is more affordable and applicable to more areas—primarily due to advancements in the underlying technology.

ImageNet: Top-1 Accuracy

Top-1 accuracy tests for how well an AI system can assign the correct label to an image, specifically whether its single most highly probable prediction (out of all possible labels) is the same as the target label. In recent years, researchers have started to focus on improving performance on ImageNet by pre-training their systems on extra training data, for instance photos from Instagram or other social media sources. By pre-training on these datasets, they're able to more effectively use ImageNet data, which further improves performance. Figure 2.1.1 shows that recent systems with extra training data make 1 error out of every 10 tries on top-1 accuracy, versus 4 errors out of every 10 tries in December 2012. The model from the Google Brain team achieved 90.2% on top-1 accuracy in January 2021.

[1] Though it is worth noting that the human baseline for this metric comes from a single Stanford graduate student who took roughly the same test as the AI systems took.

IMAGENET CHALLENGE: TOP-1 ACCURACY

Source: Papers with Code, 2020; AI Index, 2021 | Chart: 2021 AI Index Report

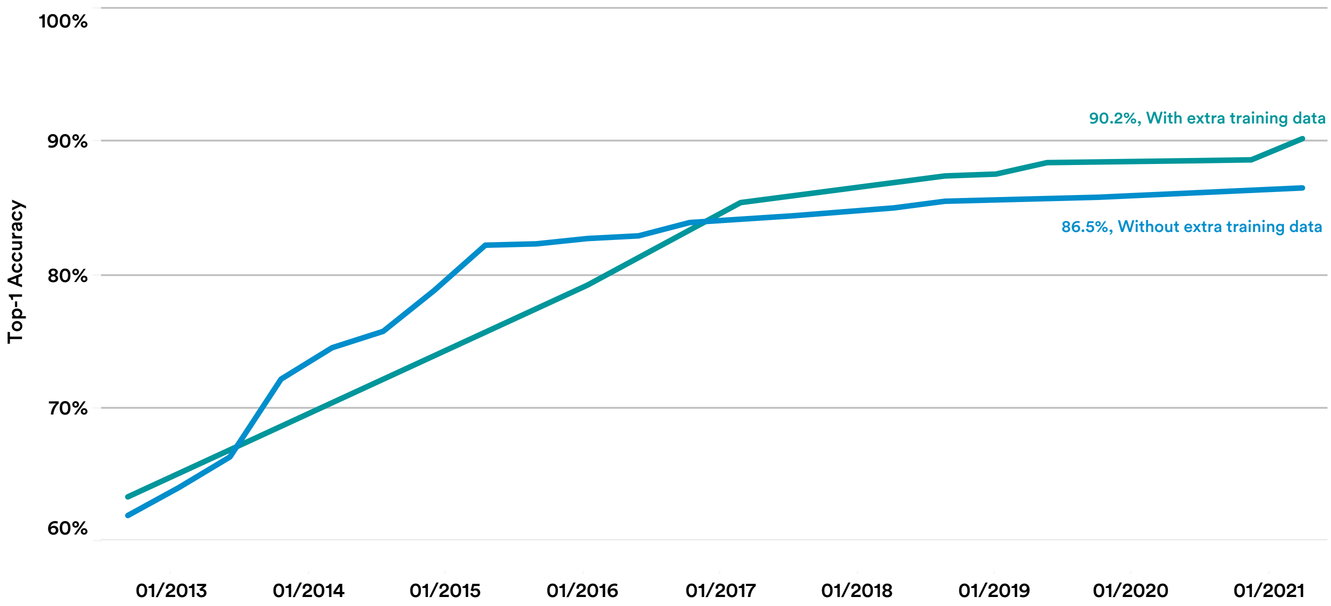

Figure 2.1.1

ImageNet: Top-5 Accuracy

Top-5 accuracy asks whether the correct label is in at least the classifier's top five predictions. Figure 2.1.2 shows that the error rate has improved from around 85% in 2013 to almost 99% in 2020.[2]

IMAGENET CHALLENGE: TOP-5 ACCURACY

Source: Papers with Code, 2020; AI Index, 2021 | Chart: 2021 AI Index Report

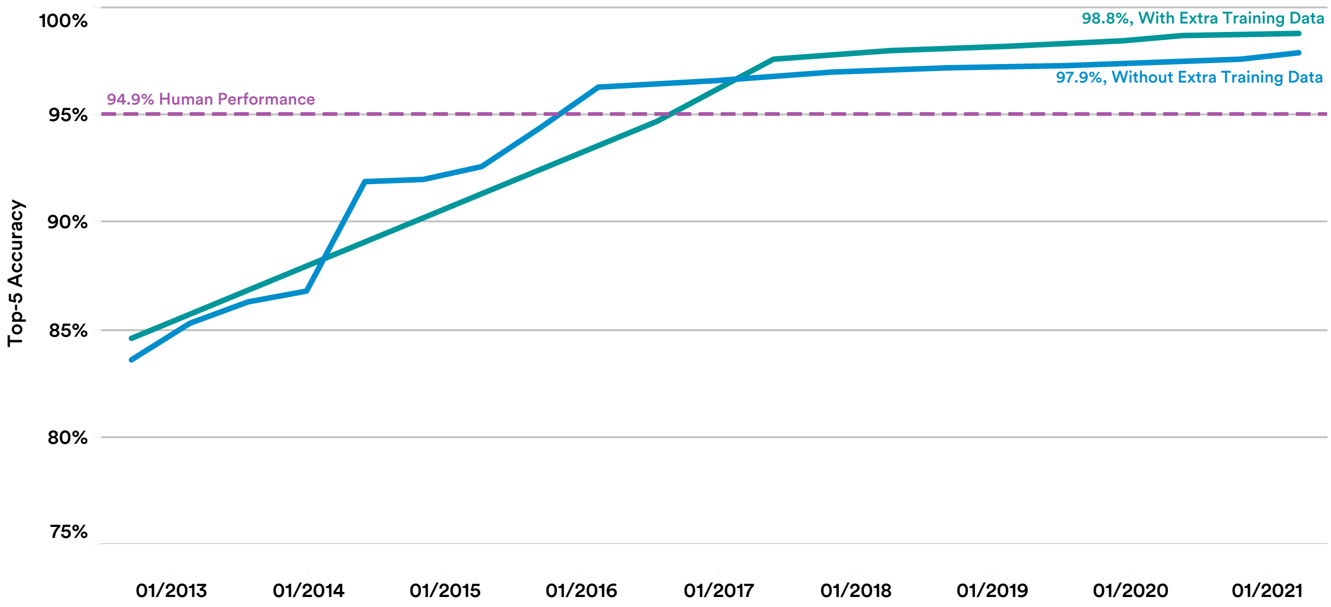

Figure 2.1.2

[2] Note: For data on human error, a human was shown 500 images and then was asked to annotate 1,500 test images; their error rate was 5.1% for Top-5 classification. This is a very rough baseline, but it gives us a sense of human performance on this task.

ImageNet: Training Time

Along with measuring the raw improvement in accuracy over time, it is useful to evaluate how long it takes to train image classifiers on ImageNet to a standard performance level as it sheds light on advances in the underlying computational infrastructure for large-scale AI training. This is important to measure because the faster you can train a system, the more quickly you can evaluate it and update it with new data. Therefore, the faster ImageNet systems can be trained, the more productive organizations can become at developing and deploying AI systems. Imagine the difference between waiting a few seconds for a system to train versus waiting a few hours, and what that difference means for the type and volume of ideas researchers explore and how risky they might be.

What follows are the results from MLPerf, a competition run by the MLCommons organization that challenges entrants to train an ImageNet network using a common (residual network) architecture, and then ranks systems according to the absolute "wall clock" time it takes them to train a system.[3]

As shown in Figure 2.1.3, the training time on ImageNet has fallen from 6.2 minutes (December 2018) to 47 seconds (July 2020). At the same time, the amount of hardware used to achieve these results has increased dramatically; frontier systems have been dominated by the use of "accelerator" chips, starting with GPUs in the 2018 results, and transitioning to Google's TPUs for the best-in-class results from 2019 and 2020.

Imagine the difference between waiting a few seconds for a system to train versus waiting a few hours, and what that difference means for the type and volume of ideas researchers explore and how risky they might be.

Distribution of Training Time: MLPerf does not just show the state of the art for each competition period; it also makes available all the data behind each entry in each competition cycle. This, in turn, reveals the distribution of training times for each period (Figure 2.1.3). (Note that in each MLPerf competition, competitors typically submit multiple entries that use different permutations of hardware.)

Figure 2.1.4 shows that in the past couple of years, training times have shortened, as has the variance between MLPerf entries. At the same time, competitors have started to use larger and larger numbers of accelerator chips to speed training times. This is in line with broader trends in AI development, as large-scale training becomes better understood, with a higher degree of shared best practices and infrastructure.

[3] The next MLPerf update is planned for June 2021.

IMAGENET: TRAINING TIME and HARDWARE of the BEST SYSTEM

Source: MLPerf, 2020 | Chart: 2021 AI Index Report

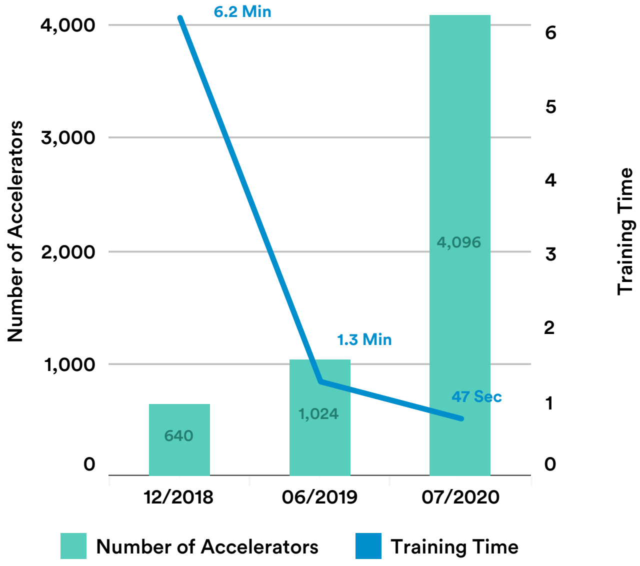

Figure 2.1.3

IMAGENET: DISTRIBUTION of TRAINING TIME

Source: MLPerf, 2020 | Chart: 2021 AI Index Report

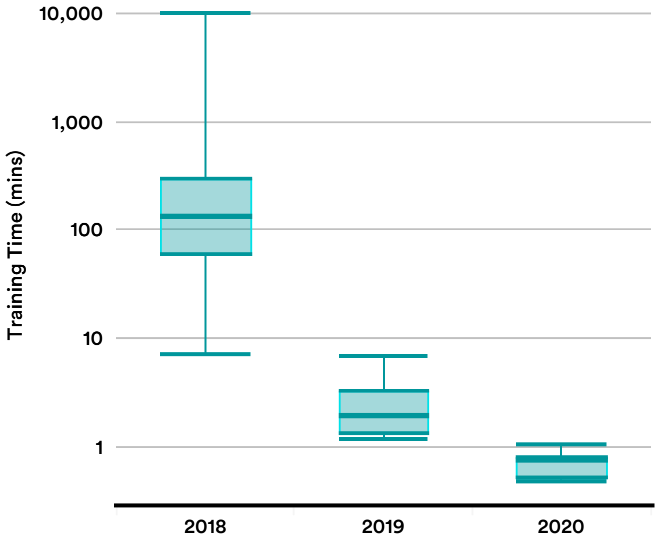

Figure 2.1.4

ImageNet: Training Costs

How much does it cost to train a contemporary image-recognition system? The answer, according to tests run by the Stanford DAWNBench team, is a few dollars in 2020, down by around 150 times from costs in 2017 (Figure 2.1.5). To put this in perspective, what cost one entrant around USD 1,100 to do in October 2017 now costs about USD 7.43. This represents progress in algorithm design as well as a drop in the costs of cloud-computing resources.

IMAGENET: TRAINING COST (to 93% ACCURACY)

Source: DAWNBench, 2020 | Chart: 2021 AI Index Report

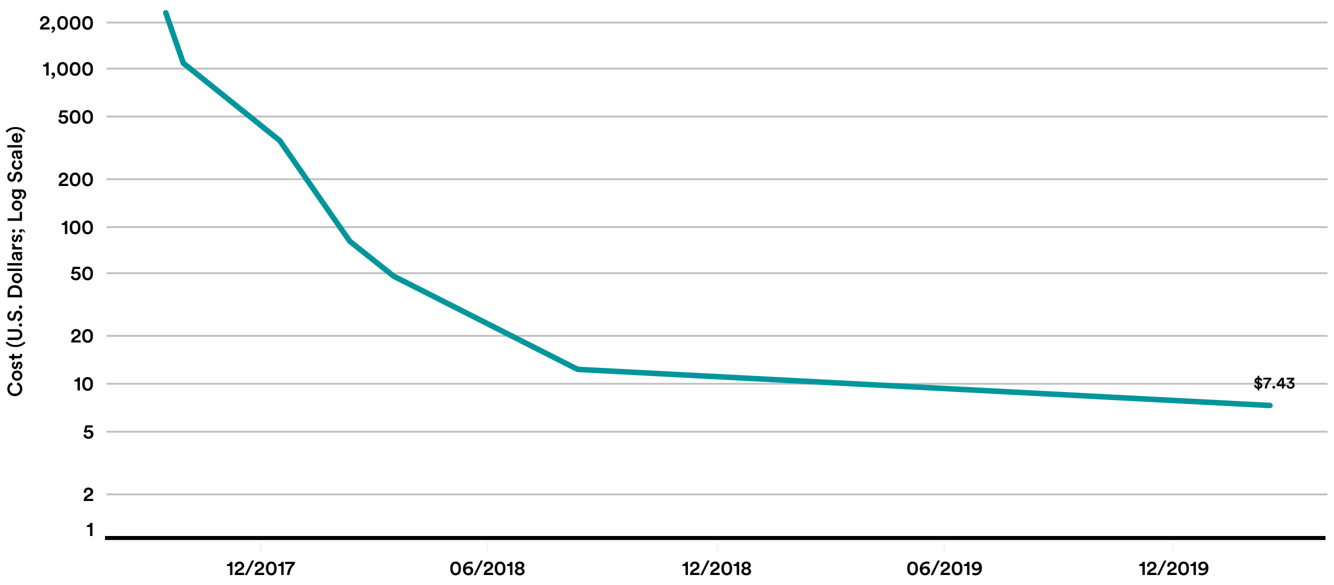

Figure 2.1.5

Harder Tests Beyond ImageNet

In spite of the progress in performance on ImageNet, current computer vision systems are still not perfect. To better study their limitations, researchers have in recent years started to develop more challenging image classification benchmarks. But since ImageNet is already a large dataset, which requires a nontrivial amount of resources to use, it does not intuitively make sense to simply expand the resolution of the images in ImageNet or the absolute size of the dataset—as either action would further increase the cost to researchers when training systems on ImageNet. Instead, people have tried to figure out new ways to test the robustness of image classifiers by creating custom datasets, many of which are compatible with ImageNet (and are typically smaller). These include

IMAGENET ADVERSARIAL:
This is a dataset of images similar to those found in ImageNet but incorporating natural confounders (e.g., a butterfly sitting on a carpet with a similar texture to the butterfly), and images that are persistently misclassified by contemporary systems. These images "cause consistent classification mistakes due to scene complications encountered in the long tail of scene configurations and by exploiting classifier blind spots," according to the researchers. Therefore, making progress on ImageNet Adversarial could improve the ability of models to generalize.

IMAGENET-C:
This is a dataset of common ImageNet images with 75 visual corruptions applied to them (e.g., changes in brightness, contrast, pixelations, fog effects, etc.). By testing systems against this, researchers can provide even more information about the generalization capabilities of these models.

IMAGENET-RENDITION:
This tests generalization by seeing how well ImageNet-trained models can categorize 30,000 illustrations of 200 ImageNet classes. Since ImageNet is designed to be built out of photos, generalization here indicates that systems have learned something more subtle about what they're trying to classify, because they're able to "understand" the relationship between illustrations and the photographed images they've been trained on.

What is the Time Table for Tracking This Data? As these benchmarks are relatively new, the plan is to wait a couple of years for the community to test a range of systems against them, which will generate the temporal information necessary to make graphs tracking progress overtime.

IMAGE GENERATION

Image generation is the task of generating images that look indistinguishable from "real" images. Image generation systems have a variety of uses, ranging from augmenting search capabilities (it is easier to search for a specific image if you can generate other images like it) to serving as an aid for other generative uses (e.g., editing images, creating content for specific purposes, generating multiple variations of a single image to help designers brainstorm, and so on).

In recent years, image generation progress has accelerated as a consequence of the continued improvement in deep learning–based algorithms, as well as the use of increased computation and larger datasets.

STL-10: Fréchet Inception Distance (FID) Score

One way to measure progress in image generation is via a technique called Fréchet Inception Distance (FID), which roughly correlates to the difference between how a given AI system "thinks" about a synthetic image versus a real image, where a real image has a score of 0 and synthetic images that look similar have scores that approach 0.

Figure 2.1.6 shows the progress of generative models over the past two years at generating convincing synthetic images in the STL-10 dataset, which is designed to test how effective systems are at generating images and gleaning other information about them.

STL-10: FRÉCHET INCEPTION DISTANCE (FID) SCORE

Source: Papers with Code, 2020 | Chart: 2021 AI Index Report

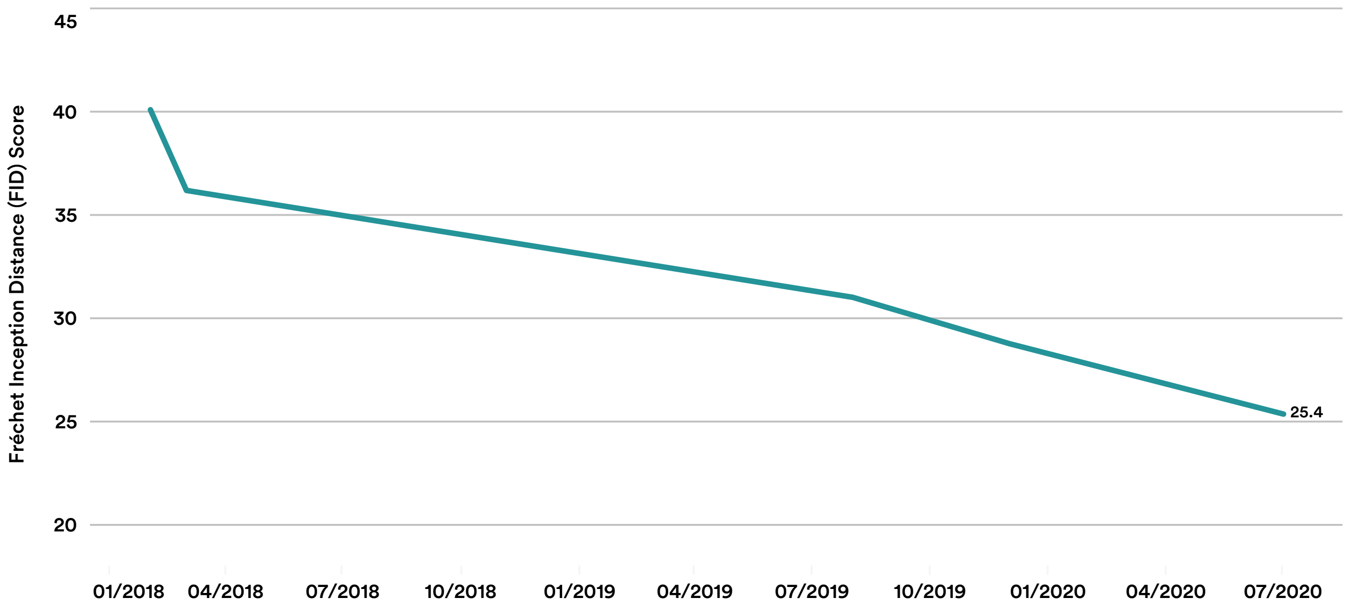

Figure 2.1.6

FID Versus Real Life

FID has drawbacks as an evaluation technique—specifically, it assesses progress on image generation via quantitative metrics that use data from the model itself, rather than other evaluation techniques. Another approach is using teams of humans to evaluate the outputs of these models; for instance, the Human eYe Perceptual Evaluation (HYPE) method tries to judge image quality by showing synthetically generated images to humans and using their qualitative ratings to drive the evaluation methodology. This approach is more expensive and slower to run than typical evaluations, but it may become more important as generative models get better.

Qualitative Examples: To get a sense of progress, you can look at the evolution in the quality of synthetically generated images over time. In Figure 2.1.7, you can see the best-in-class examples of synthetic images of human faces, ordered over time. By 2018, performance of this task had become sufficiently good that it is difficult for humans to easily model further progress (though it is possible to train machine learning systems to spot fakes, it is becoming more challenging). This provides a visceral example of recent progress in this domain and underscores the need for new evaluation methods to gauge future progress. In addition, in recent years people have turned to doing generative modeling on a broader range of categories than just images of people's faces, which is another way to test for generalization.

GAN PROGRESS ON FACE GENERATION

Source: Goodfellow et al., 2014; Radford et al., 2016; Liu & Tuzel, 2016; Karras et al., 2018; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020; AI Index, 2021

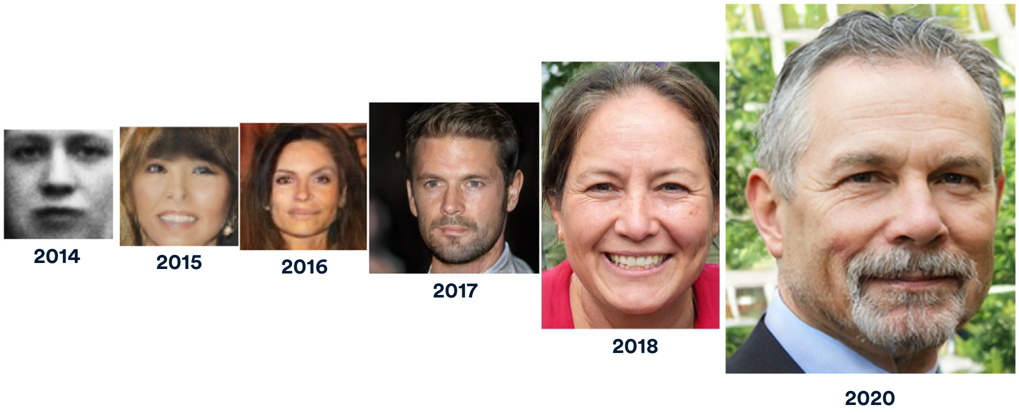

Figure 2.1.7

DEEPFAKE DETECTION

Advances in image synthesis have created new opportunities as well as threats. For instance, in recent years, researchers have harnessed breakthroughs in synthetic imagery to create AI systems that can generate synthetic images of human faces, then superimpose those faces onto the faces of other people in photographs or movies. People call this application of generative technology a "deepfake." Malicious uses of deepfakes include misinformation and the creation of (predominantly misogynistic) pornography. To try to combat this, researchers are developing deepfake-detection technologies.

Deepfake Detection Challenge (DFDC)

Created in September 2019 by Facebook, the Deepfake Detection Challenge (DFDC) measures progress on deepfake-detection technology. A two-part challenge, DFDC asks participants to train and test their models from a public dataset of around 100,000 clips. The submissions are scored on log loss, a classification metric based on probabilities. A smaller log loss means a more accurate prediction of deepfake videos. According to Figure 2.1.8, log loss dropped by around 0.5 as the challenge progressed between December 2019 and March 2020.

DEEPFAKE DETECTION CHALLENGE: LOG LOSS

Source: Kaggle, 2020 | Chart: 2021 AI Index Report

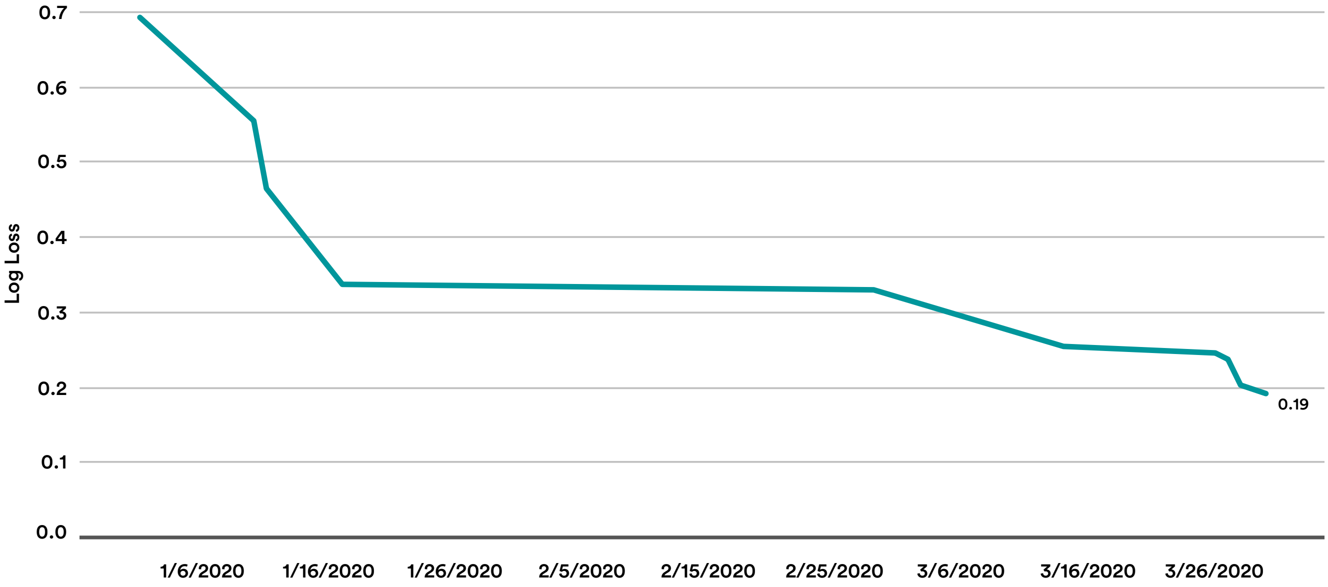

Figure 2.1.8

HUMAN POSE ESTIMATION

Human pose estimation is the problem of estimating the positions of human body parts or joints (wrists, elbows, etc.) from a single image. Human pose estimation is a classic "omni-use" AI capability. Systems that are good at this task can be used for a range of applications, such as creating augmented reality applications for the fashion industry, analyzing behaviors gleaned from physical body analysis in crowds, surveilling people for specific behaviors, aiding with analysis of live sporting and athletic events, mapping the movements of a person to a virtual avatar, and so on.

Common Objects in Context (COCO): Keypoint Detection Challenge

Common Objects in Context (COCO) is a large-scale dataset for object detection, segmentation, and captioning with 330,000 images and 1.5 million object instances. Its Keypoint Detection Challenge requires machines to simultaneously detect an object or a person and localize their body keypoints—points in the image that stand out, such as a person's elbows, knees, and other joints. The task evaluates algorithms based on average precision (AP), a metric that can be used to measure the accuracy of object detectors. Figure 2.1.9 shows that the accuracy of algorithms in this task has improved by roughly 33% in the past four years, with the latest machine scoring 80.8% on average precision.

COCO KEYPOINT CHALLENGE: AVERAGE PRECISION

Source: COCO Leaderboard, 2020 | Chart: 2021 AI Index Report

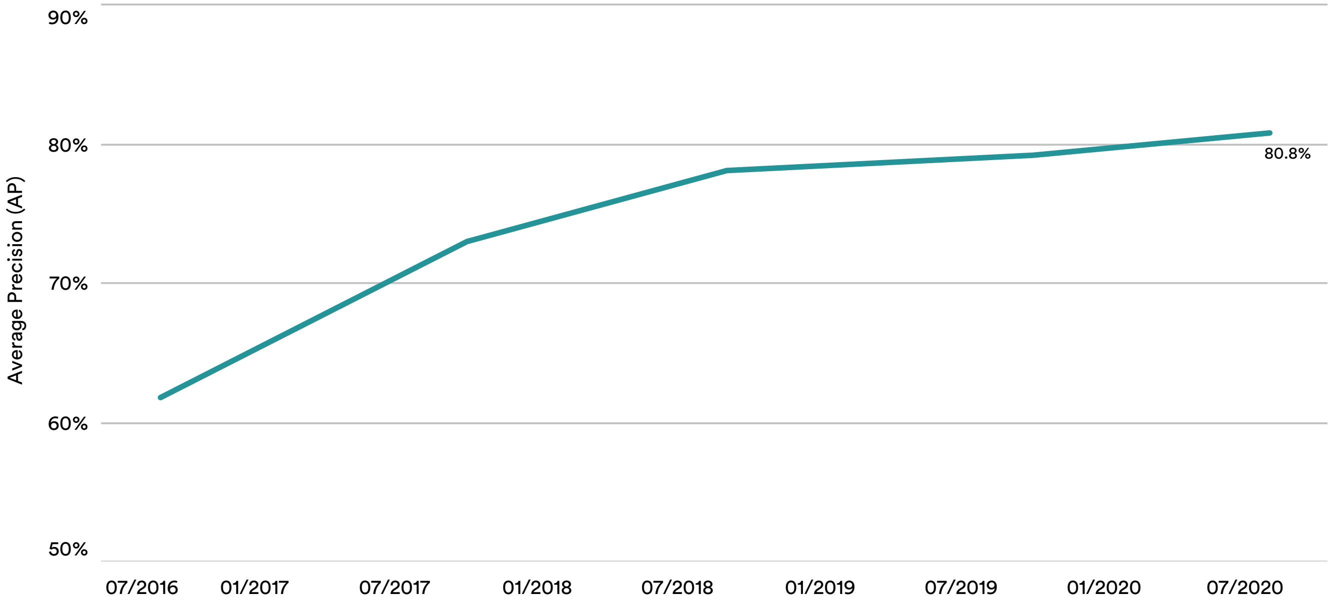

Figure 2.1.9

Common Objects in Context (COCO): DensePose Challenge

DensePose, or dense human pose estimation, is the task of extracting a 3D mesh model of a human body from a 2D image. After open-sourcing a system called DensePose in 2018, Facebook built DensePose-COCO, a large-scale dataset of image-to-surface correspondences annotated on 50,000 COCO images. Since then, DensePose has become a canonical benchmark dataset.

The COCO DensePose Challenge involves tasks of simultaneously detecting people, segmenting their bodies, and estimating the correspondences between image pixels that belong to a human body and a template 3D model. The average precision is calculated based on the geodesic point similarity (GPS) metric, a correspondence matching score that measures the geodesic distances between the estimated points and the true location of the body points on the image. The accuracy has grown from 56% in 2018 to 72% in 2019 (Figure 2.1.10).

COCO DENSEPOSE CHALLENGE: AVERAGE PRECISION

Source: arXiv & CodaLab, 2020 | Chart: 2021 AI Index Report

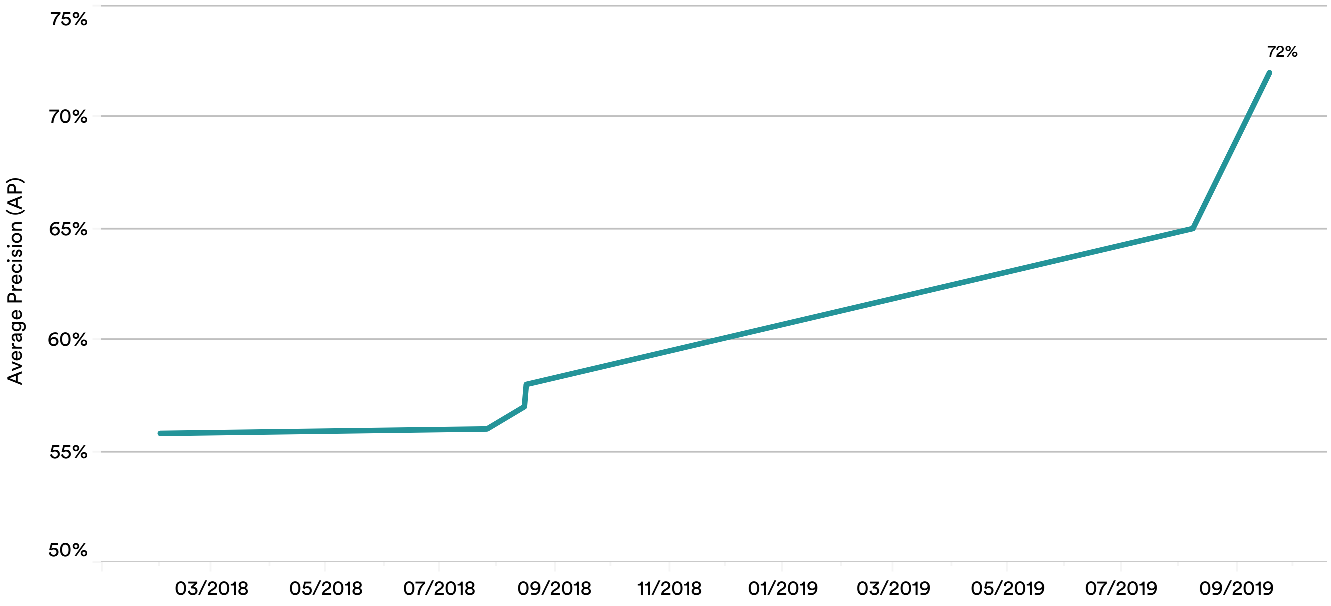

Figure 2.1.10

SEMANTIC SEGMENTATION

Semantic segmentation is the task of classifying each pixel in an image to a particular label, such as person, cat, etc. Where image classification tries to assign a label to the entire image, semantic segmentation tries to isolate the distinct entities and objects in a given image, allowing for more fine-grained identification. Semantic segmentation is a basic input technology for self-driving cars (identifying and isolating objects on roads), image analysis, medical applications, and more.

Cityscapes

Cityscapes is a large-scale dataset of diverse urban street scenes across 50 different cities recorded during the daytime over several months (during spring, summer, and fall) of the year. The dataset contains 5,000 images with high-quality, pixel-level annotations and 20,000 weekly labeled ones. Semantic scene understanding, especially in the urban space, is crucial to the environmental perception of autonomous vehicles. Cityscapes is useful for training deep neural networks to understand the urban environment.

One Cityscapes task that focuses on semantic segmentation is the pixel-level semantic labeling task. This task requires an algorithm to predict the per-pixel semantic labeling of the image, partitioning an image into different categories, like cars, buses, people, trees, and roads. Participants are evaluated based on the intersection-over-union (IoU) metric. A higher IoU score means a better segmentation accuracy. Between 2014 and 2020, the mean IoU increased by 35% (Figure 2.1.11). There was a significant boost to progress in 2016 and 2017 when people started using residual networks in these systems.

CITYSCAPES CHALLENGE: PIXEL-LEVEL SEMANTIC LABELING TASK

Source: Papers with Code, 2020 | Chart: 2021 AI Index Report

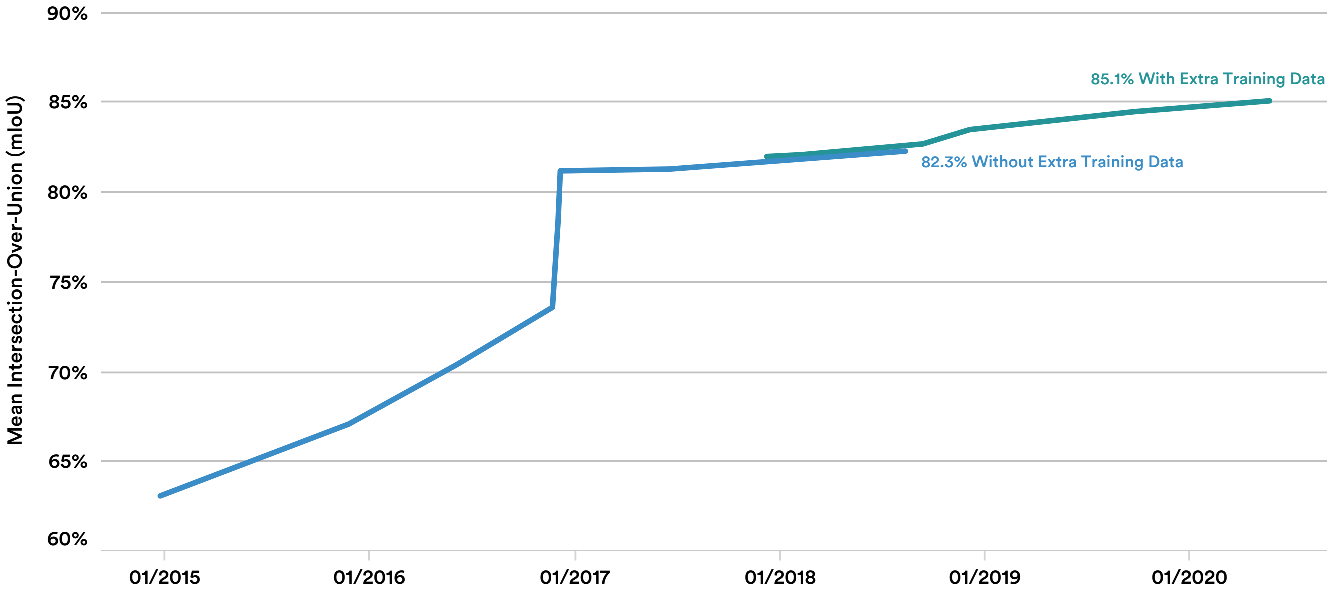

Figure 2.1.11

EMBODIED VISION

The performance data so far shows that computer vision systems have advanced tremendously in recent years. Object recognition, semantic segmentation, and human pose estimation, among others, have now achieved significant levels of performance. Note that these visual tasks are somewhat passive or disembodied. That is, they can operate on images or videos taken from camera systems that are not physically able to interact with the surrounding environment. As a consequence of the continuous improvement in those passive tasks, researchers have now started to develop more advanced AI systems that can be interactive or embodied—that is, systems that can physically interact with and modify the surrounding environment in which they operate: for example, a robot that can visually survey a new building and autonomously navigate it, or a robot that can learn to assemble pieces by watching visual demonstrations instead of being manually programmed for this.

Progress in this area is currently driven by the development of sophisticated simulation environments, where researchers can deploy robots in virtual spaces, simulate what their cameras would see and capture, and develop AI algorithms for navigation, object search, and object grasping, among other interactive tasks. Because of the relatively early nature of this field, there are few standardized metrics to measure progress. Instead, here are brief highlights of some of the available simulators, their year of release, and any other significant feature.

- **Thor** (AI2, **2017**) focuses on sequential abstract reasoning with predefined "magic" actions that are applicable to objects.
- **Gibson** (Stanford, **2018**) focuses on visual navigation in photorealistic environments obtained with 3D scanners.
- **iGibson** (Stanford, **2019**) focuses on full interactivity in large realistic scenes mapped from real houses and made actable: navigation + manipulation (known in robotics as "mobile manipulation").
- **AI Habitat** (Facebook, **2019**) focuses on visual navigation with an emphasis on much faster execution, enabling more computationally expensive approaches.
- **ThreeDWorld** (MIT and Stanford, **2020**) focuses on photorealistic environments through game engines, as well as adds simulation of flexible materials, fluids, and sounds.
- **SEAN-EP** (Yale, **2020**) is a human-robot interaction environment with simulated virtual humans that enables the collection of remote demonstrations from humans via a web browser.
- **Robosuite** (Stanford and UT Austin, **2020**) is a modular simulation framework and benchmark for robot learning.

Video analysis is the task of making inferences over sequential image frames, sometimes with the inclusion of an audio feed. Though many AI tasks rely on single-image inferences, a growing body of applications require computer vision machines to reason about videos. For instance, identifying a specific dance move benefits from seeing a variety of frames connected in a temporal sequence; the same is true of making inferences about an individual seen moving through a crowd, or a machine carrying out a sequence of movements over time.

2.2 COMPUTER VISION—VIDEO

ACTIVITY RECOGNITION

The task of activity recognition is to identify various activities from video clips. It has many important everyday applications, including surveillance by video cameras and autonomous navigation of robots. Research on video understanding is still focused on short events, such as videos that are a few seconds long. Longer-term video understanding is slowly gaining traction.

ActivityNet

Introduced in 2015, ActivityNet is a large-scale video benchmark for human-activity understanding. The benchmark tests how well algorithms can label and categorize human behaviors in videos. By improving performance on tasks like ActivityNet, AI researchers are developing systems that can categorize more complex behaviors than those that can be contained in a single image, like characterizing the behavior of pedestrians on a self-driving car's video feed or providing better labeling of specific movements in sporting events.

ActivityNet: Temporal Action Localization Task

The temporal action localization task in the ActivityNet challenge asks machines to detect time segments in a 600-hour, untrimmed video sequence that contains several activities. Evaluation on this task focuses on (1) localization: how well can the system localize the interval with the precise start time and end time; and (2) recognition: how well can the system recognize the activity and classify it into the correct category (such as throwing, climbing, walking the dog, etc.). Figure 2.2.1 shows that the highest mean average precision of the temporal action localization task among submissions has grown by 140% in the last five years.

ACTIVITYNET: TEMPORAL ACTION LOCALIZATION TASK
Source: ActivityNet, 2020 | Chart: 2021 AI Index Report

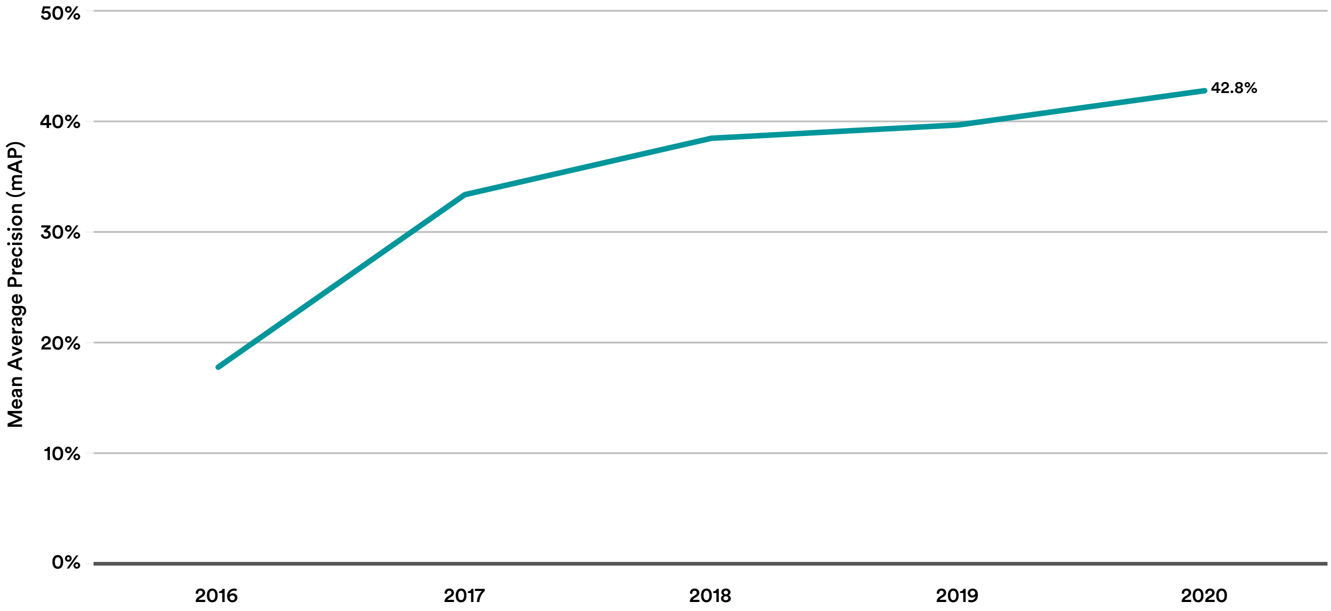

Figure 2.2.1

ActivityNet: Hardest Activity

Figure 2.2.2 shows the hardest activities of the temporal action location task in 2020 and how their mean average precision compares with the 2019 result. Drinking coffee remained the hardest activity in 2020. Rock-paper-scissors, though still the 10th hardest activity, saw the greatest improvement among all activities, increasing by 129.2%—from 6.6% in 2019 to 15.22% in 2020.

ACTIVITYNET: HARDEST ACTIVITIES, 2019-20

Source: ActivityNet, 2020 | Chart: 2021 AI Index Report

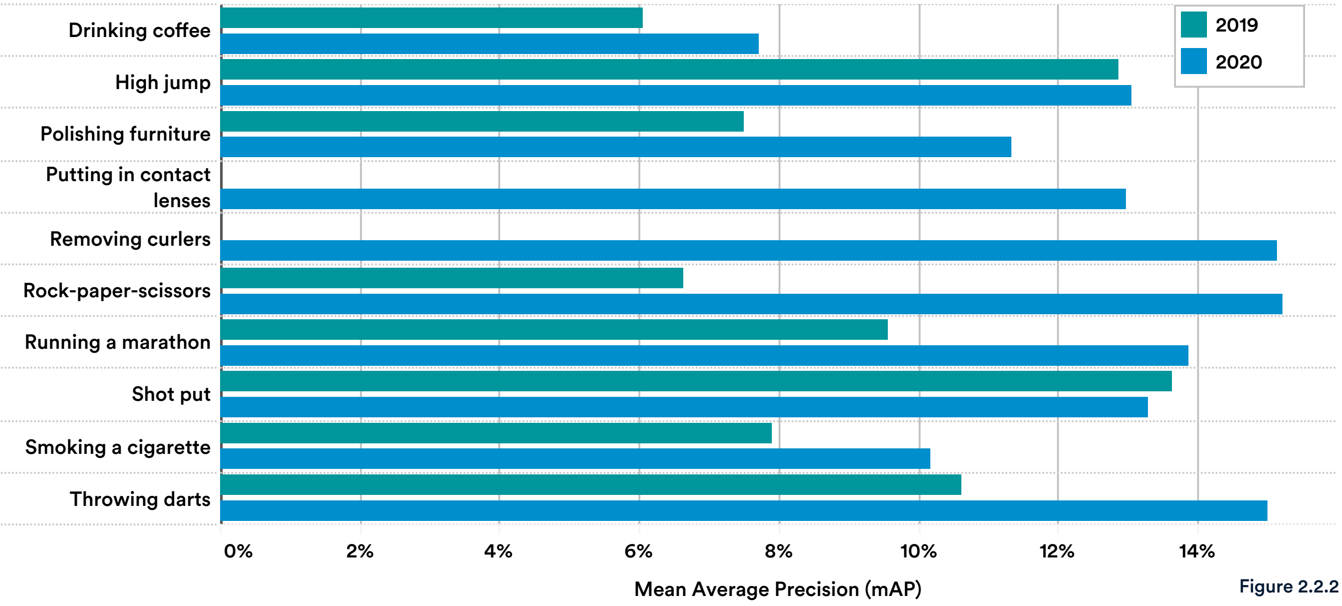

Figure 2.2.2

OBJECT DETECTION

Object detection is the task of identifying a given object in an image. Frequently, image classification and image detection are coupled together in deployed systems. One way to get a proxy measure for the improvement in deployed object recognition systems is to study the advancement of widely used object detection systems.

You Only Look Once (YOLO)

You Only Look Once (YOLO) is a widely used open source system for object detection, so its progress has been included on a standard task on YOLO variants to give a sense of how research percolates into widely used open source tools. YOLO has gone through multiple iterations since it was first published in 2015. Over time, YOLO has been optimized along two constraints: performance and inference latency, as shown in Figure 2.2.3. What this means, specifically, is that by measuring YOLO, one can measure the advancement of systems that might not have the best absolute performance but are designed around real-world needs, like low-latency inference over video streams. Therefore, YOLO systems might not always contain the absolute best performance as defined in the research literature, but they will represent good performance when faced with trade-offs such as inference time.

YOU ONLY LOOK ONCE (YOLO): MEAN AVERAGE PRECISION

Source: Redmon & Farhadi (2016 & 2018), Bochkovskiy et al. (2020), Long et al. (2020) | Chart: 2021 AI Index Report

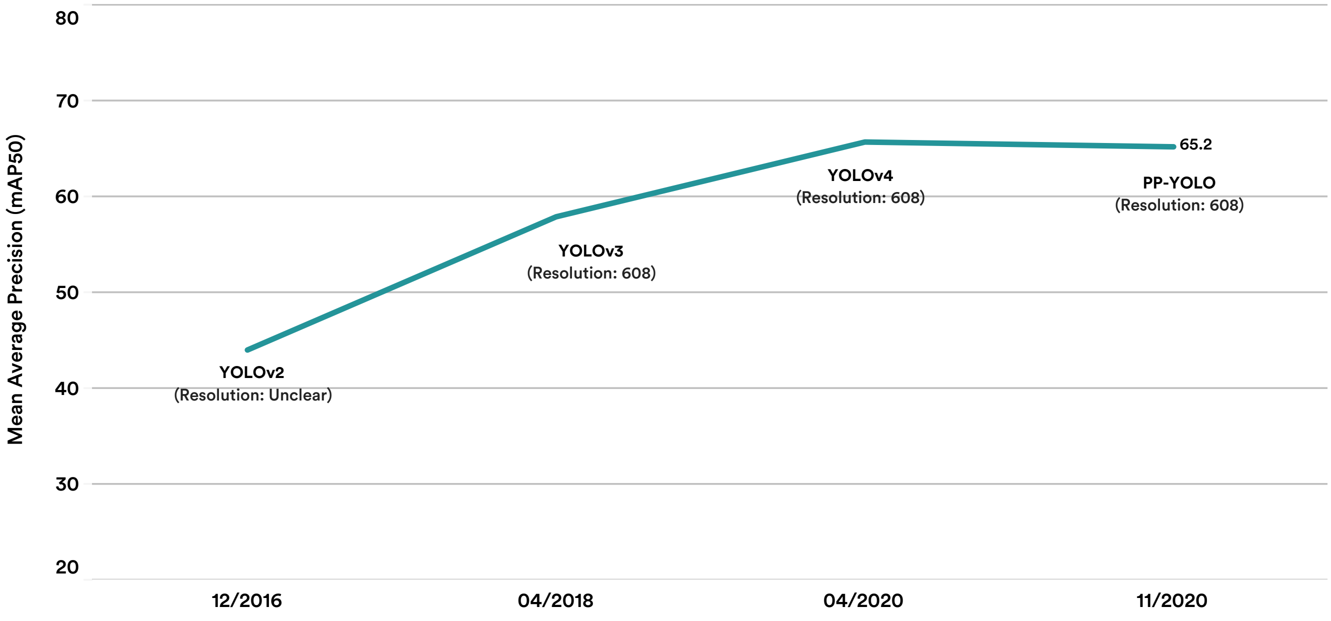

Figure 2.2.3

FACE DETECTION AND RECOGNITION

Facial detection and recognition is one of the use-cases for AI that has a sizable commercial market and has generated significant interest from governments and militaries. Therefore, progress in this category gives us a sense of the rate of advancement in economically significant parts of AI development.

National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT)

The Face Recognition Vendor Tests (FRVT) by the National Institute of Standards and Technology (NIST) provide independent evaluations of commercially available and prototype face recognition technologies. FRVT measures the performance of automated face recognition technologies used for a wide range of civil and governmental tasks (primarily in law enforcement and homeland security), including verification of visa photos, mug shot images, and child abuse images.

Figure 2.2.4 shows the results of the top-performing 1:1 algorithms measured on false non-match rate (FNMR) across several different datasets. FNMR refers to the rate at which the algorithm fails when attempting to match the image with the individual. Facial recognition technologies on mug-shot-type and visa photos have improved the most significantly in the past four years, falling from error rates of close to 50% to a fraction of a percent in 2020.[4]

NIST FRVT 1:1 VERIFICATION ACCURACY by DATASET, 2017-20

Source: National Institute of Standards and Technology, 2020 | Chart: 2021 AI Index Report

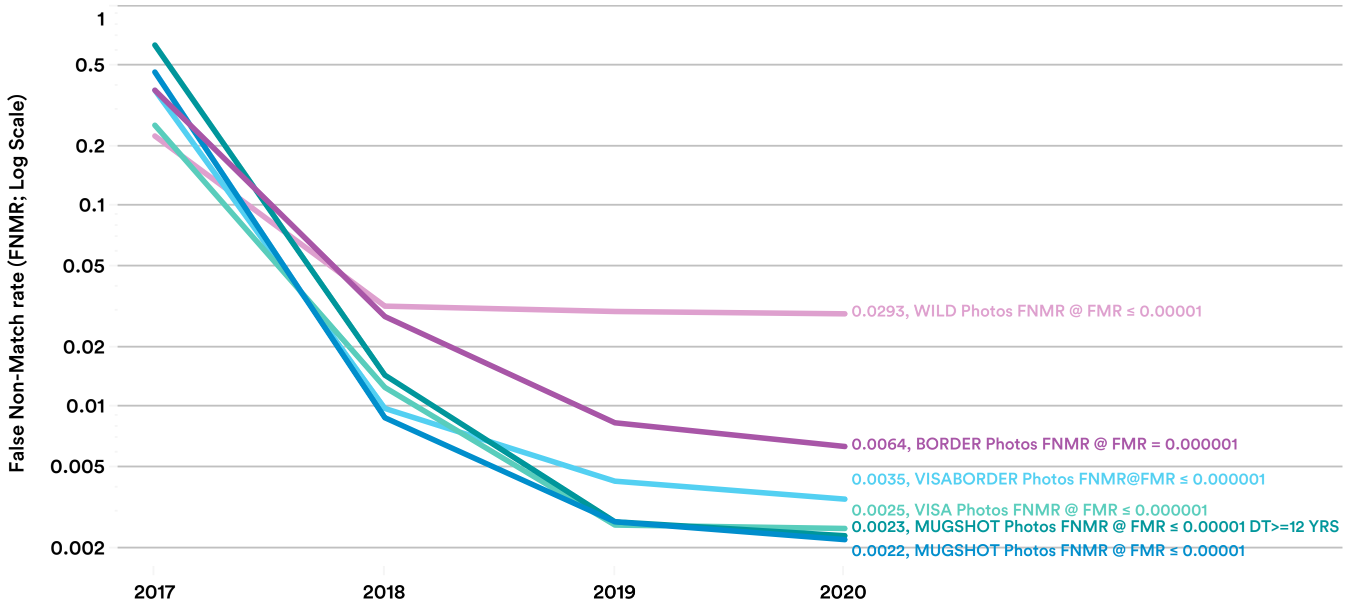

Figure 2.2.4

[4] You can view details and examples of various datasets on periodically updated FRVT 1:1 verification reports.

Natural language processing (NLP) involves teaching machines to interpret, classify, manipulate, and generate language. From the early use of handwritten rules and statistical techniques to the recent adoption of generative models and deep learning, NLP has become an integral part of our lives, with applications in text generation, machine translation, question answering, and other tasks.

2.3 LANGUAGE

In recent years, advances in natural language processing technology have led to significant changes in large-scale systems that billions of people access. For instance, in late 2019, Google started to deploy its BERT algorithm into its search engine, leading to what the company said was a significant improvement in its in-house quality metrics. Microsoft followed suit, announcing later in 2019 that it was using BERT to augment its Bing search engine.

ENGLISH LANGUAGE UNDERSTANDING BENCHMARKS

SuperGLUE

Launched in May 2019, SuperGLUE is a single-metric benchmark that evaluates the performance of a model on a series of language understanding tasks on established datasets. SuperGLUE replaced the prior GLUE benchmark (introduced in 2018) with more challenging and diverse tasks.

The SuperGLUE score is calculated by averaging scores on a set of tasks. Microsoft's DeBERTa model now tops the SuperGLUE leaderboard, with a score of 90.3, compared with an average score of 89.8 for SuperGLUE's "human baselines." This does not mean that AI systems have surpassed human performance on all SuperGLUE tasks, but it does mean that the average performance across the entire suite has exceeded that of a human baseline. The rapid pace of progress (Figure 2.3.1) suggests that SuperGLUE may need to be made more challenging or replaced by harder tests in the future, just as SuperGLUE replaced GLUE.

SUPERGLUE BENCHMARK

Source: SuperGLUE Leaderboard, 2020 | Chart: 2021 AI Index Report

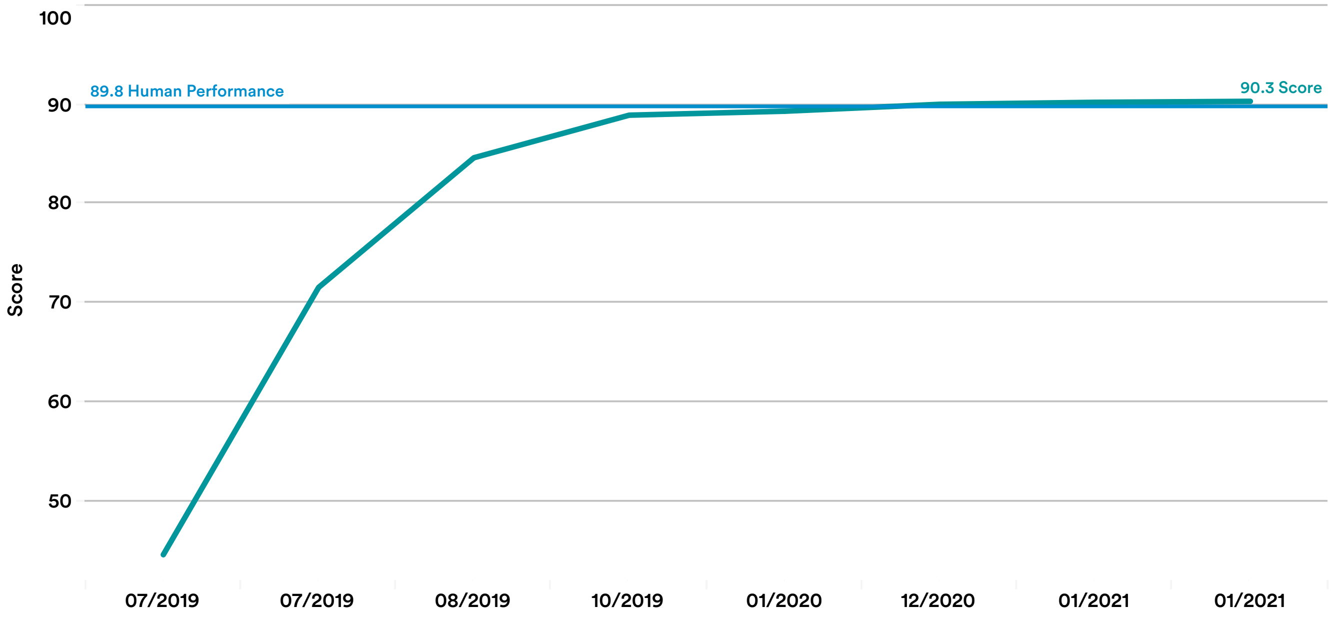

Figure 2.3.1

SQuAD

The Stanford Question Answering Dataset, or SQuAD, is a reading-comprehension benchmark that measures how accurately a NLP model can provide short answers to a series of questions pertaining to a small article of text. The SQuAD test makers established a human performance benchmark by having a group of people read Wikipedia articles on a variety of topics and then answer multiple-choice questions about those articles. Models are given the same task and are evaluated on the F1 score, or the average overlap between the model prediction and the correct answer. Higher scores indicate better performance.

Two years after the introduction of the original SQuAD, in 2016, SQuAD 2.0 was developed once the initial benchmark revealed increasingly fast performances by the participants (mirroring the trend seen in GLUE and SuperGLUE). SQuAD 2.0 combines the 100,000 questions in SQuAD 1.1 with over 50,000 unanswerable questions written by crowdworkers to resemble answerable ones. The objective is to test how well systems can answer questions and to determine when systems know that no answer exists.

As Figure 2.3.2 shows, the F1 score for SQuAD 1.1 improved from 67.75 in August 2016 to surpass human performance of 91.22 in September 2018—a 25-month period—whereas SQuAD 2.0 took just 10 months to beat human performance (from 66.3 in May 2018 to 89.47 in March 2019). In 2020, the most advanced models of SQuAD 1.1 and SQuAD 2.0 reached the F1 scores of 95.38 and 93.01, respectively.

SQUAD 1.1 and SQUAD 2.0: F1 SCORE
Source: CodaLab Worksheets, 2020 | Chart: 2021 AI Index Report

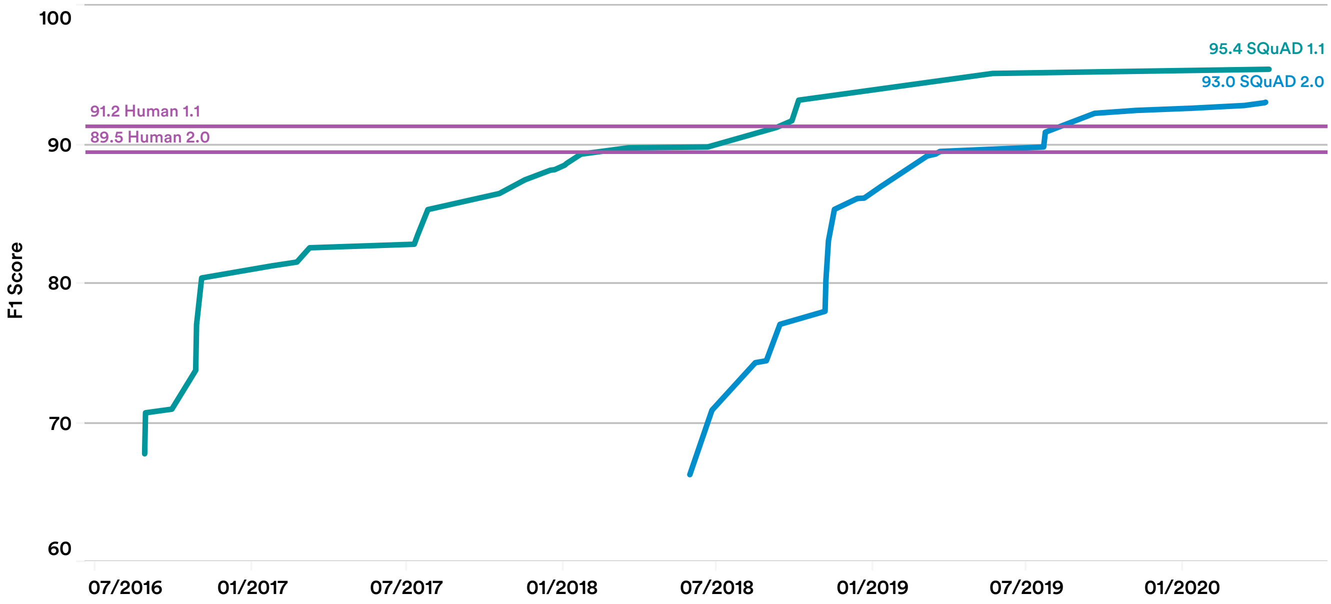

Figure 2.3.2

COMMERCIAL MACHINE TRANSLATION (MT)

Machine translation (MT), the subfield of computational linguistics that investigates the use of software to translate text or speech from one language to another, has seen significant improvement due to advances in machine learning. Recent progress in MT has prompted developers to shift from symbolic approaches toward ones that use both statistical and deep learning approaches.

Number of Commercially Available MT Systems

The trend in the number of commercially available systems speaks to the significant growth of commercial machine translation technology and its rapid adoption in the commercial marketplace. In 2020, the number of commercially available independent cloud MT systems with pre-trained models increased to 28, from 8 in 2017, according to Intento, a startup that evaluates commercially available MT services (Figure 2.3.3).

NUMBER of INDEPENDENT MACHINE TRANSLATION SERVICES

Source: Intento, 2020 | Chart: 2021 AI Index Report

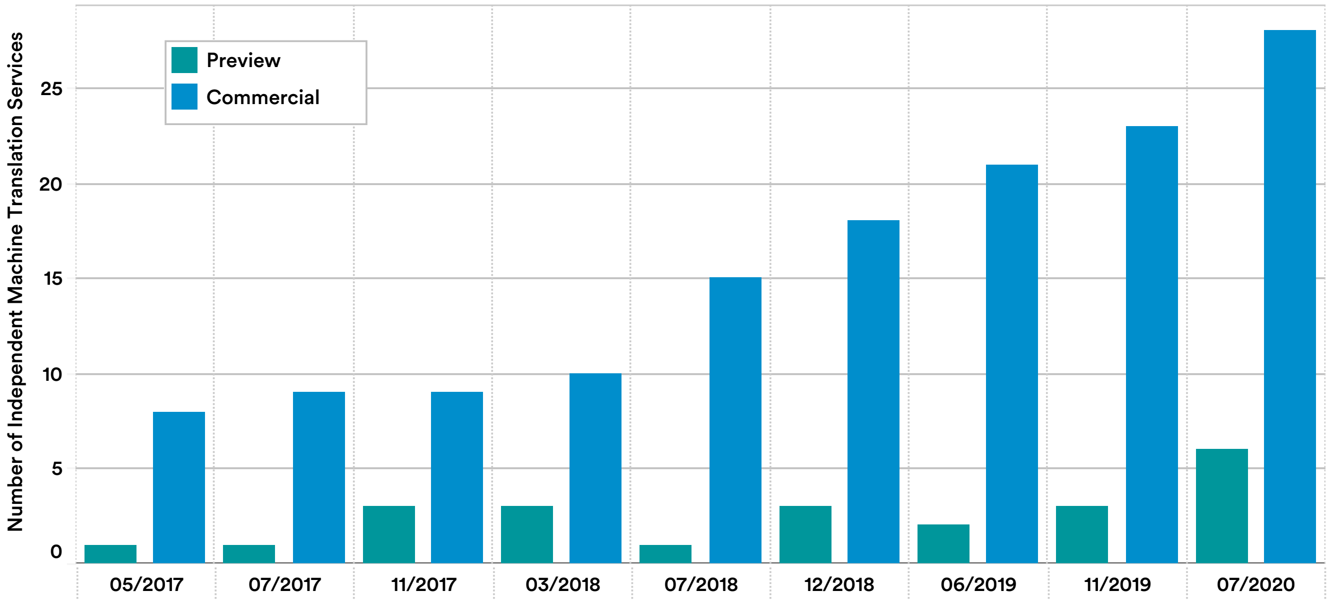

Figure 2.3.3

GPT-3

In July 2020, OpenAI unveiled GPT-3, the largest known dense language model. GPT-3 has 175 billion parameters and was trained on 570 gigabytes of text. For comparison, its predecessor, GPT-2, was over 100 times smaller, at 1.5 billion parameters. This increase in scale leads to surprising behavior: GPT-3 is able to perform tasks it was not explicitly trained on with zero to few training examples (referred to as zero-shot and few-shot learning, respectively). This behavior was mostly absent in the much smaller GPT-2. Furthermore, for some tasks (but not all; e.g., SuperGLUE and SQuAD2), GPT-3 outperforms state-of-the-art models that were explicitly trained to solve those tasks with far more training examples.

Figure 2.3.4, adapted from the GPT-3 paper, demonstrates the impact of scale (in terms of model parameters) on task accuracy (higher is better) in zero-, one-, and few-shot learning regimes. Each point on the curve corresponds to an average performance accuracy, aggregated across 42 accuracy-oriented benchmarks. As model size increases, average accuracy in all task regimes increases accordingly. Few-shot learning accuracy increases more rapidly with scale, compared with zero-shot learning, which suggests that large models can perform surprisingly well given minimal context.

GPT-3: AVERAGE PERFORMANCE across 42 BENCHMARKS

Source: OpenAI (Brown et al.), 2020 | Chart: 2021 AI Index Report

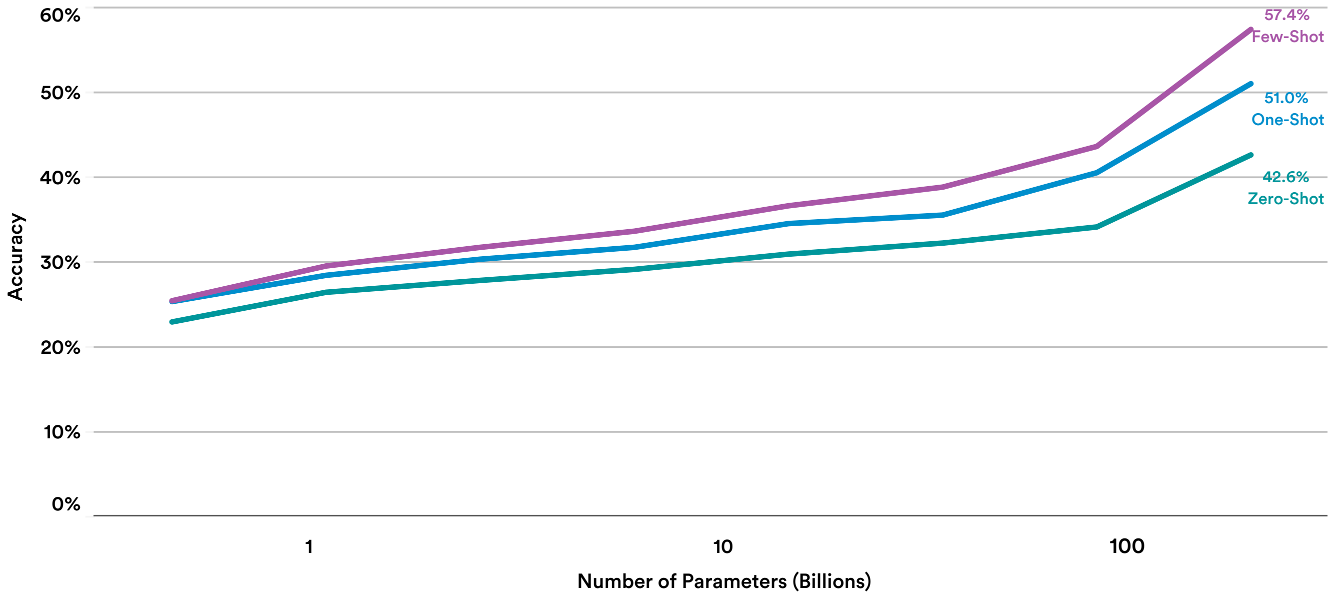

Figure 2.3.4

That a single model can achieve state-of-the-art or close to state-of-the-art performance in limited-training-data regimes is impressive. Most models until now have been designed for a single task, and thus can be evaluated effectively by a single metric. In light of GPT-3, we anticipate novel benchmarks that are explicitly designed to evaluate zero- to few-shot learning performance for language models. This will not be straightforward. Developers are increasingly finding model novel capabilities (e.g., the ability to generate a website from a text description) that will be difficult to define, let alone measure performance on. Nevertheless, the AI Index is committed to tracking performance in this new context as it evolves.

Despite its impressive capabilities, GPT-3 has several shortcomings, many of which are outlined in the original paper. For example, it can generate racist, sexist, and otherwise biased text. Furthermore, GPT-3 (and other language models) can generate unpredictable and factually inaccurate text. Techniques for controlling and “steering” such outputs to better align with human values are nascent but promising. GPT-3 is also expensive to train, which means that only a limited number of organizations with abundant resources can currently afford to develop and deploy such models. Finally, GPT-3 has an unusually large number of uses, from chatbots to computer code generation to search. Future users are likely to discover more applications, both good and bad, making it difficult to identify the range of possible uses and forecast their impact on society.

Nevertheless, research to address harmful outputs and uses is ongoing at several universities and industrial research labs, including OpenAI. For more details, refer to work by Bender and Gebru et al. and the proceedings from a recent Stanford Institute for Human-Centered Artificial Intelligence (HAI) workshop (which included researchers from OpenAI), “Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models.”

That a single model can achieve state-of-the-art or close to state-of-the-art performance in limited-training-data regimes is impressive. Most models until now have been designed for a single task, and thus can be evaluated effectively by a single metric.

2.4 LANGUAGE REASONING SKILLS

VISION AND LANGUAGE REASONING

Vision and language reasoning is a research area that addresses how well machines jointly reason about visual and text data.

Visual Question Answering (VQA) Challenge

The VQA challenge, introduced in 2015, requires machines to provide an accurate natural language answer, given an image and a natural language question about the image based on a public dataset. Figure 2.4.1 shows that the accuracy has grown by almost 40% since its first installment at the International Conference on Computer Vision (ICCV) in 2015. The highest accuracy of the 2020 challenge is 76.4%. This achievement is closer to the human baseline of 80.8% accuracy and represents a 1.1% absolute increase in performance from the top 2019 algorithm.

VISUAL QUESTION ANSWERING (VQA) CHALLENGE: ACCURACY

Source: VQA Challenge, 2020 | Chart: 2021 AI Index Report

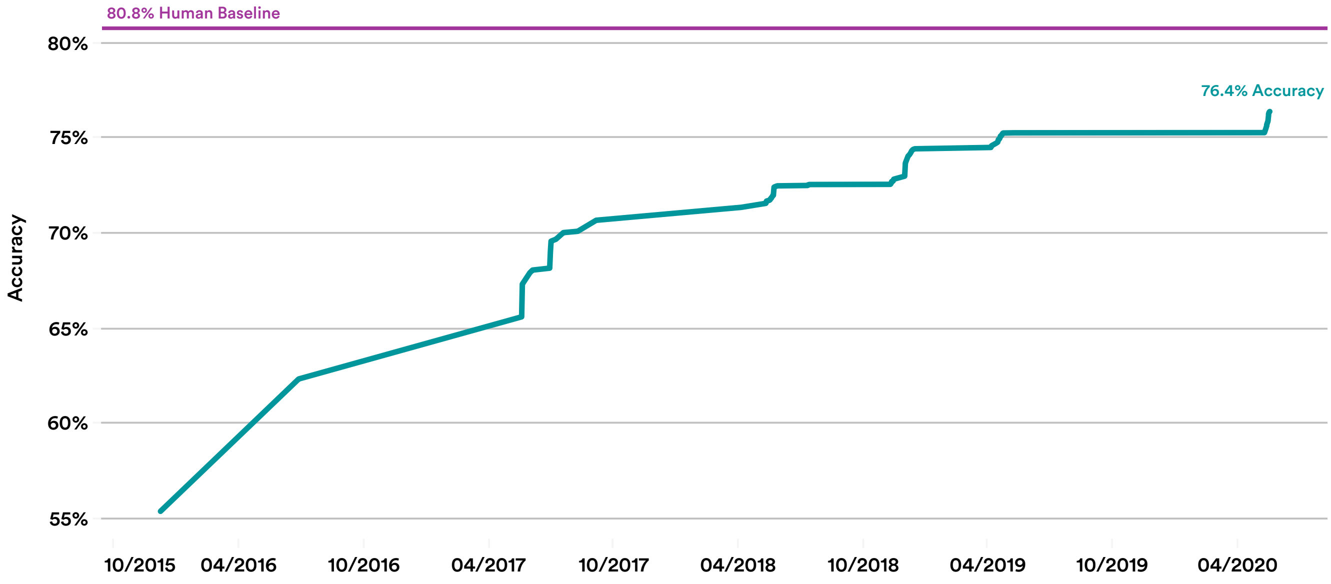

Figure 2.4.1

Visual Commonsense Reasoning (VCR) Task

The Visual Commonsense Reasoning (VCR) task, first introduced in 2018, asks machines to answer a challenging question about a given image and justify that answer with reasoning (whereas VQA just requests an answer). The VCR dataset contains 290,000 pairs of multiple-choice questions, answers, and rationales, as well as over 110,000 images from movie scenes.

The main evaluation mode for the VCR task is the Q->AR score, requiring machines to first choose the right answer (A) to a question (Q) among four answer choices (Q->A) and then select the correct rationale (R) among four rationale choices based on the answer. A higher score is better, and human performance on this task is measured by a QA->R score of 85. The best-performing machine has improved on the Q->AR score from 44 in 2018 to 70.5 in 2020 (Figure 2.4.2), which represents a 60.2% increase in performance from the top competitor in 2019.

VISUAL COMMONSENSE REASONING (VCR) TASK: Q->AR Score

Source: VCR Leaderboard, 2020 | Chart: 2021 AI Index Report

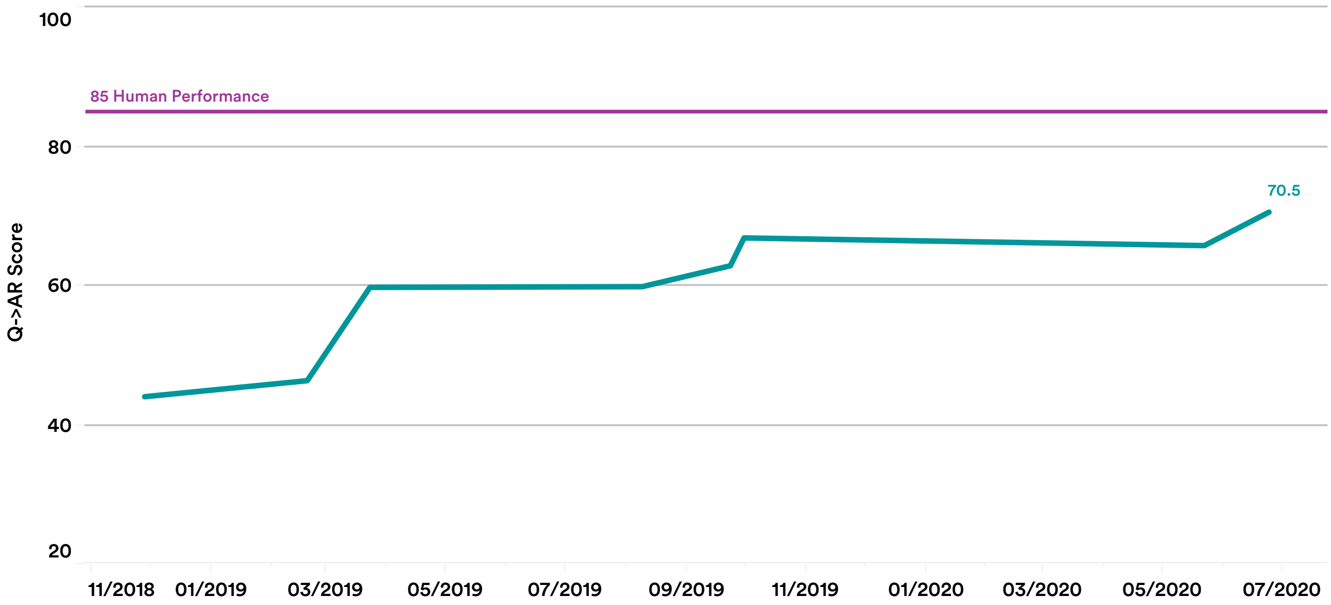

Figure 2.4.2

A major aspect of AI research is the analysis and synthesis of human speech conveyed via audio data. In recent years, machine learning approaches have drastically improved performance across a range of tasks.

2.5 SPEECH

SPEECH RECOGNITION

Speech recognition, or automatic speech recognition (ASR), is the process that enables machines to recognize spoken words and convert them to text. Since IBM introduced its first speech recognition technology in 1962, the technology has evolved with voice-driven applications such as Amazon Alexa, Google Home, and Apple Siri becoming increasingly prevalent. The flexibility and predictive power of deep neural networks, in particular, has allowed speech recognition to become more accessible.

Transcribe Speech: LibriSpeech

LibriSpeech is a dataset, first introduced in 2015, made up of 1,000 hours of speech from audiobooks. It has become widely used for the development and testing of speech recognition technologies. In recent years, neural-network-based AI systems have started to dramatically improve performance on LibriSpeech, lowering the word error rate (WER; 0% is optimal performance) to around 2% (Figure 2.5.1a and Figure 2.5.1b).

Developers can test out their systems on LibriSpeech in two ways:

- Test Clean determines how well their systems can transcribe speech from a higher-quality subset of the LibriSpeech dataset. This test gives clues about how well AI systems might perform in more controlled environments.
- Test Other determines how systems can deal with lower-quality parts of the LibriSpeech dataset. This test suggests how well AI systems might perform in noisier (and perhaps more realistic) environments.

There has been substantial progress recently on both datasets, with an important trend emerging in the past two years: The gap between performance on Test Clean and Test Other has started to close significantly for frontier systems, shifting from an absolute performance difference of more than seven points in late 2015 to a difference of less than one point in 2020. This reveals dramatic improvements in the robustness of ASR systems over time and suggests that we might be saturating performance on LibriSpeech—in other words, harder tests may be needed.

Speaker Recognition: VoxCeleb

Speaker identification tests how well machine learning systems can attribute speech to a particular person. The VoxCeleb dataset, first introduced in 2017, contains over a million utterances from 6,000 distinct speakers, and its associated speaker-identification task tests the error rate for systems that try to attribute a particular utterance to a particular speaker. A better (lower) score in VoxCeleb provides a proxy for how well a machine can distinguish one voice among 6,000. Evaluation method for VoxCeleb is Equal Error Rate (EER), a commonly used metric for identity verification systems. EER provides a measure for both the false positive rate (assigning a label incorrectly) and the false negative rate (failing to assign a correct label).

In recent years, progress on this task has come from hybrid systems—systems that fuse contemporary deep learning approaches with more structured algorithms, developed by the broader speech-processing community. As of 2020, error rates have dropped such that computers have a very high (99.4%) ability to attribute utterances to a given speaker (Figure 2.5.2)

Still, obstacles remain: These systems face challenges processing speakers with different accents and in differentiating among speakers when confronted with a large dataset (it is harder to identify one person in a set of a billion people than to pick out one person across the VoxCeleb training set of 6,000).

LIBRISPEECH: WORD ERROR RATE, TEST CLEAN

Source: Papers with Code, 2020 | Chart: 2021 AI Index Report

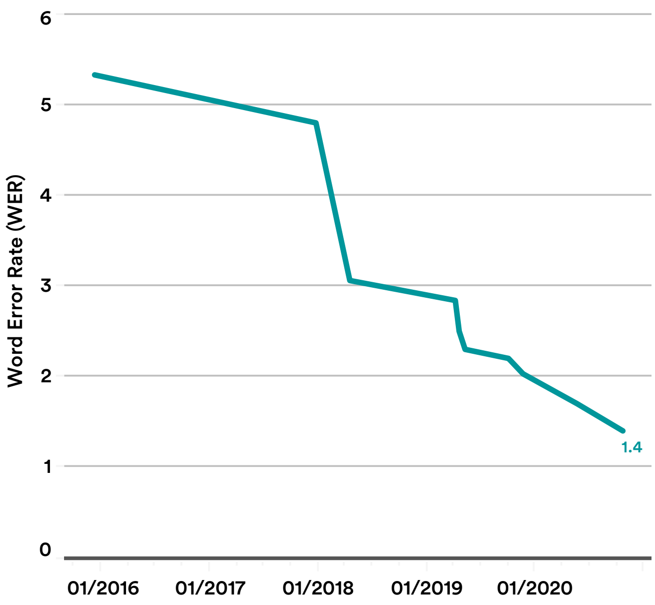

Figure 2.5.1a

LIBRISPEECH: WORD ERROR RATE, TEST OTHER

Source: Papers with Code, 2020 | Chart: 2021 AI Index Report

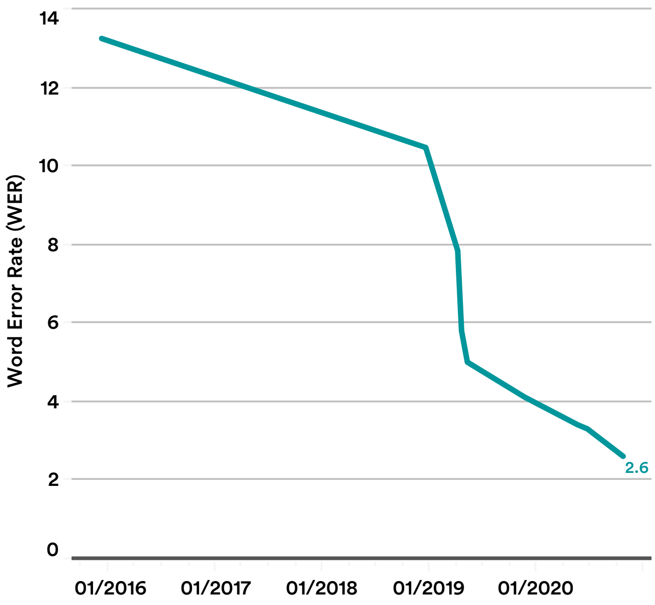

Figure 2.5.1b

VOXCELEB: EQUAL ERROR RATE

Source: VoxCeleb, 2020 | Chart: 2021 AI Index Report

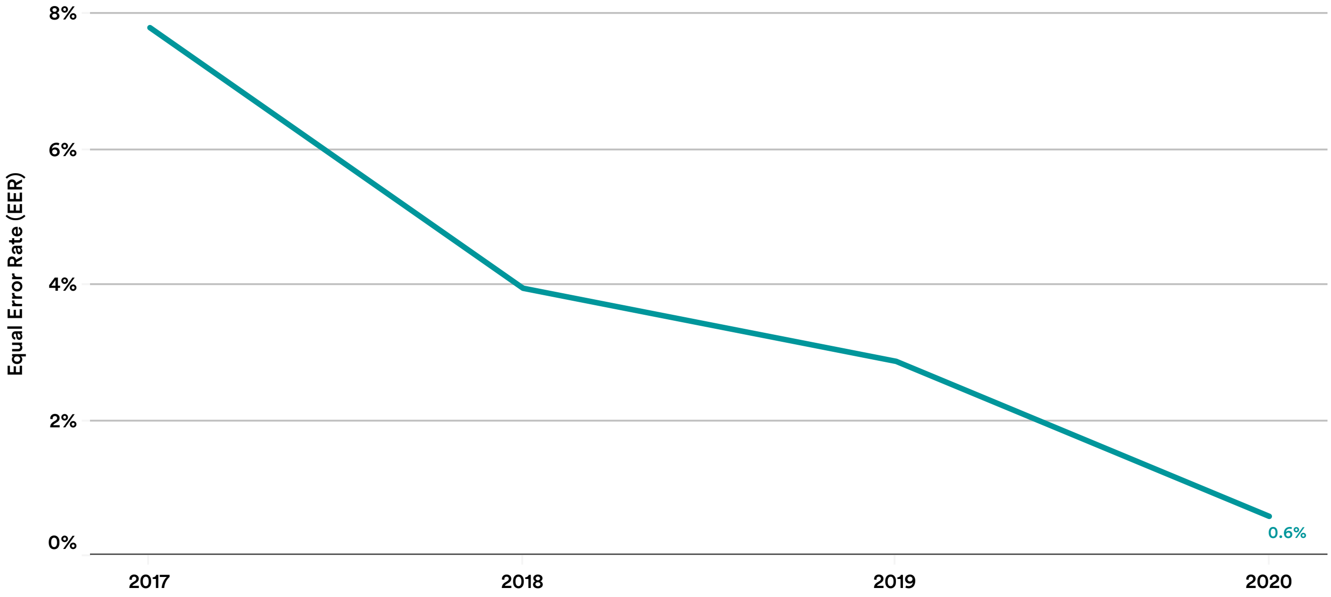

Figure 2.5.2

The Race Gap in Speech Recognition Technology

Researchers from Stanford University found that state-of-the-art ASR systems exhibited significant racial and gender disparity—they misunderstand Black speakers twice as often as white speakers. In the paper, titled "Racial Disparities in Automated Speech Recognition," authors ran thousands of audio snippets of white and Black speakers, transcribed from interviews conducted with 42 white speakers and 73 Black speakers, through leading speech-to-text services by Amazon, Apple, Google, IBM, and Microsoft.

The results suggest that, on average, systems made 19 errors every hundred words for white speakers and 35 errors for Black speakers—nearly twice as many. Moreover, the systems performed particularly poorly for Black men, with more than 40 errors for every hundred words (Figure 2.5.3). The breakdown by ASR systems shows that gaps are similar across companies (Figure 2.5.4). This research emphasizes the importance of addressing the bias of AI technologies and ensuring equity as they become mature and deployed.

TESTINGS on LEADING SPEECH-to-TEXT SERVICES: WORD ERROR RATE by RACE and GENDER, 2019

Source: Koenecke et al., 2020 | Chart: 2021 AI Index Report

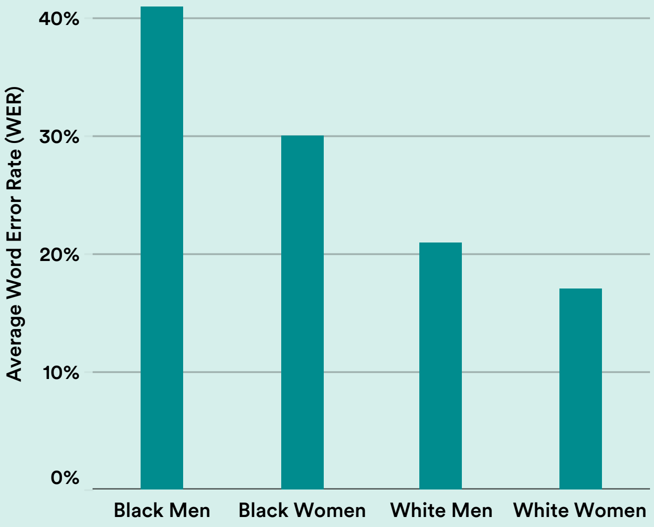

Figure 2.5.3

TESTINGS on LEADING SPEECH-to-TEXT SERVICES: WORD ERROR RATE by SERVICE and RACE, 2019

Source: Koenecke et al., 2020 | Chart: 2021 AI Index Report

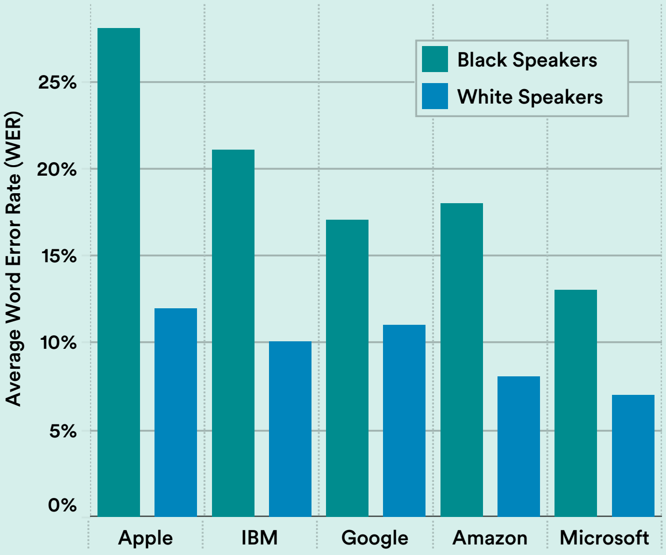

Figure 2.5.4

This section measures progress on symbolic (or logical) reasoning in AI, which is the process of drawing conclusions from sets of assumptions. We consider two major reasoning problems, Boolean Satisfiability (SAT) and Automated Theorem Proving (ATP). Each has real-world applications (e.g., circuit design, scheduling, software verification, etc.) and poses significant measurement challenges. The SAT analysis shows how to assign credit for the overall improvement in the field to individual systems over time. The ATP analysis shows how to measure performance given an evolving test set.

All analyses below are original to this report. Lars Kotthoff wrote the text and performed the analysis for the SAT section. Geoff Sutcliffe, Christian Suttner, and Raymond Perrault wrote the text and performed the analysis for the ATP section. This work had not been published at the time of writing; consequently, a more academically rigorous version of this section (with references, more precise details, and further context) is included in the Appendix.

2.6 REASONING

BOOLEAN SATISFIABILITY PROBLEM

Analysis and text by Lars Kotthoff

The SAT problem considers whether there is an assignment of values to a set of Boolean variables, joined by logical connectives, that makes the logical formula it represents true. Many real-world problems, such as circuit design, automated theorem proving, and scheduling, can be represented and solved efficiently as SAT problems.

The performance of the top-, median-, and bottom-ranked SAT solvers was examined from each of the last five years (2016–2020) of the SAT Competition, which has been running for almost 20 years, to measure a snapshot of state-of-the-art performance. In particular, all 15 solvers were run on all 400 SAT instances from the main track of the 2020 competition and the time (in CPU seconds) it took to solve all instances was measured.[5] Critically, each solver was run on the same hardware, such that comparisons across years would not be confounded by improvements in hardware efficiency over time.

While performance of the best solvers from 2016 to 2018 did not change significantly, large improvements are evident in 2019 and 2020 (Figure 2.6.1). These improvements affect not only the best solvers but also their competitors. The performance of the median-ranked solver in 2019 is better than that of the top-ranked solvers in all previous years, and the performance of the median-ranked solver in 2020 is almost on par with the top-ranked solver in 2019.

Performance improvements in SAT—and more generally, hard computational AI problems—come primarily from two areas of algorithmic improvements: novel techniques and more efficient implementations of existing techniques. Typically, performance improvements arise primarily from novel techniques. However, more efficient implementations (which can arise with performance improvements in hardware over time) can also increase performance. Therefore, it is difficult to assess whether performance improvements arise primarily from novel techniques or more efficient implementations. To address this problem, the temporal Shapley value, which is the contribution of an individual system to state-of-the-art performance over time, was measured (see the Appendix for more details).

Figure 2.6.2 shows the temporal Shapley value contributions of each solver for the different competition years. Note that the contributions of the solvers in 2016 are highest because there is no previous state-of-the-art to compare them with in our evaluation and that their contribution is not discounted.

[5] Acknowledgments: The Advanced Research Computing Center at the University of Wyoming provided resources for gathering the computational data. Austin Stephen performed the computational experiments.

TOTAL TIME to SOLVE ALL 400 INSTANCES for EACH SOLVER and YEAR (LOWER IS BETTER), 2016-20

Source: Kotthoff, 2020 | Chart: 2021 AI Index Report

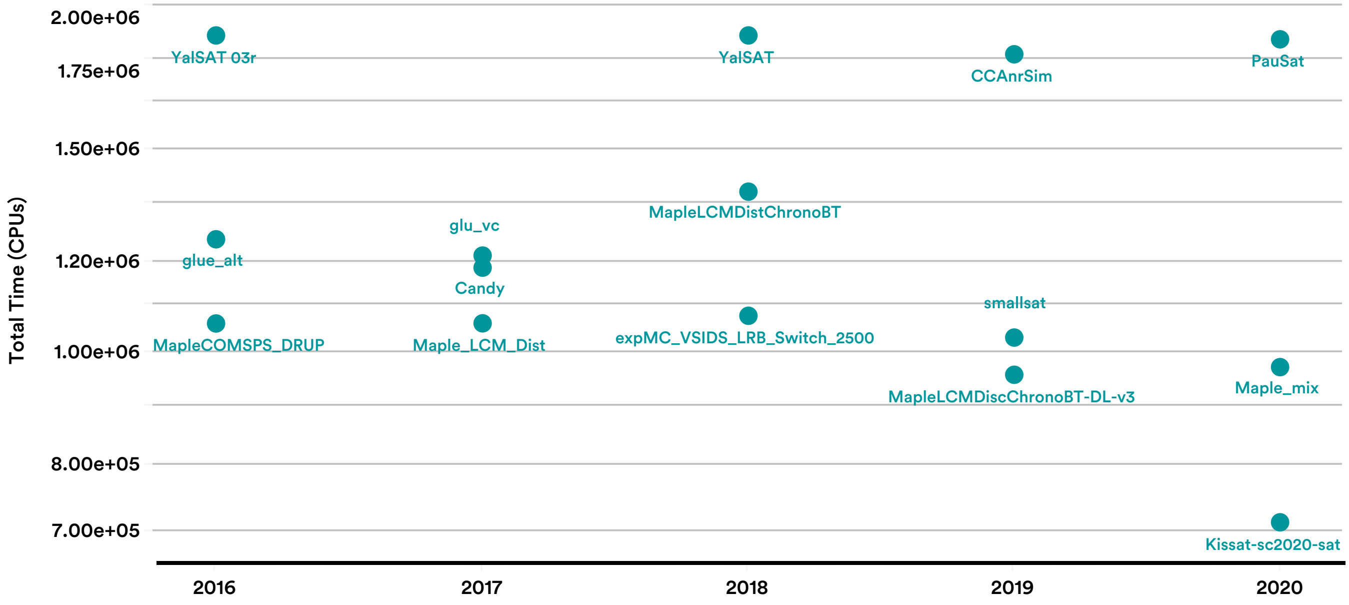

Figure 2.6.1

TEMPORAL SHAPLEY VALUE CONTRIBUTIONS of INDIVIDUAL SOLVERS to the STATE of the ART OVER TIME (HIGHER IS BETTER), 2016-20

Source: Kotthoff, 2020 | Chart: 2021 AI Index Report

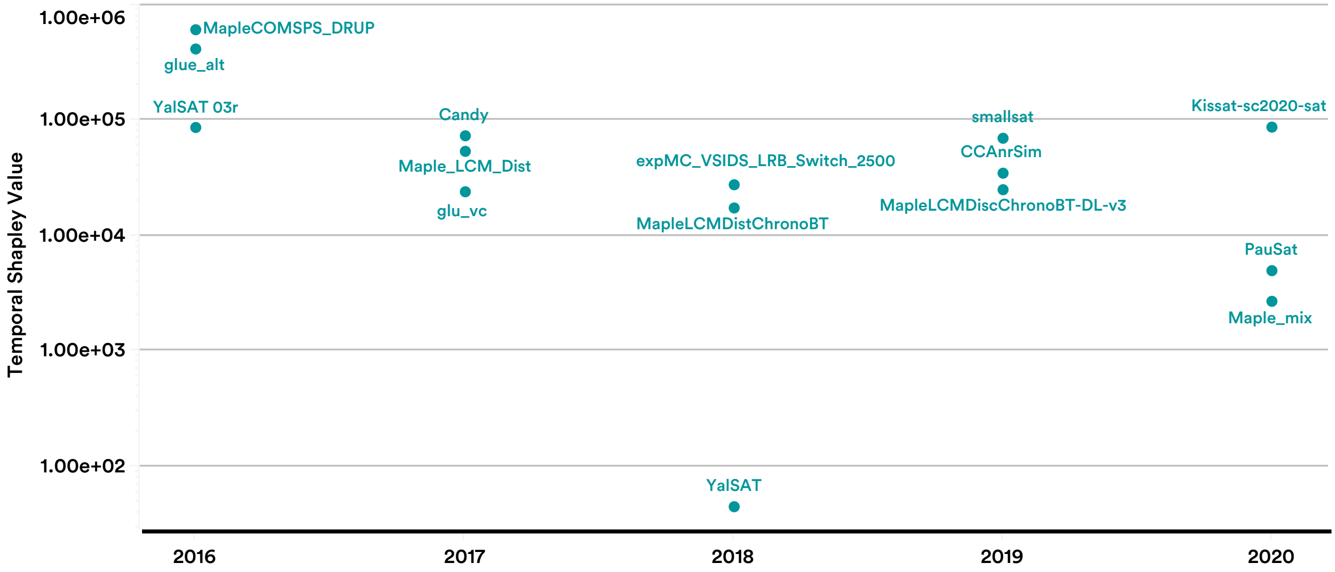

Figure 2.6.2

According to the temporal Shapley value, in 2020 the best solver contributes significantly more than the median- and bottom-ranked solvers do. The 2020 winner, Kissat, has the highest temporal Shapley value of any solvers excluding the first year. The changes it incorporates, compared with those of previous solvers, are almost exclusively more efficient data structures and algorithms; Kissat thus impressively demonstrates the impact of good engineering on the state-of-the-art performance.

By contrast, smallsat, the solver with the largest temporal Shapley value (but not the winner) in 2019, focuses on improved heuristics instead of a more efficient implementation. The same is true of Candy, the solver with the largest temporal Shapley value in 2017, whose main novelty is to analyze the structure of a SAT instance and apply heuristics based on this analysis. Interestingly, neither solver ranked first in their respective years; both were outperformed by versions of the Maple solver, which nevertheless contributes less to the state of the art. This indicates that incremental improvements, while not necessarily exciting, are important for good performance in practice.

Based on our limited analysis of the field, novel techniques and more efficient implementations have made equally important contributions to the state of the art in SAT solving. Incremental improvements of established solvers are as likely to result in top performance as more substantial improvements of solvers without a long track record.

AUTOMATED THEOREM PROVING (ATP)

Analysis and text by Christian Suttner, Geoff Sutcliffe, and Raymond Perrault

Automated Theorem Proving (ATP) concerns the development and use of systems that automate sound reasoning, or the derivation of conclusions that follow inevitably from facts. ATP systems are at the heart of many computational tasks, including software verification. The TPTP problem library was used to evaluate the performance of ATP algorithms from 1997 to 2020 and to measure the fraction of problems solved by any system over time (see the Appendix for more details).

The analysis extends to the whole TPTP (over 23,000 problems) in addition to four salient subsets (each ranging between 500 and 5,500 problems)—clause normal form (CNF), first-order form (FOF), monomorphic typed first-order form (TF0) with arithmetic, and monomorphic typed higher-order form (TH0) theorems—all including the use of the equality operator.

Figure 2.6.3 shows that the fraction of problems solved climbs consistently, indicating progress in the field. The noticeable progress from 2008 to 2013 included strong progress in the FOF, TF0, and TH0 subsets. In FOF, which has been used in many domains (e.g., mathematics, real-world knowledge, software verification), there were significant improvements in the Vampire, E, and iProver systems. In TF0 (primarily used for solving problems in mathematics and computer science) and TH0 (useful in subtle and complex topics such as philosophy and logic), there was rapid initial progress as systems developed techniques that solved "low-hanging fruit" problems. In 2014–2015, there was another burst of progress in TF0, as the Vampire system became capable of processing TF0 problems. It is noteworthy that, since 2015, progress has continued but slowed, with no indication of rapid advances or breakthroughs in the last few years.

PERCENTAGE of PROBLEMS SOLVED, 1997-2020

Source: Sutcliffe, Suttner & Perrault, 2020 | Chart: 2021 AI Index Report

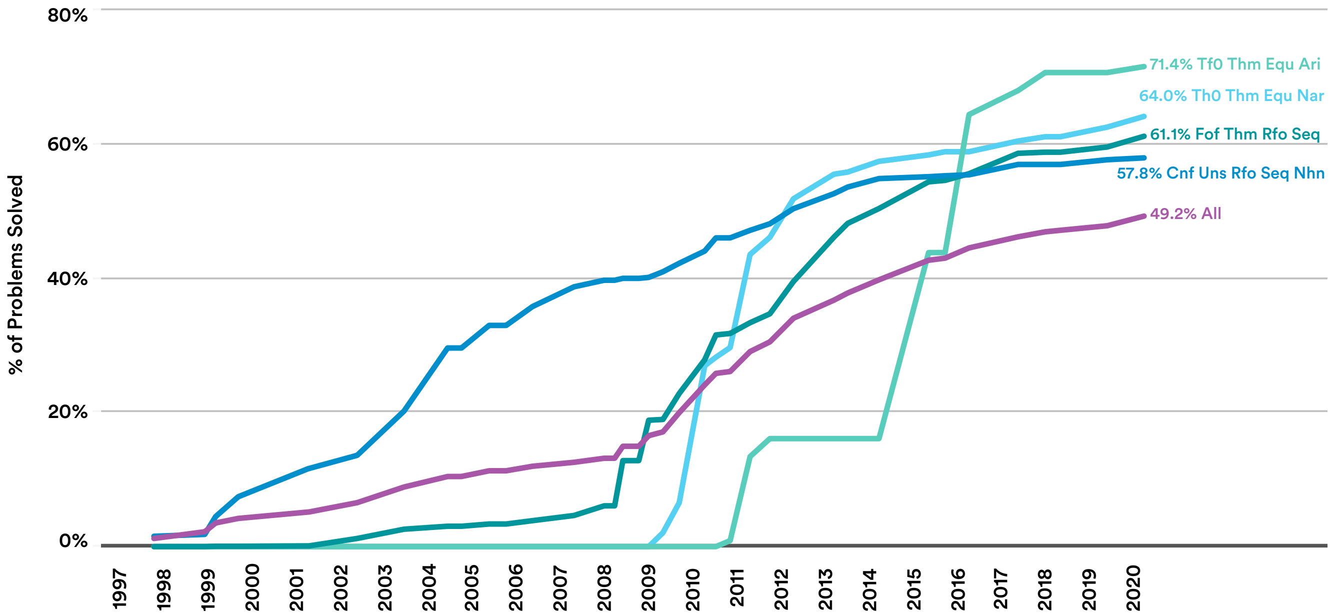

Figure 2.6.3

While this analysis demonstrates progress in ATP, there is obviously room for much more. Two keys to solve ATP problems are axiom selection (given a large set of axioms, only some of which are needed for a proof of the conjecture, how to select an adequate subset of the axioms); and search choice (at each stage of an ATP system's search for a solution, which logical formula(e) should be selected for attention). The latter issue has been at the forefront of ATP research since its inception in the 1960s, while the former has become increasingly important as large bodies of knowledge are encoded for ATP. In the last decade, there has been growing use of machine learning approaches to addressing these two key challenges (e.g., in the MaLARea and Enigma ATP systems). Recent results from the CADE ATP System Competition (CASC) have shown that the emergence of machine learning is a potential game-changer for ATP.

2.7 HEALTHCARE AND BIOLOGY

In collaboration with the "State of AI Report"

MOLECULAR SYNTHESIS

Text by Nathan Benaich and Philippe Schwaller

Over the last 25 years, the pharmaceutical industry has shifted from developing drugs from natural sources (e.g., plants) to conducting large-scale screens with chemically synthesized molecules. Machine learning allows scientists to determine what potential drugs are worth evaluating in the lab and the most effective way of synthesizing them. Various ML models can learn representations of chemical molecules for the purposes of chemical synthesis planning.

A way to approach chemical synthesis planning is to represent chemical reactions with a text notation and cast the task as a machine translation problem. Recent work since 2018 makes use of the transformer architecture trained on large datasets of single-step reactions. Later work in 2020 approached model forward prediction and retrosynthesis as a sequence of graph edits, where the predicted molecules were built from scratch.

Notably, these approaches offer an avenue to rapidly sweep through a list of candidate drug-like molecules in silico and output synthesizability scores and synthesis plans. This enables medicinal chemists to prioritize candidates for empirical validation and could ultimately let the pharmaceutical industry mine the vast chemical space to unearth novel drugs to benefit patients.

Test Set Accuracy for Forward Chemical Synthesis Planning

Figure 2.7.1 shows the top-1 accuracy of models benchmarked on a freely available dataset of one million reactions in the U.S. patents.[6] Top-1 accuracy means that the product predicted by the model with the highest likelihood corresponds to the one that was reported in the ground truth. Data suggests that progress in chemical synthesis planning has seen steady growth in the last three years, as the accuracy grew by 15.6% in 2020 from 2017. The latest molecular transformer scored 92% on top-1 accuracy in November 2020.

CHEMICAL SYNTHESIS PLANS BENCHMARK: TOP-1 TEST ACCURACY

Source: Schwaller, 2020 | Chart: 2021 AI Index Report

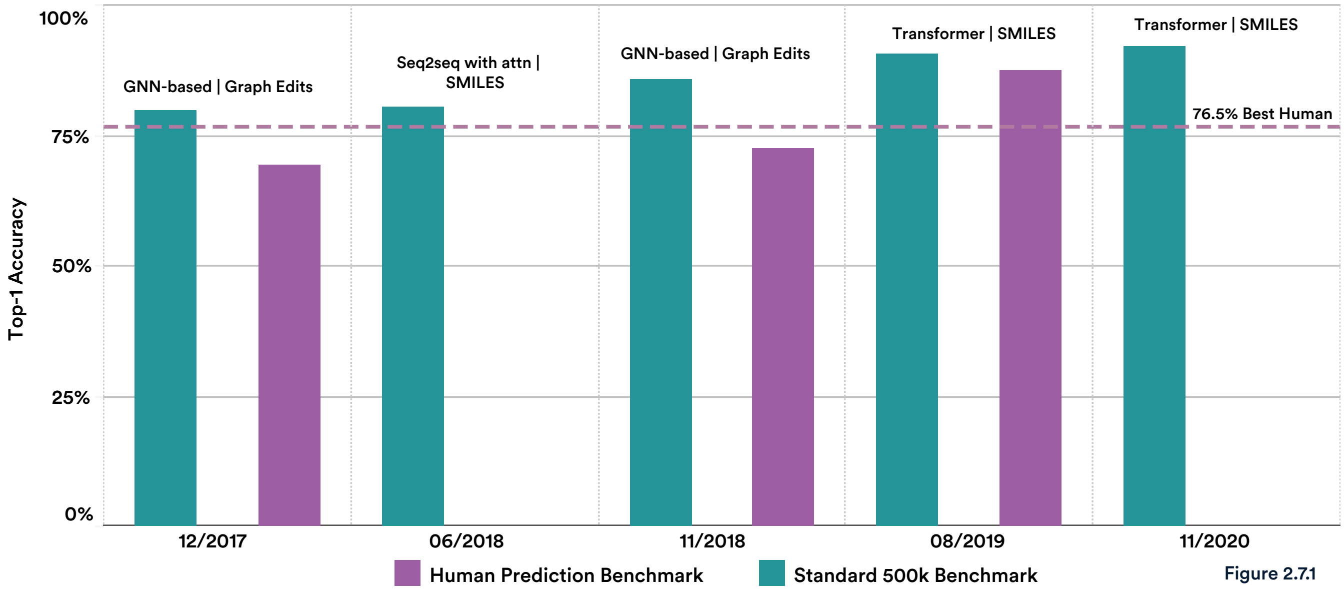

Figure 2.7.1

[6] Acknowledgment: Philippe Schwaller at IBM Research–Europe and the University of Bern provided instructions and resources for gathering and analyzing the data.

COVID-19 AND DRUG DISCOVERY

AI-powered drug discovery has gone open source to combat the COVID-19 pandemic. COVID Moonshot is a crowdsourced initiative joined by over 500 international scientists to accelerate the development of a COVID-19 antiviral. The consortium of scientists submits their molecular designs pro bono, with no claims. PostEra, an AI startup, uses machine learning and computational tools to assess how easily compounds can be made using the submissions from the scientists and generates synthetic routes. After the first week, Moonshot received over 2,000 submissions, and PostEra designed synthetic routes in under 48 hours. Human chemists would have taken three to four weeks to accomplish the same task.

Figure 2.7.2 shows the accumulated number of submissions by scientists over time. Moonshot received over 10,000 submissions from 365 contributors around the world in just four months. Toward the end of August 2020, the crowdsourcing had served its purpose, and the emphasis moved to optimize the lead compounds and set up for animal testing. As of February 2021, Moonshot aims to nominate a clinical candidate by the end of March.

POSTERA: TOTAL NUMBER of MOONSHOT SUBMISSIONS
Source: PostEra, 2020 | Chart: 2021 AI Index Report

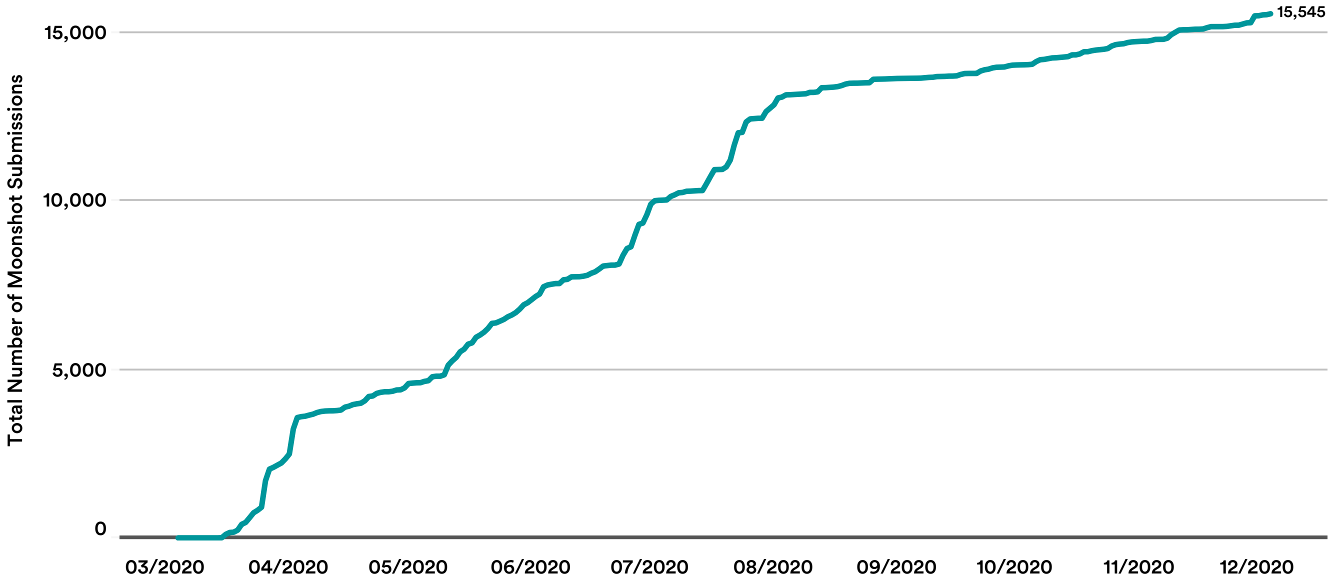

Figure 2.7.2

ALPHAFOLD AND PROTEIN FOLDING

The protein folding problem, a grand challenge in structural biology, considers how to determine the three-dimensional structure of proteins (essential components of life) from their one-dimensional representations (sequences of amino acids[7]). A solution to this problem can have wide ranging applications—from better understanding the cellular basis of life, to fueling drug discovery, to curing diseases, to engineering de-novo proteins for industrial tasks, and more.

In recent years, machine learning-based approaches have started to make a meaningful difference on the protein folding problem. Most notably, DeepMind's AlphaFold debuted in 2018 at the Critical Assessment of Protein Structure (CASP) competition, a biennial competition to foster and measure progress on protein folding. At CASP, competing teams are given amino acid sequences and tasked to predict the three-dimensional structures of the corresponding proteins, the latter of which are determined through laborious and expensive experimental methods (e.g., nuclear magnetic resonance spectroscopy, X-ray crystallography, cryo-electron microscopy, etc.) and unknown to the competitors. Performance on CASP is commonly measured by the Global Distance Test (GDT) score, a number between 0 and 100, which measures the similarity between two protein structures. A higher GDT score is better.

Figure 2.7.3, adapted from the DeepMind blog post, shows the median GDT scores of the best team on some of the harder types of proteins to predict (the 'free-modelling' category of proteins) at CASP over the last 14 years. In the past, winning algorithms were typically based on physics based models; however, in the last two competitions, Deepmind's AlphaFold and AlphaFold 2 algorithms achieved winning scores through the partial incorporation of deep learning techniques.

CASP: MEDIAN ACCURACY of PREDICTIONS in FREE-MODELING by THE BEST TEAM, 2006-20
Source: DeepMind, 2020 | Chart: 2021 AI Index Report

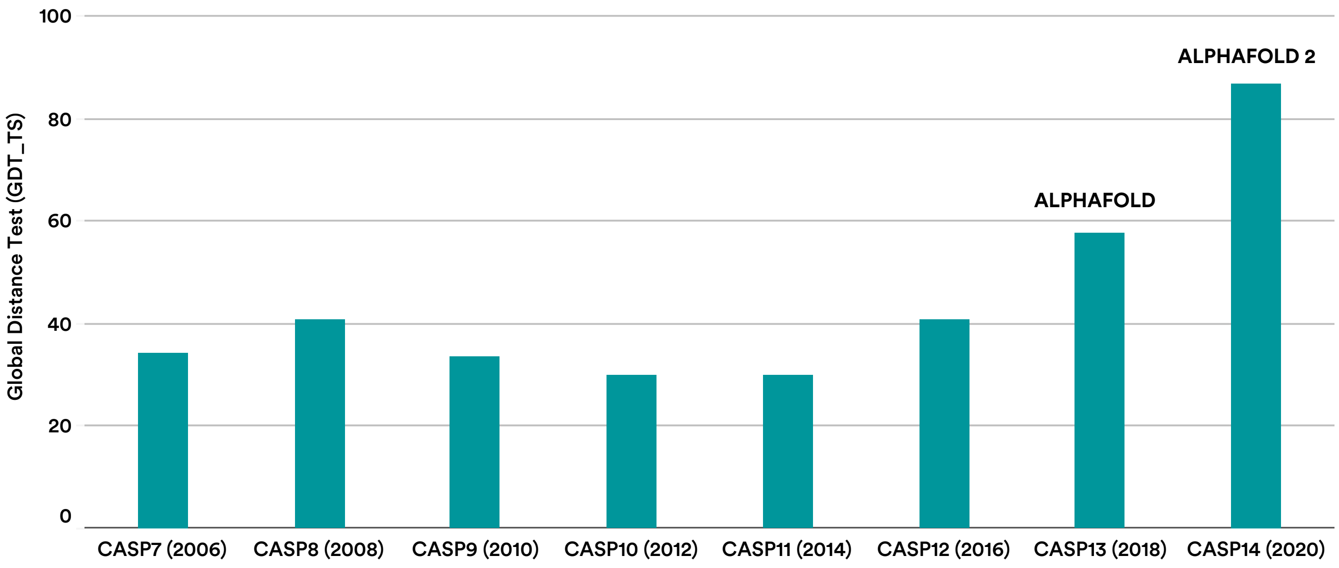

Figure 2.7.3

[7] Currently most protein folding algorithms leverage multiple sequence alignments—many copies of a protein sequence representing the same protein across evolution—rather than just a single sequence.

EXPERT HIGHLIGHTS

This year, the AI Index asked AI experts to share their thoughts on the most significant technical AI breakthroughs in 2020. Here's a summary of their responses, along with a couple of individual highlights.

What was the single most impressive AI advancement in 2020?

- The two most mentioned systems by a significant margin were AlphaFold (DeepMind), a model for molecular assay, and GPT-3 (OpenAI), a generative text model.

What single trend will define AI in 2021?

- Experts predict that more advances will be built by using pretrained models. For instance, GPT-3 is a large NLP model that can subsequently be fine-tuned for excellent performance on specific, narrow tasks. Similarly, 2020 saw various computer vision advancements built on top of models pretrained on very large image datasets.

What aspect of AI technical progress, deployment, and development are you most excited to see in 2021?

- "It's interesting to note the dominance of the Transformers architecture, which started for machine translation but has become the de facto neural network architecture. More broadly, whereas NLP trailed vision in terms of adoption of deep learning, now it seems like advances in NLP are also driving vision." — Percy Liang, Stanford University

- "The incredible recent advancements in language generation have had a profound effect on the fields of NLP and machine learning, rendering formerly difficult research challenges and datasets suddenly useless while simultaneously encouraging new research efforts into the fascinating emergent capabilities (and important failings) of these complex new models." —Carissa Schoenick, Allen Institute of AI Research

APPENDIX

IMAGENET: ACCURACY

Prepared by Jörg Hellwig and Thomas A. Collins

Source

Data on ImageNet accuracy was retrieved through an arXiv literature review. All results reported were tested on the LSRVC 2012 validation set, as the results on the test set, which are not significantly different, are not public. Their ordering may differ from the results reported on the LSRVC website, since those results were obtained on the test set. Dates we report correspond to the day when a paper was first published to arXiv, and top-1 accuracy corresponds to the result reported in the most recent version of each paper. We selected a top result at any given point in time from 2012 to Nov. 17, 2019. Some of the results we mention were submitted to LSRVC competitions over the years. Image classification was part of LSRVC through 2014; in 2015, it was replaced with an object localization task, where results for classification were still reported but no longer a part of the competition, having instead been replaced by more difficult tasks.

For papers published in 2014 and later, we report the best result obtained using a single model (we did not include ensembles) and using single-crop testing. For the three earliest models (AlexNet, ZFNet, Five Base), we reported the results for ensembles of models.

While we report the results as described above, due to the diversity in models, evaluation methods, and accuracy metrics, there are many other ways to report ImageNet performance. Some possible choices include:

- Evaluation set: validation set (available publicly) or test set (available only to LSRVC organizers)
- Performance metric: Top-1 accuracy (whether the correct label was the same as the first predicted label for each image) or top-5 accuracy (whether the correct label was present among the top five predicted labels for each image)
- Evaluation method: single-crop or multi-crop

To highlight progress here in top-5 accuracy, we have taken scores from the following papers, without extra training data:
Fixing the Train-Test Resolution Discrepancy: FixEfficientNet
Adversarial Examples Improve Image Recognition
OverFeat: Integrated Recognition, Localization and Detection Using Convolutional Networks
Local Relation Networks for Image Recognition
Densely Connected Convolutional Networks
Revisiting Unreasonable Effectiveness of Data in Deep Learning Era
Squeeze-and-Excitation Networks
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks
MultiGrain: A Unified Image Embedding for Classes and Instances
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks
Billion-Scale Semi-Supervised Learning for Image Classification
GPipe: Efficient Training of Giant Neural Networks Using Pipeline Parallelism
RandAugment: Practical Data Augmentation with No Separate Search
Fixing the Train-Rest Resolution Discrepancy

To highlight progress here in top-5 accuracy, we have taken scores from the following papers, with extra training data:
Meta Pseudo Labels
Self-Training with Noisy Student Improves ImageNet Classification
Big Transfer (BiT): General Visual Representation Learning
ImageNet Classification with Deep Convolutional Neural Networks
ESPNetv2: A Light-Weight, Power Efficient, and General Purpose Convolutional Neural Network
Xception: Deep Learning with Depthwise Separable Convolutions
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks
Self-training with Noisy Student Improves ImageNet Classification

To highlight progress here in top-1 accuracy, we have taken scores from the following papers, without extra training data:
Fixing the Train-Test Resolution Discrepancy: FixEfficientNet
Adversarial Examples Improve Image Recognition
OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks
Densely Connected Convolutional Networks
Revisiting Unreasonable Effectiveness of Data in Deep Learning Era
Dual Path Networks
Res2Net: A New Multi-Scale Backbone Architecture
Billion-Scale Semi-Supervised Learning for Image Classification
Squeeze-and-Excitation Networks
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks
MultiGrain: A Unified Image Embedding for Classes and Instances
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks
Billion-Scale Semi-Supervised Learning for Image Classification
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks
RandAugment: Practical Data Augmentation with No Separate Search
Fixing the Train-Test Resolution Discrepancy

To highlight progress here in top-1 accuracy, we have taken scores from the following papers, without extra training data:
Meta Pseudo Labels
Sharpness-Aware Minimization for Efficiently Improving Generalization
An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale
Fixing the Train-Test Resolution Discrepancy: FixEfficientNet
Self-training with Noisy Student Improves ImageNet Classification
Big Transfer (BiT): General Visual Representation Learning
ImageNet Classification with Deep Convolutional Neural Networks
ESPNetv2: A Light-Weight, Power Efficient, and General Purpose Convolutional Neural Network
Xception: Deep Learning with Depthwise Separable Convolutions
EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks
Self-training with Noisy Student Improves ImageNet Classification

The estimate of human-level performance is from Russakovsky et al, 2015. Learn more about the LSVRC ImageNet competition and the ImageNet data set.

IMAGENET: TRAINING TIME

Trends can also be observed by studying research papers that discuss the time it takes to train ImageNet on *any* infrastructure. To gather this data, we looked at research papers from the past few years that tried to optimize for training ImageNet to a standard accuracy level while competing on reducing the overall training time.

Source

The data is sourced from MLPerf. Detailed data for runs for specific years are available:
2020: MLPerf Training v0.7 Results
2019: MLPerf Training v0.6 Results
2018: MLPerf Training v0.5 Results

Notes

Data from MLPerf is available in cloud systems for rent. Available On Premise systems contain only components that are available for purchase. Preview systems must be submittable as Available In Cloud or Available on Premise in the next submission round. Research, Development, or Internal (RDI) contain experimental, in development, or internal-use hardware or software. Each row in the results table is a set of results produced by a single submitter using the same software stack and hardware platform. Each row contains the following information:

Submitter: the organization that submitted the results
System: general system description
Processor and count: the type and number of CPUs used, if CPUs perform the majority of ML compute
Accelerator and count: the type and number of accelerators used, if accelerators perform the majority of ML compute
Software: the ML framework and primary ML hardware library used
Benchmark results: training time to reach a specified target quality, measured in minutes
Details: link to metadata for submission
Code: link to code for submission
Notes: arbitrary notes from the submitter

IMAGENET: TRAINING COST

Source

DAWNBench is a benchmark suite for end-to-end, deep-learning training and inference. Computation time and cost are critical resources in building deep models, yet many existing benchmarks focus solely on model accuracy. DAWNBench provides a reference set of common deep-learning workloads for quantifying training time, training cost, inference latency, and inference cost across different optimization strategies, model architectures, software frameworks, clouds, and hardware. More details available at DawnBench.

Note

The DawnBench data source has been deprecated for the period after March 2020, and MLPerf is the most reliable and updated source for AI compute measurements.

COCO: KEYPOINT DETECTION

The data for COCO keypoint detection data is sourced from COCO keypoints leaderboard.

COCO: DENSEPOSE ESTIMATION

We gathered data from the CODALab 2020 challenge and read arXiv repository papers to build comprehensive data on technical progress in this challenge. The detailed list of papers and sources used in our survey include:
DensePose: Dense Human Pose Estimation In the Wild
COCO-DensePose 2018 CodaLab
Parsing R-CNN for Instance-Level Human Analysis
Capture Dense: Markerless Motion Capture Meets Dense Pose Estimation
Slim DensePose: Thrifty Learning from Sparse Annotations and Motion Cues
COCO-DensePose 2020 CodaLab
Transferring Dense Pose to Proximal Animal Classes
Making DensePose Fast and Light
SimPose: Effectively Learning DensePose and Surface Normals of People from Simulated Data

ACTIVITYNET: TEMPORAL LOCALIZATION TASK

In the challenge, there are three separate tasks, but they focus on the main problem of temporally localizing where activities happen in untrimmed videos from the ActivityNet benchmark. We have compiled several attributes for the task of temporal localization at the challenge over the last four rounds. Below is a link to the overall stats and trends for this task, as well as some detailed analysis (e.g., how has the performance for individual activity classes improved over the years? Which are the hardest and easiest classes now? Which classes have the most improvement over the years?). See the Performance Diagnosis (2020) tab for a detailed trends update. Please see ActivityNet Statisticsin the public data folder for more details.

YOLO (YOU ONLY LOOK ONCE)

YOLO is a neural network model mainly used for the detection of objects in images and in real-time videos. mAP (mean average precision) is a metric that is used to measure the accuracy of object detectors. It is a combination of precision and recall. mAP is the average of the precision and recall calculated over a document. The performance of YOLO has increased gradually with the development of new architectures and versions in past years. With the increase in size of model, its mean average precision increases as well, with a corresponding decrease in FPS of the video.

We conducted a detailed survey of arXiv papers and GitHub repository to segment progress in YOLO across its various versions. Below are the references for original sources:

YOLOv1:
You Only Look Once: Unified, Real-Time Object Detection

YOLOv2:
YOLO9000: Better, Faster, Stronger
YOLO: Real-Time Object Detection

YOLOv3:
YOLOv3: An Incremental Improvement
Learning Spatial Fusion for Single-Shot Object Detection
GitHub: ultralytics/yolov3

YOLOv4:
YOLOv4: Optimal Speed and Accuracy of Object Detection
GitHub: AlexeyAB/darknet

YOLOv5:
GitHub: ultralytics/yolov5

PP-YOLO:
PP-YOLO: An Effective and Efficient Implementation of Object Detector

POLY-YOLO:
Poly-YOLO: Higher Speed, More Precise Detection and Instance Segmentation for YOLOV3

VISUAL QUESTION ANSWERING (VQA)

VQA accuracy data was provided by the VQA team. Learn more about VQA here. More details on VQA 2020 are available here.

Methodology

Given an image and a natural language question about the image, the task is to provide an accurate natural language answer. The challenge is hosted on the VQA Challenge website. The challenge is hosted on EvalAI. The challenge link is here.

The VQA v2.0 training, validation, and test sets, containing more than 250,000 images and 1.1 million questions, are available on the download page. All questions are annotated with 10 concise, open-ended answers each. Annotations on the training and validation sets are publicly available.

VQA Challenge 2020 is the fifth edition of the VQA Challenge. Results from previous versions of the VQA Challenge were announced at the VQA Challenge Workshop in CVPR 2019, CVPR 2018, CVPR 2017, and CVPR 2016. More details about past challenges can be found here: VQA Challenge 2019, VQA Challenge 2018, VQA Challenge 2017, VQA Challenge 2016.

VQA had 10 humans answer each question. More details about the VQA evaluation metric and human accuracy can be found here (see Evaluation Code section) and in sections three ("Answers") and four ("Inter-Human Agreement") of the paper.

See slide 56 for the progress graph in VQA in the 2020 Challenge. The values corresponding to the progress graph are available in a sheet. Here is the information about the teams that participated in the 2020 challenge and their accuracies. For more details about the teams, please refer to the VQA website.

PAPERS WITH CODE: PAPER AND CODE LINKING

We used paperswithcode (PWC) for referencing technical progress where available. Learn more about PWC here and see the public link here.

Methodology

For papers, we follow specific ML-related categories on arxiv (see [1] below for the full list) and the major ML conferences (NeurIPS, ICML, ICLR, etc.). For code, we follow GitHub repositories mentioning papers. We have good coverage of core ML topics but are missing some applications—for instance, applications of ML in medicine or bioinformatics, which are usually in journals behind paywalls. For code, the dataset is fairly unbiased (as long as the paper is freely available).

For tasks (e.g., "image classification"), the dataset has annotated those on 1,600 state-of-the-art papers from the database, published in 2018 Q3.

For state-of-the-art tables (e.g., "image classification on ImageNet"), the data has been scraped from different sources (see the full list here), and a large number focusing on CV and NLP were hand-annotated. A significant portion of our data was contributed by users, and they have added data based on their own preferences and interests. Arxiv categories we follow:
ARXIV_CATEGORIES = "cs.CV", "cs.AI", "cs.LG", "cs.CL", "cs. NE", "stat.ML","cs.IR"}

Process of Extracting Dataset at Scale

1) Follow various paper sources (as described above) for new papers.
2) Conduct a number of predefined searches on GitHub (e.g., for READMEs containing links to arxiv).
3) Extract GitHub links from papers.
4) Extract paper links from GitHub.
5) Run validation tests to decide if links from 3) and 4) are bona fide links or false positives.
6) Let the community fix any errors and/or add any missing values.

NIST FRVT

Source

There are two FRVT evaluation leaderboards available here: 1:1 Verification and 1:N Identification

Nuances about FRVT evaluation metrics

Wild Photos have some identity labeling errors as the best algorithm has a low false non-match rate (FNMR), but obtaining complete convergence is difficult. This task will be retired in the future. The data became public in 2018 and has become easier over time. Wild is coming from public web sources. So it is possible those same images have been scrapped from the web by developers. There is no training in the FRVT data, only test data.

The 1:1 and 1:N should be studied separately. The differences include algorithmic approaches, particularly fast search algorithms are especially useful in 1:N whereas speed is not a factor in 1:1.

SUPERGLUE

The SuperGLUE benchmark data was pulled from the SuperGLUE leaderboard. Details about the SuperGLUE benchmark are in the SuperGLUE paper and SuperGLUE software toolkit. The tasks and evaluation metrics for SuperGLUE are:

NAME	IDENTIFIER	METRIC
Broad Coverage Diagnostics	AX-b	Matthew's Corr
CommitmentBank	CB	Avg. F1 / Accuracy
Choice of Plausible Alternatives	COPA	Accuracy
Multi-Sentence Reading Comprehension	MultiRC	F1a / EM
Recognizing Textual Entailment	RTE	Accuracy
Words in Context	WiC	Accuracy
The Winograd Schema Challenge	WSC	Accuracy
BoolQ	BoolQ	Accuracy
Reading Comprehension with Commonsense Reasoning	ReCoRD	F1 / Accuracy
Winogender Schema Diagnostics	AX-g	Gender Parity / Accuracy

VISUAL COMMONSENSE REASONING (VCR)

Technical progress for VCR is taken from the VCR leaderboard. VCR has two different subtasks:

- Question Answering (Q->A): A model is provided a question and has to pick the best answer out of four choices. Only one of the four is correct.
- Answer Justification (QA->R): A model is provided a question, along with the correct answer, and it must justify it by picking the best rationale among four choices.

The two parts with the Q->AR metrics are combined in which a model only gets a question right if it answers correctly and picks the right rationale. Models are evaluated in terms of accuracy (%).

VOXCELEB

VoxCeleb is an audio-visual dataset consisting of short clips of human speech, extracted from interview videos uploaded to YouTube. VoxCeleb contains speech from 7,000-plus speakers spanning a wide range of ethnicities, accents, professions, and ages—amounting to over a million utterances (face-tracks are captured "in the wild," with background chatter, laughter, overlapping speech, pose variation, and different lighting conditions) recorded over a period of 2,000 hours (both audio and video). Each segment is at least three seconds long. The data contains an audio dataset based on celebrity voices, shorts, films, and conversational pieces (e.g., talk shows). The initial VoxCeleb 1 (100,000 utterances taken from 1,251 celebrities on YouTube) was expanded to VoxCeleb 2 (1 million utterances from 6,112 celebrities).

However, in earlier years of the challenge, top-1 and top-5 scores were also reported. For top-1 score, the system is correct if the target label is the class to which it assigns the highest probability. For top-5 score, the system is correct if the target label is one of the five predictions with the highest probabilities. In both cases, the top score is computed as the number of times a predicted label matches the target label, divided by the number of data points evaluated.

The data is extracted from different years of the submission challenges, including:

- 2017: VoxCeleb: A Large-Scale Speaker Identification Dataset
- 2018: VoxCeleb2: Deep Speaker Recognition
- 2019: Voxceleb: Large-Scale Speaker Verification in the Wild
- 2020: Query ExpansionSystem for the VoxCeleb Speaker Recognition Challenge 2020

BOOLEAN SATISFIABILITY PROBLEM

Analysis and text by Lars Kotthoff

Primary Source and Data Sets

The Boolean Satisfiability Problem (SAT) determines whether there is an assignment of values to a set of Boolean variables joined by logical connectives that makes the logical formula it represents true. SAT was the first problem to be proven NP-complete, and the first algorithms to solve it were developed in the 1960s. Many real-world problems, such as circuit design, automated theorem proving, and scheduling, can be represented and solved efficiently as SAT. The annual SAT competition is designed to present a snapshot of the state-of-the-art and has been running for almost 20 years.

We took the top-ranked, median-ranked, and bottom-ranked solvers from each of the last five years (2016-2020) of the SAT competition. We ran all 15 solvers on all 400 SAT instances from the main track of the 2020 competition. More information on the competition, as well as the solvers and instances, is available at the SAT competition website.

Results

We ran each solver on each instance on the same hardware, with a time limit of 5,000 CPU seconds per instance, and measured the time it took a solver to solve an instance in CPU seconds. Ranked solvers always return correct results, hence we do not consider correctness as a metric. Except for the 2020 competition solvers, we evaluated the performance of the SAT solvers on a set of instances different from the set of instances they competed on. Further, our hardware is different from what was used for the SAT competition. The results we report here will therefore differ from the exact results reported for the respective SAT competitions.

The Shapley value is a concept from cooperative game theory that assigns a contribution to the total value that a coalition generates to each player. It quantifies how important each player is for the coalition and has several desirable properties that make the distribution of the total value to the individual players fair. For example, the Shapley value is used to distribute airport costs to its users, allocate funds to different marketing campaigns, and in machine learning, where it helps render complex black-box models more explainable.

In our context, it quantifies the contribution of a solver to the state-of-the-art through the average performance improvement it provides over a set of other solvers and over all subsets of solvers (Fréchette et al. (2016)). For a given set of solvers, we choose the respective best for each instance to solve. By including another solver and being able to choose it, overall solving performance improves, with the difference to the original set of solvers being the marginal contribution of the added solver. The average marginal contribution to all sets of solvers is the Shapley value.

Quantifying the contribution of a solver through the Shapley value compares solvers from earlier competitions to solvers in later competitions. This is often not a fair comparison, as later solvers are often improved versions of earlier solvers, and the contribution of the solver to the future state-of-the-art will always be low. The temporal Shapley value (Kotthoff et al. (2018)) solves this problem by considering the time a particular solver was introduced when quantifying its contribution to the state-of-the-art.

AUTOMATED THEOREM PROVING

Analysis and text by Christian Suttner, Geoff Sutcliffe, and Raymond Perrault

1. Motivation

Automated Theorem Proving (ATP) (also referred to as Automated Deduction) is a subfield of automated reasoning, concerned with the development and use of systems that automate sound reasoning: the derivation of conclusions that follow inevitably from facts. ATP systems are at the heart of many computational tasks and are used commercially, e.g., for integrated circuit design and computer program verification. ATP problems are typically solved by showing that a conjecture is or is not a logical consequence of a set of axioms. ATP problems are encoded in a chosen logic, and an ATP system for that logic is used to (attempt to) solve the problem. A key concern of ATP research is the development of more powerful systems, capable of solving more difficult problems within the same resource limits. In order to assess the merits of new techniques, sound empirical evaluations of ATP systems are key.

2. Analysis

For the evaluation of ATP systems, there exists a large and growing collection of problems called the TPTP problem library. The current release v7.4.0 (released June 10, 2020) contains 23,291 ATP problems, structured into 54 topic domains (e.g., Set Theory, Software Verification, Philosophy, etc.). Orthogonally, the TPTP is divided into Specialist Problem Classes (SPCs), each of which contains problems with a specified set of logical, language, and syntactic characteristics (e.g. first-order logic theorems with some use of equality). The SPCs allow ATP system developers to select problems and evaluate their systems appropriately. Since its first release in 1993, many researchers have used the TPTP as an appropriate and convenient basis for ATP system evaluation. Over the years, the TPTP has also increasingly been used as a conduit for ATP users to contribute samples of their problems to ATP system developers. This exposes the problems to ATP system developers, who can then improve their systems' performances on the problems, which completes a cycle to provide users with more effective tools.

Associated with the TPTP is the TSTP solution library, which maintains updated results from running all current versions of ATP systems (available to the maintainer) on all the TPTP problems. One use of the TSTP is to compute TPTP problem difficulty ratings: Easy problems, which are solved by all ATP systems, have a rating of 0.0; difficult problems, which are solved by some ATP systems, have ratings between 0.0 and 1.0; unsolved problems, which are not solved by any ATP system, have a rating of 1.0. Note that the rating for a problem is not strictly decreasing, as different ATP systems and versions become available for populating the TSTP. The history of each TPTP problem's ratings is saved with the problem, which makes it possible to tell when the problem was first solved by any ATP system (the point at which its rating dropped below 1.0). That information has been used here to obtain an indication of progress in the field.

The simplest way to measure progress takes a fixed set of problems that has been available (and unchanged) in the TPTP from some chosen initial TPTP release, and then for the TPTP releases from then on, counts how many of the problems had been solved from that release. The analysis reports the fraction of problems solved for each release. This simple approach is unambiguous, but it does not take into account new problems that are added to the TPTP after the initial release.

The analysis used here extends the “Fixed Set” analysis, taking into account new problems added after the initial release. As it is not possible to run all previously available ATP systems on new problems when they are added, this approach assumes that if a problem is unsolved by current ATP systems when it is added to the TPTP, then it would have been unsolved by previously available ATP systems. Under that assumption, the new problem is retrospectively “added” to prior TPTP releases for the analysis. If a problem is solved when it is added to the TPTP, it is ignored because it may have been solved in prior versions as well, and therefore should not serve as an indicator of progress. This analysis reports the fraction of problems solved for each release, but note that the fraction is with respect to both the number of problems actually in the release and also the problems retrospectively “added.”

The growing set analysis is performed on the whole TPTP and on four SPCs. These were chosen because many ATP problems in those forms have been contributed to the TPTP, and correspondingly there are many ATP systems that can attempt them; they represent the “real world” demand for ATP capability.

The table here in the public data folder shows the breakdown of TPTP problems by content fields, as well as by SPCs used in the analysis. The totals are slightly larger than those shown in the analysis, as some problems were left out for technical reasons (no scores available, problems revised over time, etc.).