Chapter Preview

Chapter 3: Technical Performance

[Table_of_Contents]
[Technical_Performance_Technical_Appendix]

Introduction

The technical performance chapter tracks technical progress in tasks across Computer Vision (Images, Videos, and Image+Language), Natural Language, potential limitations (Omniglot Challenge), and trends in computational capabilities.

Image Classification: ImageNet

ImageNet is a public image dataset of over 14 million images, created by Fei-Fei Li and her collaborators in 2009, to address the issue of scarcity of training data in the field of computer vision. The dataset, and an accompanying yearly competition (ImageNet Large Scale Visual Recognition Challenge, or ILSVRC), have been important catalysts to the developments of computer vision over the last 10 years. It was a 2012 submission to ILSVRC by Krizhevsky et al. that lead to a revival of interest in convolutional neural networks and deep learning.

The database is organized according to the WordNet hierarchy, with images depicting both higher- ("animal") and lower-level concepts ("cat"). A key computer vision task that is studied with this dataset is image classification, where an algorithm must infer whether any of the 1000 object categories of interest is present in the image.

The graph below shows accuracy scores for image classification on the ImageNet dataset over time, which can be viewed as a proxy for broader progress in supervised learning for image recognition.

ImageNet performance is being tracked by looking at scores on the validation set from the ImageNet 2012 dataset reported in published papers. The appendix documents variants of evaluation metrics to assess performance on ImageNet. The graph (Figure 3.1) shows ImageNet performance of the best performing models models trained on the ImageNet Competition training data only (grey points). The first method surpassing human performance[5] was published in 2015, and the ImageNet challenge discontinued in 2017. The dataset continues to be an important benchmark for new computer vision models, and gradual improvements continue to be reported. Three of the most recently published successful methods on this task used additional data for training - they are included as a separate plot on this graph (orange points).

Alternatively, the appendix also shows the performance improvement based on Top-5 accuracy (which evaluates a prediction as successful if the 5 top predictions returned by the model included the correct classification).

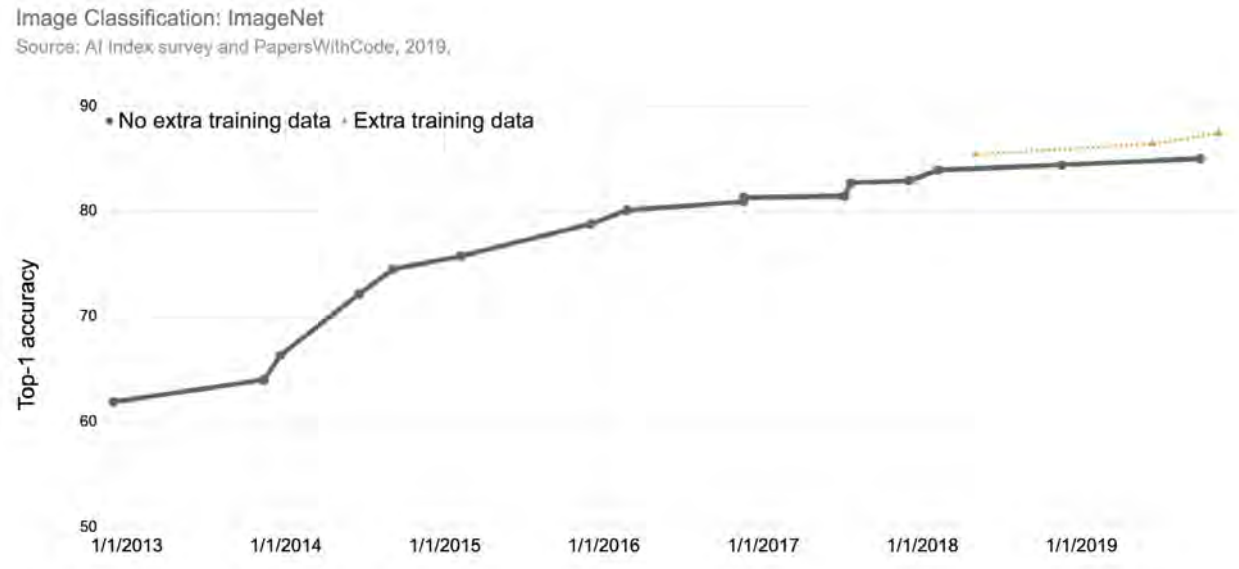

Fig. 3.1.

[5] Note: human performance here is represented by a single person annotating images. It is not representative of "human performance" for a large population.

[Technical_Performance_Technical_Appendix]_[Leaderboard_Papers_with_Code]
[Access_Data]

Image Classification: ImageNet Training Time and Cost

Training Time on Public Clouds

State-of-the-art image classification methods are largely based on supervised machine learning techniques. Measuring how long it takes to train a model and associated costs is important because it is a measurement of the maturity of AI development infrastructure, reflecting advances in software and hardware.

The graph (Figure 3.2a) below shows the time required to train an image classification model to a top-5 validation accuracy of 93% or greater on ImageNet corpora when using public cloud infrastructure. This data is from Stanford's "DAWNBench" project; the data reflects the time it takes well-resourced actors in the AI field to train systems to categorize images. Improvements here give an indication of how rapidly AI developers can re-train networks to account for new data - a critical capability when seeking to develop services, systems, and products that can be updated with new data in response to changes in the world. In a year and a half, the time required to train a network on cloud infrastructure for supervised image recognition has fallen from about three hours in October 2017 to about 88 seconds in July, 2019. Data on ImageNet training time on private cloud instances shows a similar trend (see Appendix).

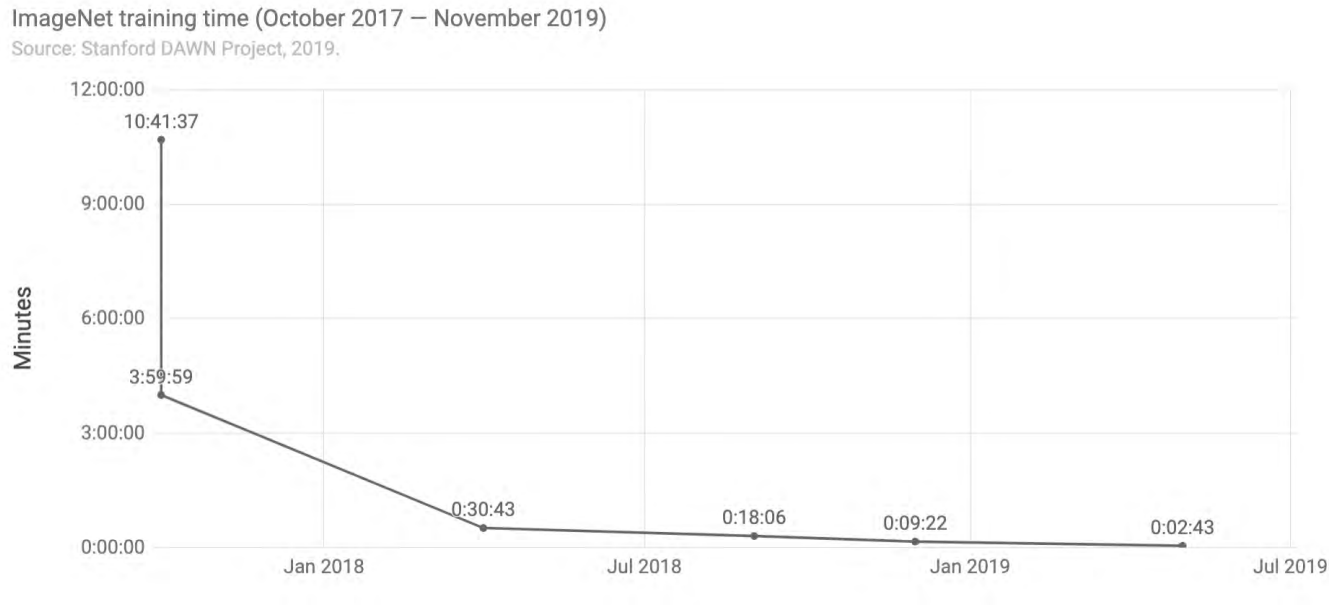

Fig. 3.2a.
Note: DAWNBench will migrate to MLperf. The latest point estimate (not shown) from ML Perf is from July, 2019 at 1 minute and 28 seconds uses Top-1 accuracy versus Top-5 accuracy benchmark shown in the graph above.

In a year and a half, the time required to train a network on cloud infrastructure has fallen from about three hours in October 2017 to about 88 seconds in July, 2019.

[Technical_Performance_Technical_Appendix]_[Benchmark_Details][MLperf]
[Access_Data]

Image Classification: ImageNet Training Time and Cost

The next graph shows the training cost as measured by the cost of public cloud instances to train an image classification model to a top-5 validation accuracy of 93% or greater on ImageNet (Figure 3.2b). The first benchmark was a ResNet model that required over 13 days of training time to reach just above 93% accuracy that cost over $2,323 in October, 2017 (see DAWNbench submissions). The latest benchmark available on Stanford DAWNBench with lowest cost was a ResNet model run on GCP cluster with cloud TPU also reaching slightly above 93% accuracy cost slightly over $12 in September, 2018.

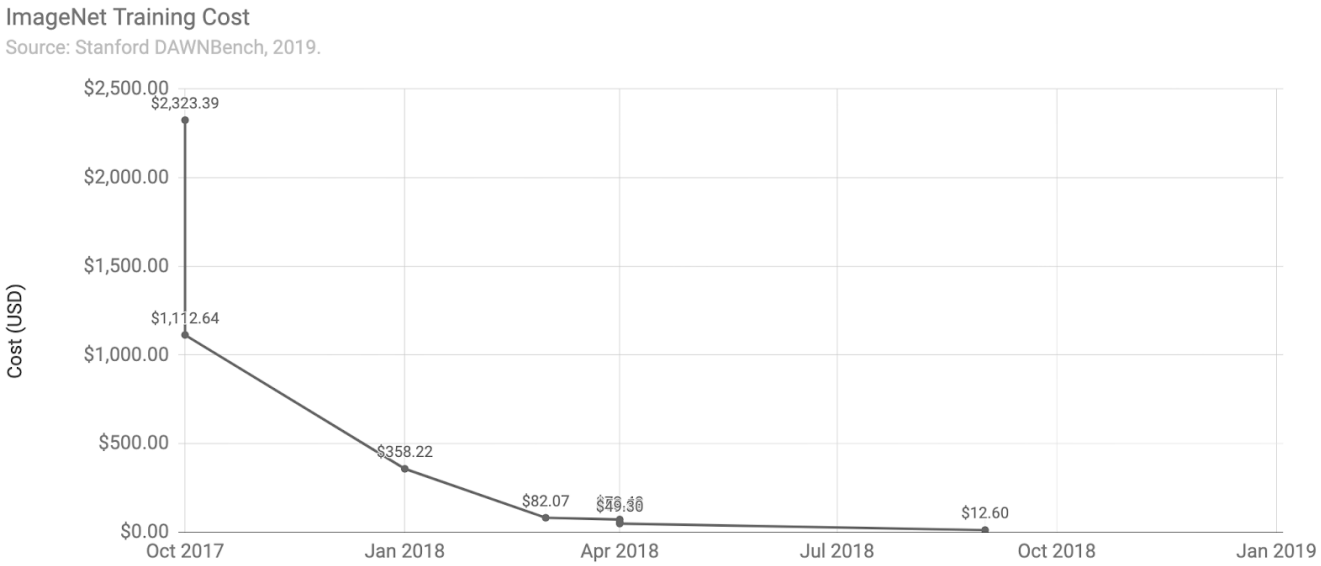

Fig. 3.2b.

[Technical_Performance_Technical_Appendix]_[Benchmark_Details][MLperf]
[Access_Data]

Image Generation: CIFAR-10

Image generation has received attention from artists to the general public and policymakers alike. Image generation (synthesis) relies on AI models generating an output image that is meant to approximate (not necessarily replicate) the data distribution the model was trained on. Progress in image generation can be taken as a proxy for the evolution of AI models' ability to generate content in a variety of domains, ranging from images to video to text. However, assessing progress here is difficult, as beyond a certain level of realism, the quality of an image is subjective. In lieu of large-scale qualitative studies, researchers have begun using a metric called FID, which calculates the distance between the feature vectors; using the Inception v3 image model, activations are calculated on real and generated images, then the distance between these activations is calculated, giving a sense of similarity between these two groups of images. When evaluating FID, a lower score tends to correlate with images that better map their underlying data distribution and is therefore a proxy for image quality. (Figure 3.3).[6] Inception score is also reported (see Appendix Graph).

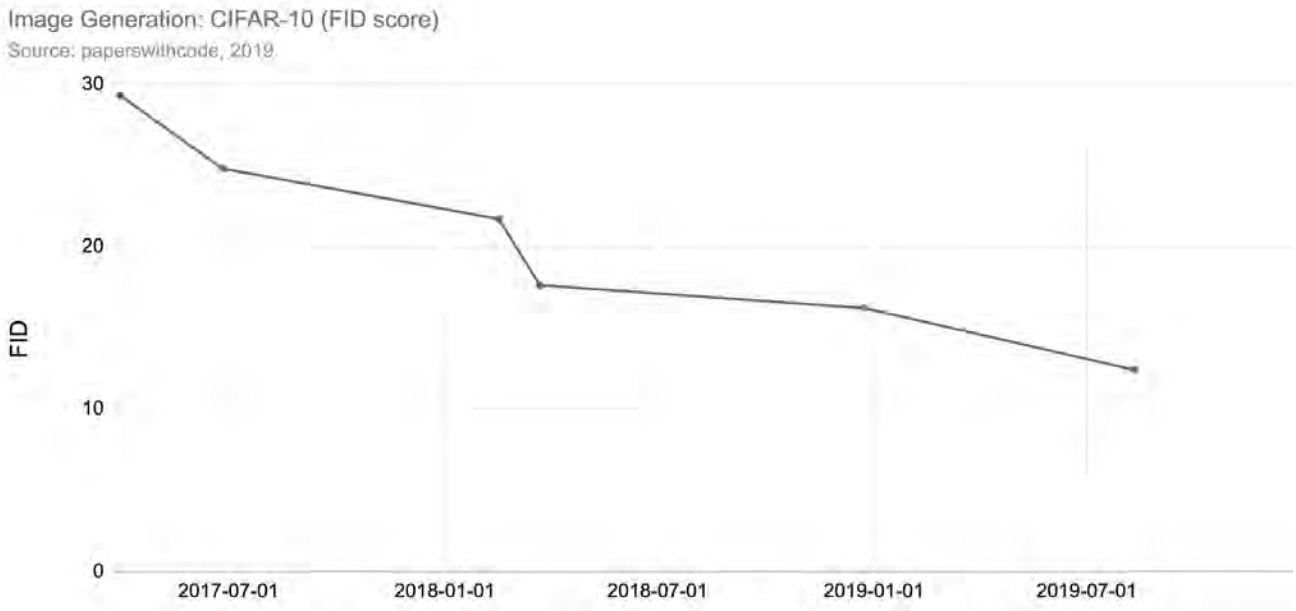

Fig. 3.3.

[1]The inception score is an attempt to remove the subjective human evaluation of images and uses a pre-trained deep learning neural network model for image classification to classify the generated images.

Semantic Segmentation

While image classification can produce a list of objects in the image, many applications require more detailed knowledge of the image contents. For instance, a robot or self-driving car may require to detect the precise boundaries and object categories for all pixels within the image. This corresponds to the task of semantic segmentation, where the algorithm must divide the image into regions and classify each region into one of the categories of interest, producing a pixel-level map of the image contents.

Progress in semantic segmentation is an input to progress in real-world AI vision systems, such as those being developed for self-driving cars. Progress is measured in this domain using the mean intersection over union (IoU) metric on two datasets: Cityscapes (Figure 3.4). Some systems were trained with extra data. See Appendix for details on individual datasets and progress in PASCAL Context

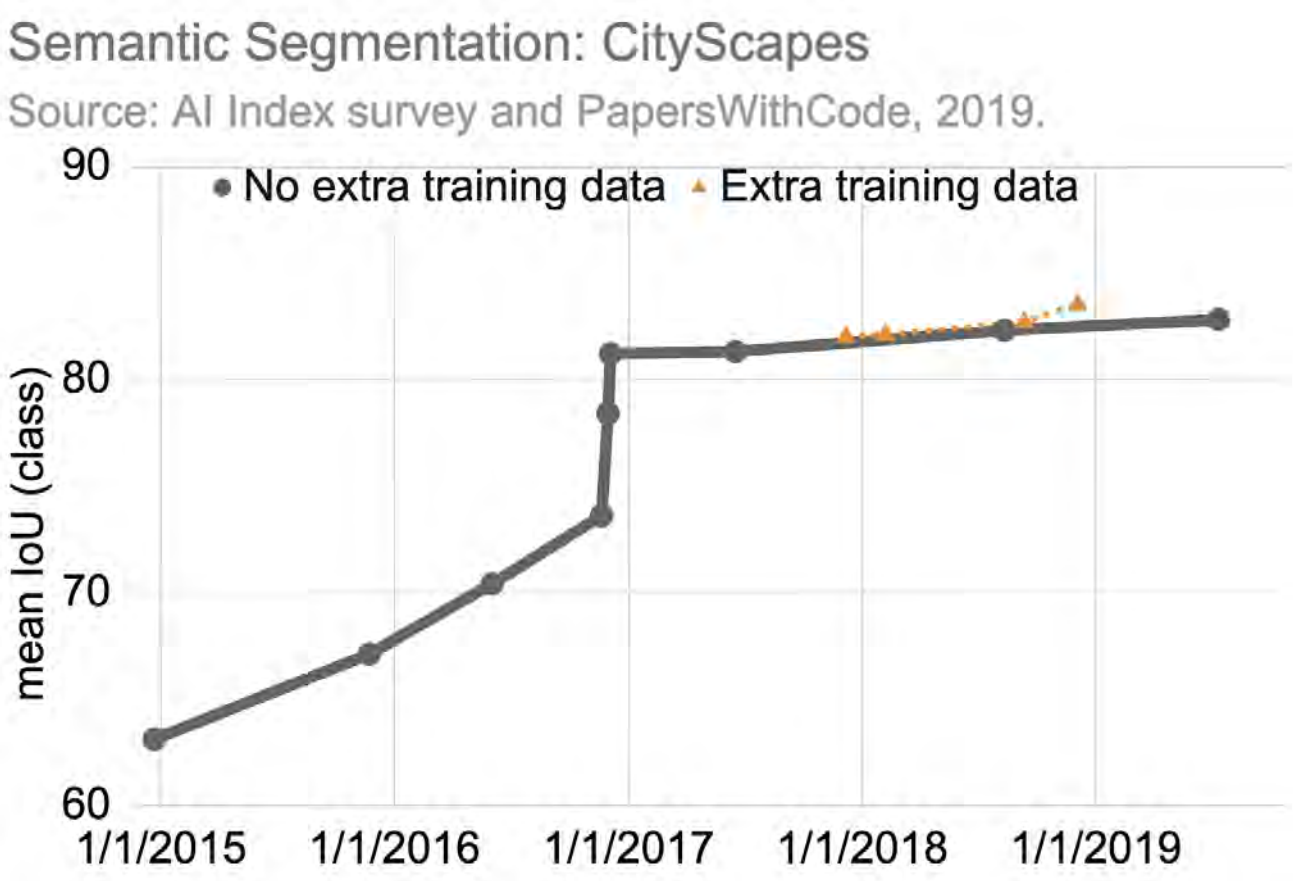

Fig. 3.4.
Note: The orange dots denote tests with additional training data.

ActivityNet

In addition to image analysis, algorithms for understanding and analyzing videos are an important focus in the computer vision research community. Particularly, algorithms that can recognize human actions and activities from videos would enable many important applications. Further discussion of progress in activity recognition in videos appears in the ActivityNet Challenge.

A key task in the ActivityNet Challenge is that of Temporal Activity Localization. In this task, algorithms are given long video sequences that depict more than one activity, and each activity is performed in a sub-interval of the video but not during its entire duration. Algorithms are then evaluated on how precisely they can temporally localize each activity within the video as well as how accurately they can classify the interval into the correct activity category.

ActivityNet has compiled several attributes for the task of temporal localization at the challenge over the last four rounds. Below detailed analysis and trends for this task are presented (e.g. how has the performance for individual activity classes improved over the years (Figure 3.5a)? Which are the hardest and easiest classes now (Figure 3.5b & 3.5c)? Which classes have the leastmost improvement over the years (figure 3.5d)? The ActivityNet statistics are available **here**.

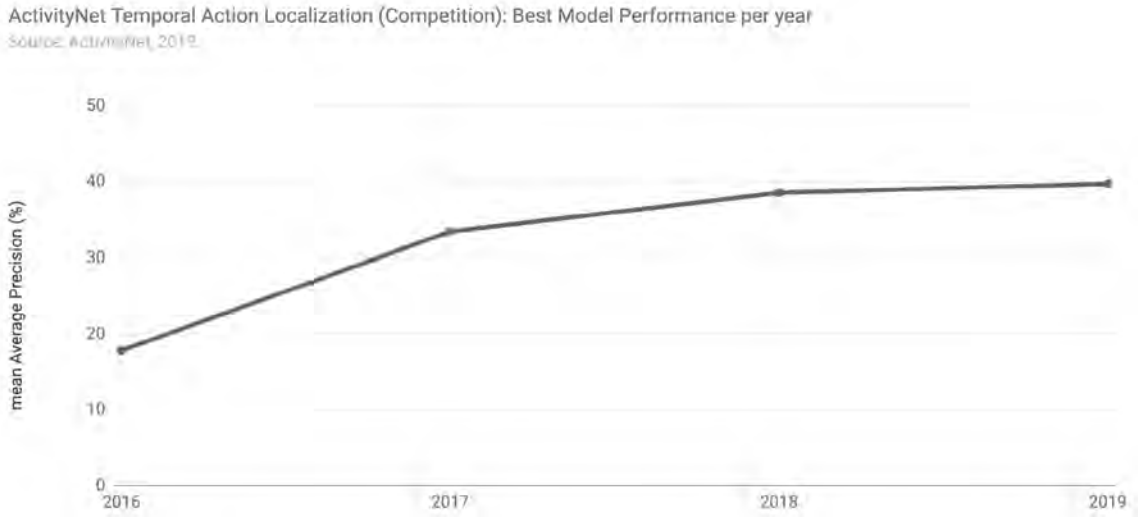

Fig. 3.5a.

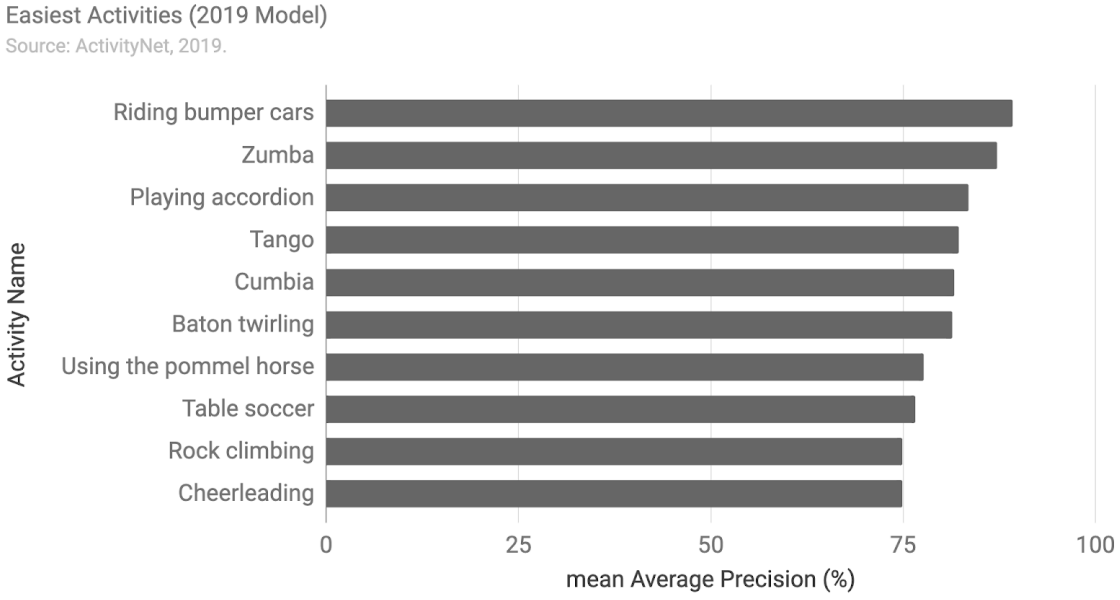

Fig. 3.5b.

[Technical_Performance_Technical_Appendix]_[ActivityNet]
[Access_Data]

Activity Recognition in Videos

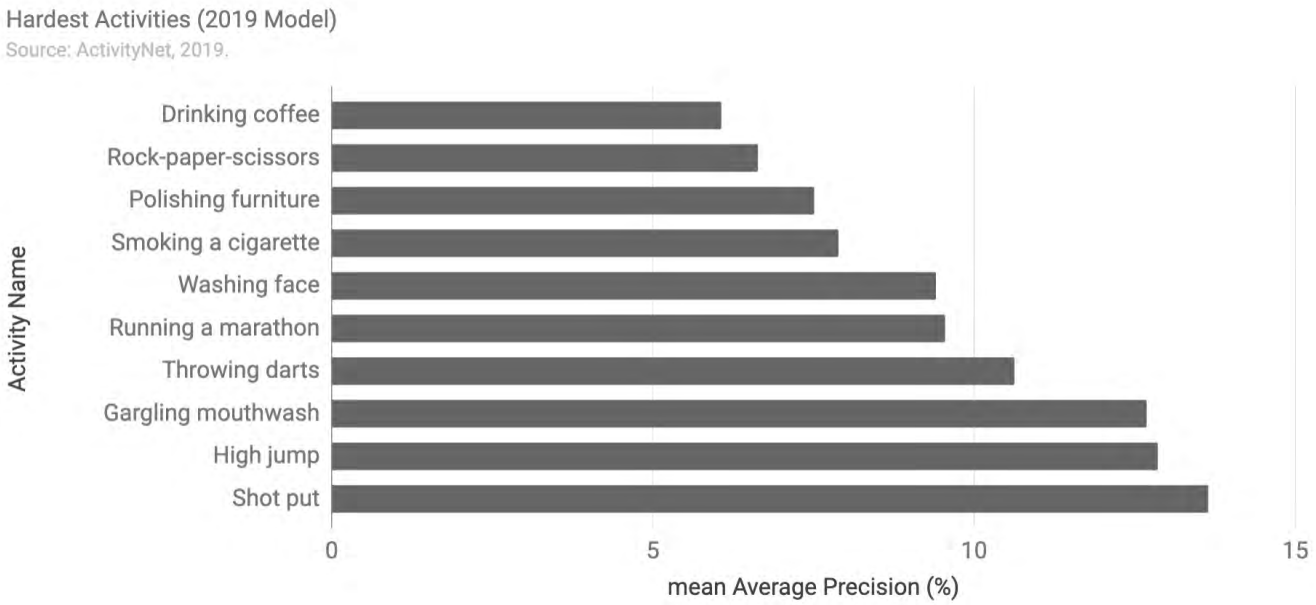

Fig. 3.5c.

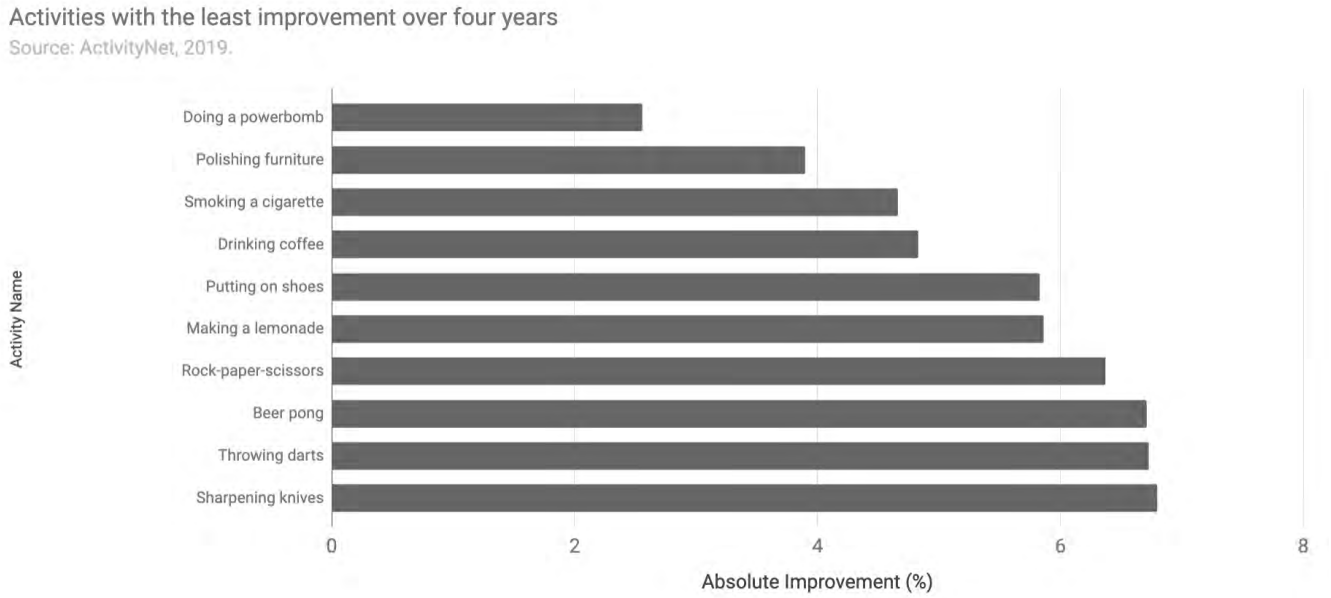

Fig. 3.5d.

"The emergence of large-scale datasets such as ActivityNet and Kinetics has equipped computer vision researchers with valuable data and benchmarks to train and develop innovative algorithms that push the limits of automatic activity understanding. These algorithms can now accurately recognize hundreds of complex human activities such as bowling or sailing, and they do so in real-time. However, after organizing the International Activity Recognition Challenge (ActivityNet) for the last four years, we observe that more research is needed to develop methods that can reliably discriminate activities, which involve fine-grained motions and/or subtle patterns in motion cues, objects, and human-object interactions. Looking forward, we foresee the next generation of algorithms to be one that accentuates learning without the need for excessively large manually curated data. In this scenario, benchmarks and competitions will remain a cornerstone to track progress in this self-learning domain."
Bernard Ghanem, Associate Professor of Electrical Engineering
King Abdullah University of Science and Technology

[Technical_Performance_Technical_Appendix]_[ActivityNet]
[Access_Data]

Visual-Question Answering (VQA)

The VQA challenge incorporates both computer vision and natural language understanding. The VQA challenge tests how well computers can jointly reason over these two distinct data distributions. The VQA challenge uses a dataset containing open-ended questions about the contents of images. Successfully answering these questions requires an understanding of vision, language and commonsense knowledge. In 2019, the overall accuracy grew by +2.85% to 75.28% (Figure 3.6). The 2019 VQA challenge had 41 teams representing more than 34 institutions and 11 countries. Reader refer to the VQA challenge website and Appendix for more details.

Can you beat the VQA challenge?
To get a sense of the challenge, you can try online VQA demos out at https://vqa.cloudcv.org/. Upload an image, ask the model a question, and see what it does.

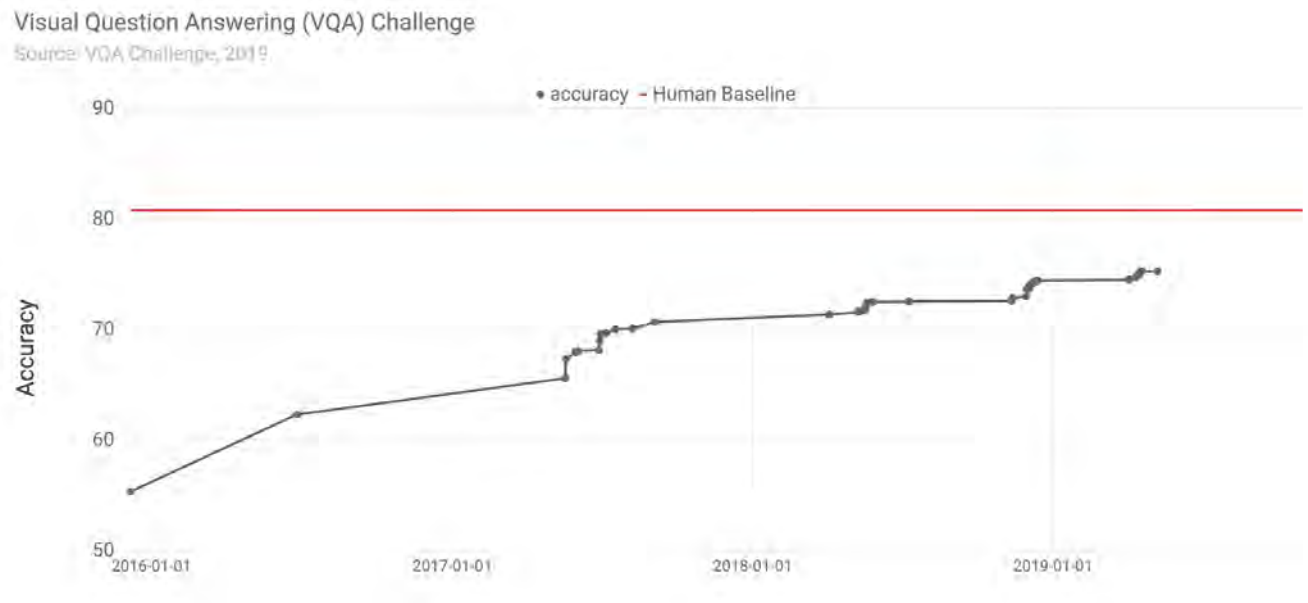

Fig. 3.6.
Note: Human performance is measured by having humans answer questions for images and evaluating their answers using the same metrics as we use to evaluate machines that answer the same questions. Inter-human disagreement, paraphrased answers, spelling errors, etc, contribute to human performance being (quite a bit lower) than 100%.

What explains progress in this domain? *"There's been no silver bullet. Progress has been the consequence of open exploratory research and consistent iterations by researchers in the community -- the vision and language community, the vision community, and the language community. As a community we identified effective multimodal fusion techniques, image representations that are more appropriate for tasks that link to language, convolutional neural network architectures for improved perception, pre-training mechanisms to learn language representations that can be transferred to other tasks."*
Devi Parikh
Georgia Tech | Facebook AI Research (FAIR)

[Technical_Performance_Technical_Appendix]_[VQA_Challenge][VQA_Workshop_Presentation]
[Access_Data]

GLUE

Being able to analyze text is a crucial, multi-purpose AI capability. In recent years, progress in natural language processing and natural language understanding has caused the AI community to develop new, harder tests for AI capabilities. In the language domain, a good example is GLUE, the General Language Understanding Evaluation benchmark. GLUE tests single AI systems on nine distinct tasks in an attempt to measure the general text-processing performance of AI systems. GLUE consists of nine sub-tasks — two on single sentences (measuring linguistic acceptability and sentiment), three on similarity and paraphrase, and four on natural language inference, including the Winograd Schema Challenge. As an illustration of the pace of progress in this domain, though the benchmark was only released in May 2018, performance of submitted systems crossed non-expert human performance in June, 2019. Performance has continued to improve in 2019 (Figure 3.7) with models like RoBERTa from Facebook and T5 from Google. More details on GLUE tasks with greater (or shorter) distance to human performance frontier are available (see Appendix Graph).

Fig. 3.7.

[Technical_Performance_Technical_Appendix]_[GLUE_Leaderboard]
[Access_Data]

SuperGLUE

Progress in language-oriented AI systems has been so dramatic that the creators of the GLUE benchmark needed to create a new, more challenging benchmark, so they could test performance after some systems surpassed human performance on GLUE. SuperGLUE contains a new set of more diverse and difficult language understanding tasks, improved resources, and a new public leaderboard.

Within five months of its launch in May, 2019, the T5 model published by Google almost reached human baseline of 89.9 with their at the score of 88.9 (Figure 3.8). This was achieved using a task-agnostic text-to-text framework that utilized an encoder-decoder architecture. The model was pre-trained on a mixture of NLP tasks and fine-tuned on SuperGLUE.

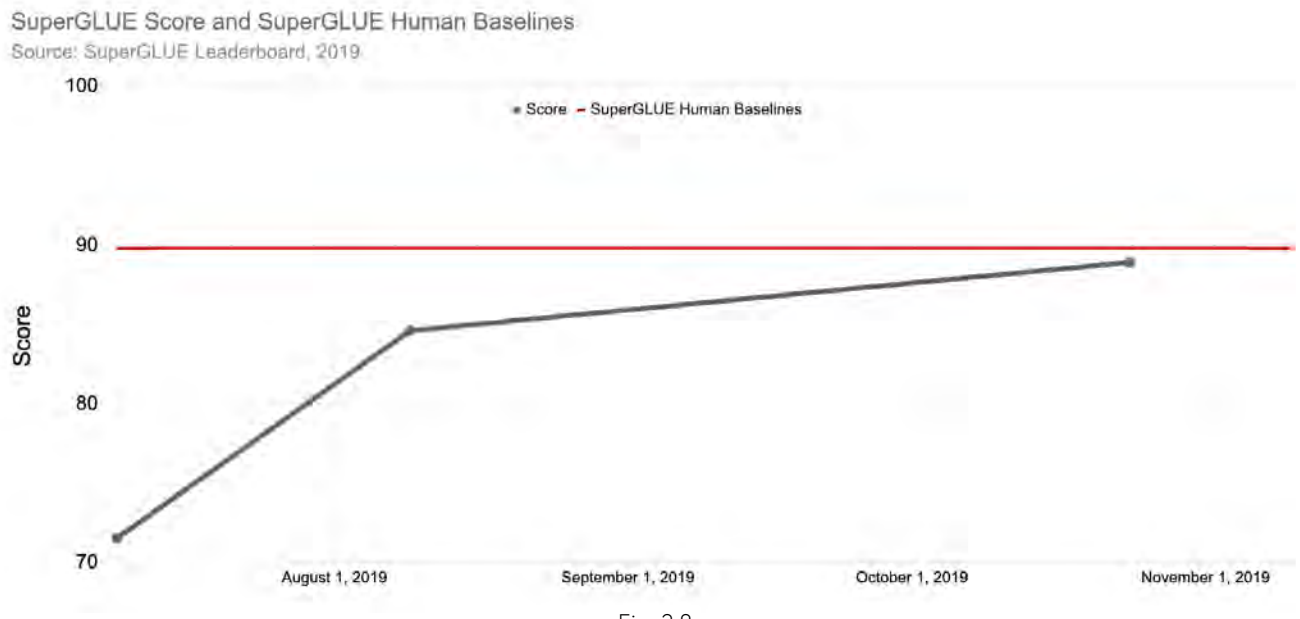

Fig. 3.8.
Notes: Human baseline was estimated by hiring crowdworker annotators through Amazon's Mechanical Turk platform to reannotate a sample of each test set to estimate. More details can be found here.

Since being launched in May, 2019, the T5 Team at Google has almost reached human baseline at the score of 88.9 within five months on SuperGLUE. Human baseline is 89.8.

[Technical_Performance_Technical_Appendix]_[SuperGLUE_Paper][SuperGLUE_Leaderboard]
[Access_Data]

GLUE and superGLUE

What does progress in natural language understanding mean?

What is the best way to interpret the rapid progress in natural language and what might measures like GLUE and SuperGLUE tell us about progress in this domain? Sam Bowman, an assistant professor at NYU whose group has developed GLUE and SuperGLUE offers:

"We know now how to solve an overwhelming majority of the sentence- or pararaph-level text classification benchmark datasets that we've been able to come up with to date. GLUE and SuperGLUE demonstrate this out nicely, and you can see similar trends across the field of NLP. I don't think we have been in a position even remotely like this before: We're solving hard, AI-oriented challenge tasks just about as fast as we can dream them up," Sam says. "I want to emphasize, though, that we haven't solved language understanding yet in any satisfying way."

While GLUE and SuperGLUE may indicate progress in the field, it is important to remember that successful models could be exploiting statistical patterns in their underlying datasets, are likely to display harmful biases, and when they demonstrate better-than-human performance, they may be doing this unevenly, displaying good performance on some tasks and faulty or inhuman reasoning on others.

"This leaves us in an odd position," Bowman says. "Especially for these classification-style tasks, we see clear weaknesses with current methods, but we don't yet have clear, fair ways to quantify those weaknesses. I'm seeing what looks like a new surge of interest in data collection methods and evaluation metrics, and I think that's a healthy thing for us to be focusing on."

Human Expectations for the SuperGLUE Benchmark

The AI Index has partnered with Metaculus, a crowd forecasting initiative, to source 'crowd predictions' from the general public for the 2019 report. The question went public on August 9, 2019 and will close on closes Dec 30, 2019. Respondents don't predict "yes" or "no," but rather the percent likelihood. At the time of writing this, there were 127 human predictions. Metaculus users were asked the following question:

By May 2020, will a single language model obtain an average score equal to or greater than 90% on the SuperGLUE benchmark?

Results: The median prediction of respondents is a 90% likelihood that a single model will obtain an average score equal to or greater than 90% on the SuperGLUE benchmark.

[Technical_Performance_Technical_Appendix]_[SuperGLUE_Paper][SuperGLUE_Leaderboard]
[Access_Data]

SQuAD

One way to highlight recent progress in natural language processing is to examine performance on the Stanford Question Answering Dataset (SQuAD) challenge. SQuAD is a reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles. The answer to every question is a segment of text, or span, from the corresponding reading passage, or the question might be unanswerable. SQuAD1.1 is the SQuAD dataset and contains 100,000+ question-answer pairs on 500+ articles. SQuAD2.0 combines the 100,000 questions in SQuAD1.1 with over 50,000 unanswerable questions written adversarially by crowdworkers to look similar to answerable ones. To do well on SQuAD2.0, systems must not only answer questions when possible, but also determine when no answer is supported by the paragraph and abstain from answering. SQuAD2.0 was developed partially because of surprising, rapid performance by entrances on the original SQuAD benchmark. The SQuAD Leaderboard and data are available. The F1 score for SQuAD1.1 went from 67 in August, 2016 to 95 in May, 2019 (Figure 3.9). Progress on SQuAD2.0 has been even faster. F1 score went from 62 in May, 2018 to 90 in June, 2019. CodaLab hosts other active NLP competitions.

The time taken to train QA model to 75 F1 score or greater on SQuAD 1.0 went down from over 7 hours in October, 2017 to less than 19 minutes in March, 2019 (Figure 3.13b). The cost to public cloud instances to train a QA model to has reduced from $8 to 57 cents by December, 2018, and inference time reduced from 638 milliseconds to 7 milliseconds (see Appendix Graph).

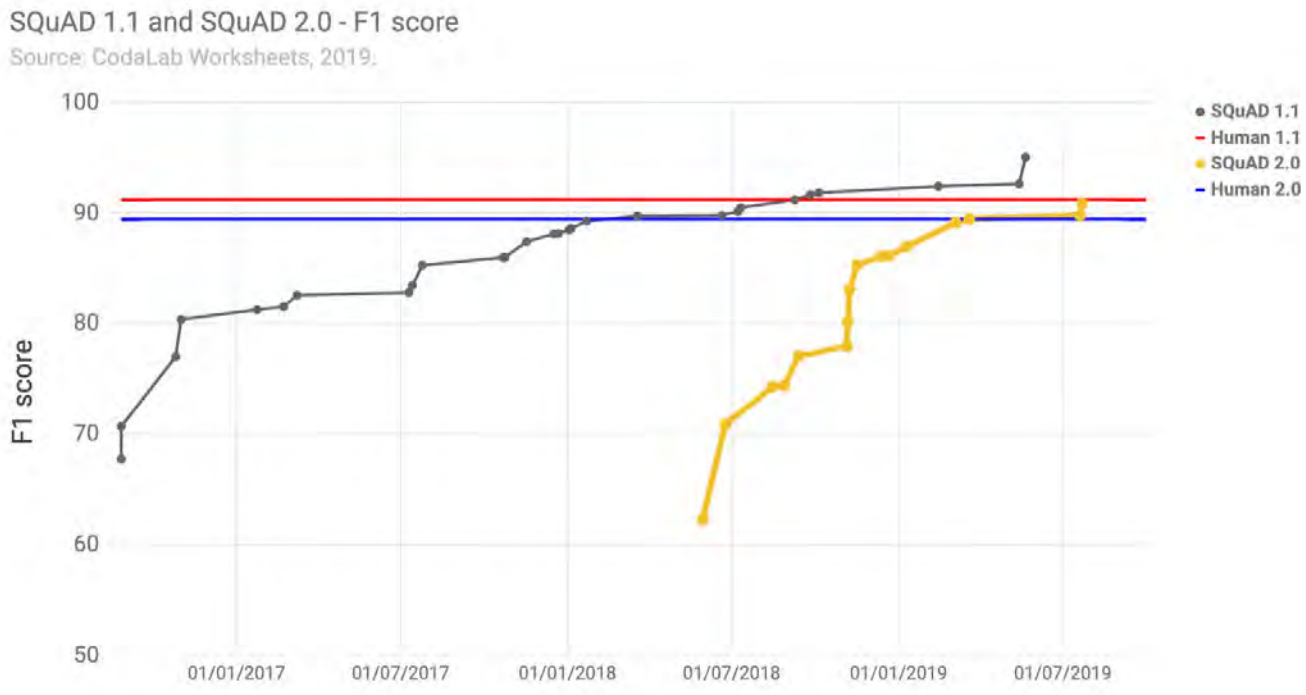

Fig. 3.9.

The F1 score for SQuAD1.1 went from 67 in August, 2016 to 95 in May, 2019. Progress on SQuAD2.0 has been even faster. F1 score went from 62 in May, 2018 to 90 in June, 2019.

[Technical_Performance_Technical_Appendix]_[SQuAD_Leaderboard]
[Access_Data]

Reasoning

The Allen Institute for Artificial Intelligence (AI2) has several initiatives that relate to measuring the advancing capabilities of AI systems and is home to several AI research initiatives including the AllenNLP, Aristo, and Mosaic projects. Several AI2 Leaderboards are publicly available for NLP and commonsense reasoning tasks. Performance improvements in selected tasks are presented below.

AI2 Reasoning Challenge (ARC)

Released in April 2018, the ARC dataset contains 7,787 genuine grade-school level, multiple-choice science questions. The questions are text-only, English language exam questions that span several grade levels. Each question has a multiple-choice structure (with typically four answer options). The questions are accompanied by the ARC Corpus, a collection of 14M unordered, science-related sentences including knowledge relevant to ARC. It is not guaranteed that answers to the questions can be found in the corpus. The ARC dataset is divided into a Challenge Set (2,590 questions) and an Easy Set (5,197 questions). The Challenge Set contains only questions that were answered incorrectly by both a retrieval-based algorithm and a word co-occurrence algorithm.

ARC Easy

The first graph from AI2 shows the progress on the ARC-Easy dataset, 5,197 questions that can be answered by retrieval or co-occurrence algorithms. More details about this task can be found in the Appendix. There have been 20 submissions to the ARC-Easy leaderboard, with the top score yielding 85.4% accuracy on the test set, updated on September 27, 2019 (Figure 3.10).

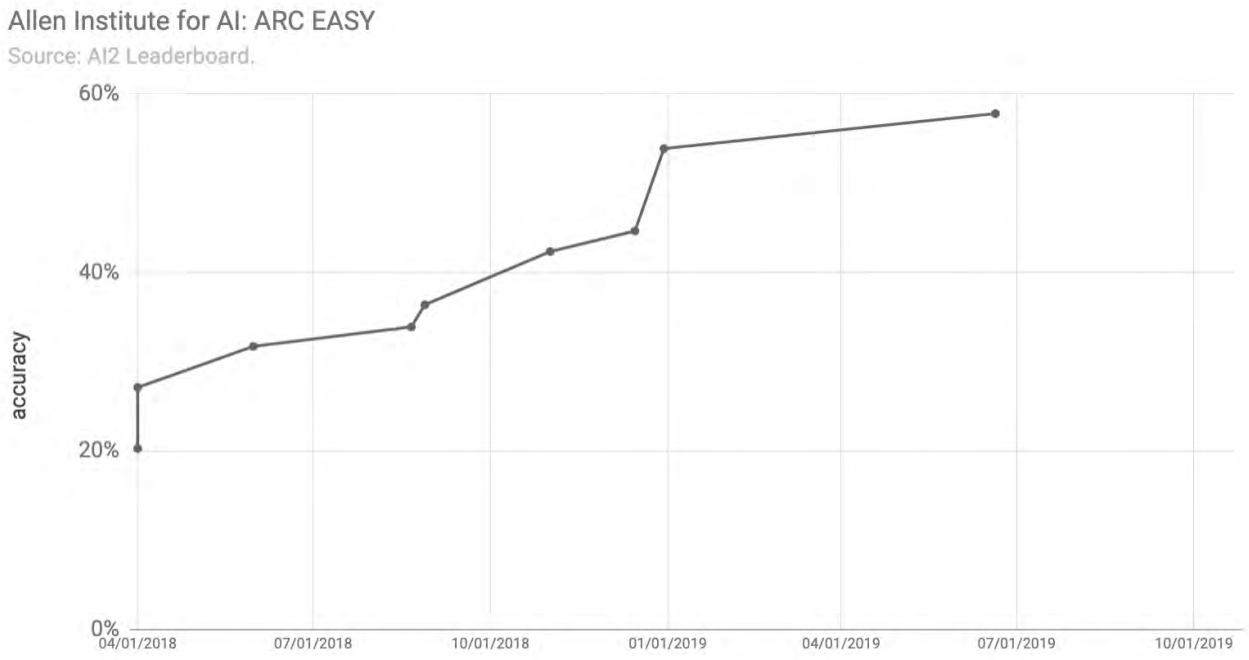

Fig. 3.10.

[Technical_Performance_Technical_Appendix]_[ARC_Easy_Leaderboard]
[Access_Data]

Reasoning

ARC Challenge Set

The graph below shows performance over time for the ARC Challenge Set. See Appendix for data and methodology. There have been 26 submissions to the ARC Challenge Set leaderboard with a top score of 67.7% last updated on September 27, 2019 (Figure 3.11).

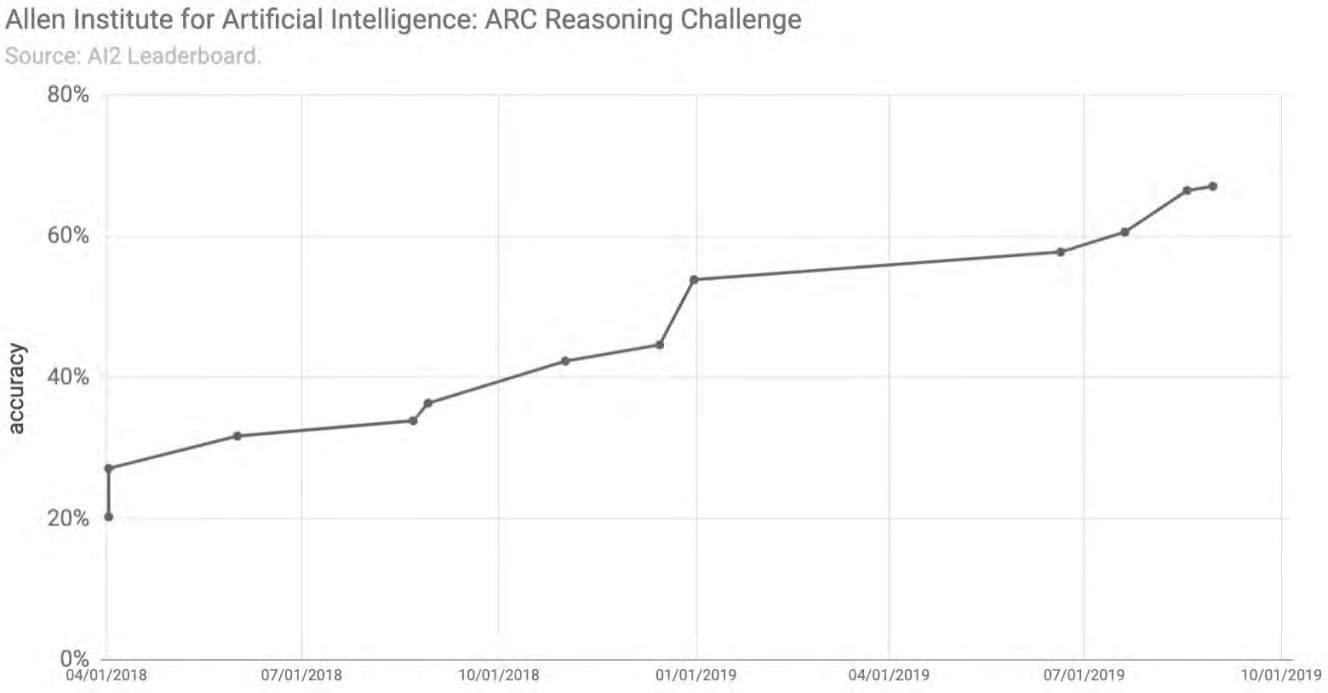

Fig. 3.11.

[Technical_Performance_Technical_Appendix]_[ARC_Reasoning_Challenge_Leaderboard]
[Access_Data]

Commercial Machine Translation (MT)

Translation is one of the more easily applicable capabilities of contemporary language-oriented AI systems. Therefore, examining the number and performance of commercially deployed translation systems gives us a sense of how rapidly technology migrates from research to production, and of what the impact is here.

According to Intento, a startup that provides simple APIs to evaluate third-party AI models in MT from many vendors, the number of commercially available MT systems with pre-trained models and public APIs has grown rapidly, from 8 in 2017 to over 24 in 2019 (Figure 3.12a). Increasingly, MT systems provide a full range of customization options: pre-trained generic models, automatic domain adaptation to build models and better engines with their own data, and custom terminology support.

The growth in commercial MT is driven by engines that excel at their geography and business-related language pairs and domains (Germany, Japan, Korea, China). Since early 2018, the increase in commercial MT system is due to two factors: (1) existing vendors of on-premise and bespoke MT are starting to provide pre-trained models available in the cloud and (2) the technology barrier to fielding translation systems is getting lower as a consequence of more neural machine translation (NMT) frameworks being made available open-source.

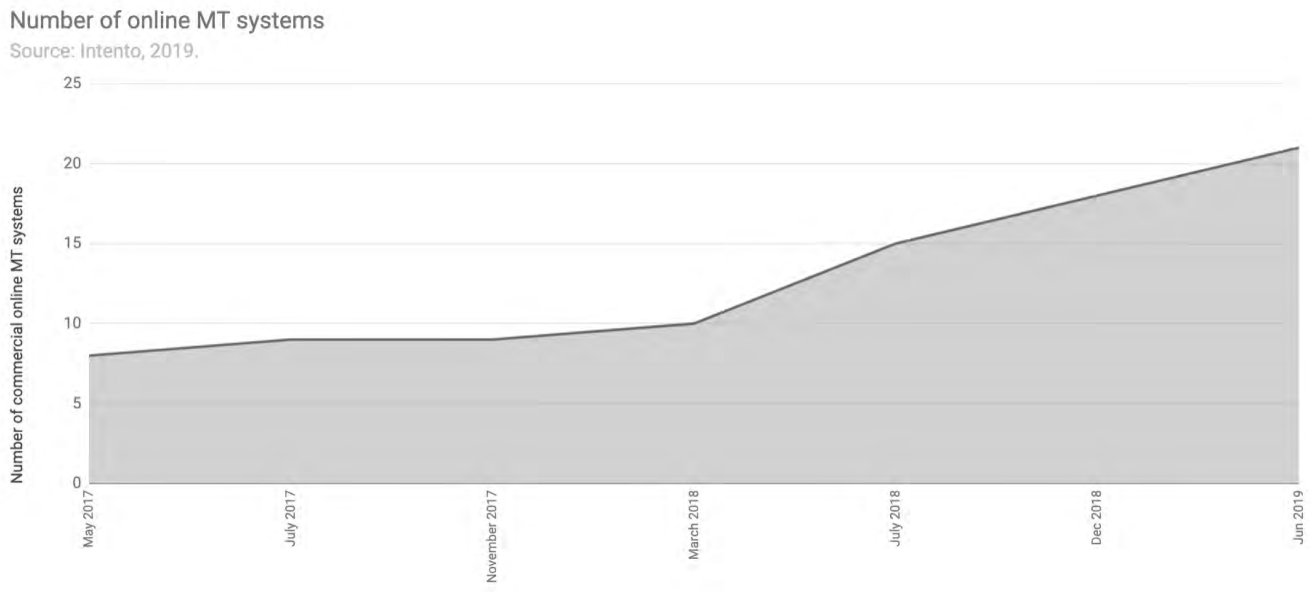

Fig. 3.12a.

[Technical_Performance_Technical_Appendix]
[Access_Data]

Commercial Machine Translation (MT)

Commercial MT quality is evaluated quality using hLEPOR metric, which measure the difference from a human reference translation. hLEPOR scores of 0.7 means almost human-level quality with just a couple of mistakes per sentence. The hLEPOR performance score in language pairs for online systems is presented below (Figure 3.12b). To make the analysis comparable, the presentation is only for pairs including English. It is based on ranking the best online MT system for 48 language-pairs tested. Portugese-English and English-Portugese are pairs with highest hLEPOR score, followed by English to German, and Italian to English. Details on data, methodology, and replicability of results can be found in the Technical Appendix. The next chart shows the ranking of language pairs based on improvement in hLEPOR score between May, 2017 and June, 2019 (figure 3.12c). The fastest improvement was for Chinese-to-English, followed by English-to-German and Russian-to-English. Performance of the baseline models varies widely between different language pairs. The main contributing factor is language pair popularity, which defines how much investment goes into data acquisition and curation. Also, the next-generation translation technology (such as Transformer) is being rolled out to the most popular language pairs first, while rare language pairs may still employ Phrase Based Machine Translation (PBMT) models.

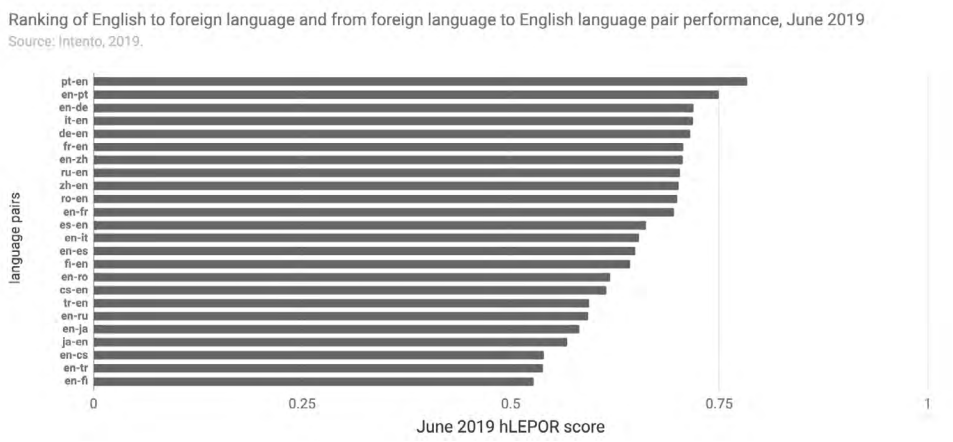

Fig. 3.12b.

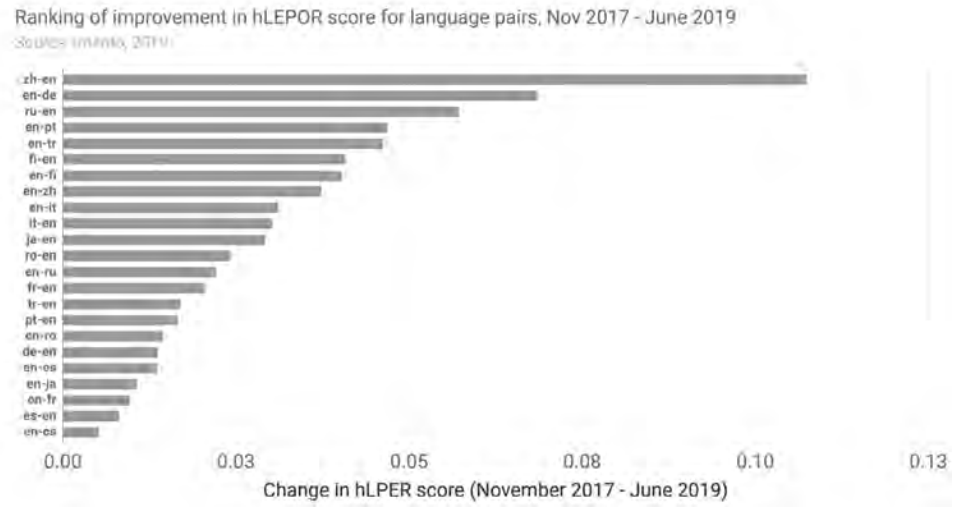

Fig. 3.12c.

"Increased data confidentiality concerns complicate data acquisition for domain-specific models. As a result, we see MT providers putting a lot of effort into building domain adaptation tools for data owners. Those are AutoML-type technology, terminology adaptation, and the ability to improve models based on end-user feedback. We expect these will be the primary technology drivers in the near term."
Konstantin Savenkov, CEO Intento, Inc.

[Technical_Performance_Technical_Appendix]
[Access_Data]

Omniglot Challenge

There has been notable progress on one-shot classification over the last three years; however, there has been less progress on the other four concept learning tasks in the Omniglot Challenge. The Omniglot Challenge requires performing many tasks with a single model, including classification, parsing, generating new exemplars, and generating whole new concepts. Bayesian program learning (BPL) performs better than neural network approaches on the original one-shot classification challenge, despite the improving capabilities of neural network models (Figure 3.13). See tge Appendix for details on the task.

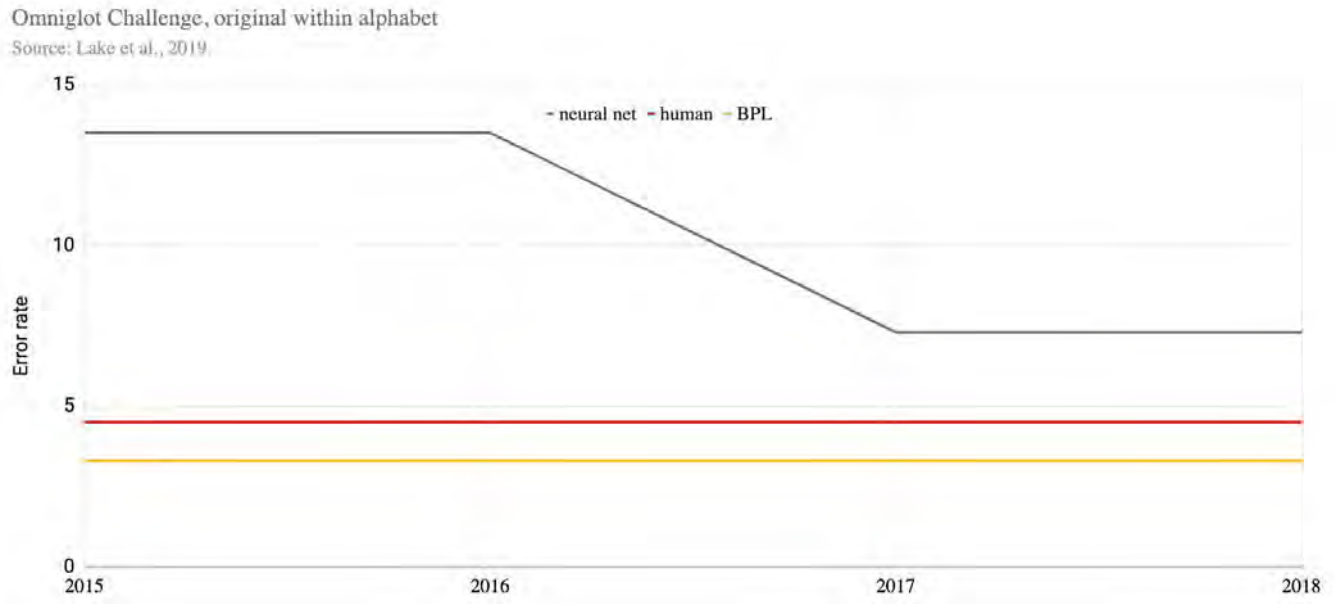

Fig. 3.16.

The Omniglot challenge: a 3-year progress report

Human-level concept learning through probabilistic program induction

"Achieving human-level concept learning will require learning richer representations from less data, and reconfiguring these representations to tackle new tasks" says Brenden Lake, an Assistant Professor at New York University and author of the Omniglot challenge and progress report. Lake further says that *"there is no official leaderboard for Omniglot, and in fact, it's difficult to define an appropriate leaderboard for the entire challenge. Progress on building machines that can learn concepts in a more human-like way cannot be boiled down to just a single number or a single task. Rather, as the progress report states, models need to be developed with a broad competence for performing a variety of different tasks using their conceptual representation."*

[Technical_Performance_Technical_Appendix]
[Access_Data]

Computational Capacity

The amount of computation used in the largest AI training runs has doubled every 3.4 months since 2012 (net increase of 300,000x). The y-axis of the chart shows the total amount of compute, in petaflop/s-days, used to train selected results (Figure 3.14a and 3.14b). A petaflop-day (pf-day) consists of performing 10^{15} neural net operations per second for one day, or a total of about 10^{20} operations. The x-axis is the publication date. Doubling time for the line of best fit shown is 3.4 months. Based on analysis of compute used in major AI results for the past decades, a structural break with two AI eras are identified by OpenAI:

1) **Prior to 2012** - AI results closely tracked Moore's Law, with compute doubling every two years (Figure 3.14a).

2) **Post-2012** - compute has been doubling every 3.4 months (Figure 3.14b). Since 2012, this compute metric has grown by more than 300,000x (a 2-year doubling period would yield only a 7x increase).

Two methodologies were used to generate these data points. When information was available, the number of FLOPs (adds and multiplies) in the described architecture per training example were directly counted and multiplied by the total number of forward and backward passes during training. When enough information to directly count FLOPs was not available, GPU training time and total number of GPUs were used and a utilization efficiency (usually 0.33) was assumed. Technical details on calculations can be found on the OpenAI blog.

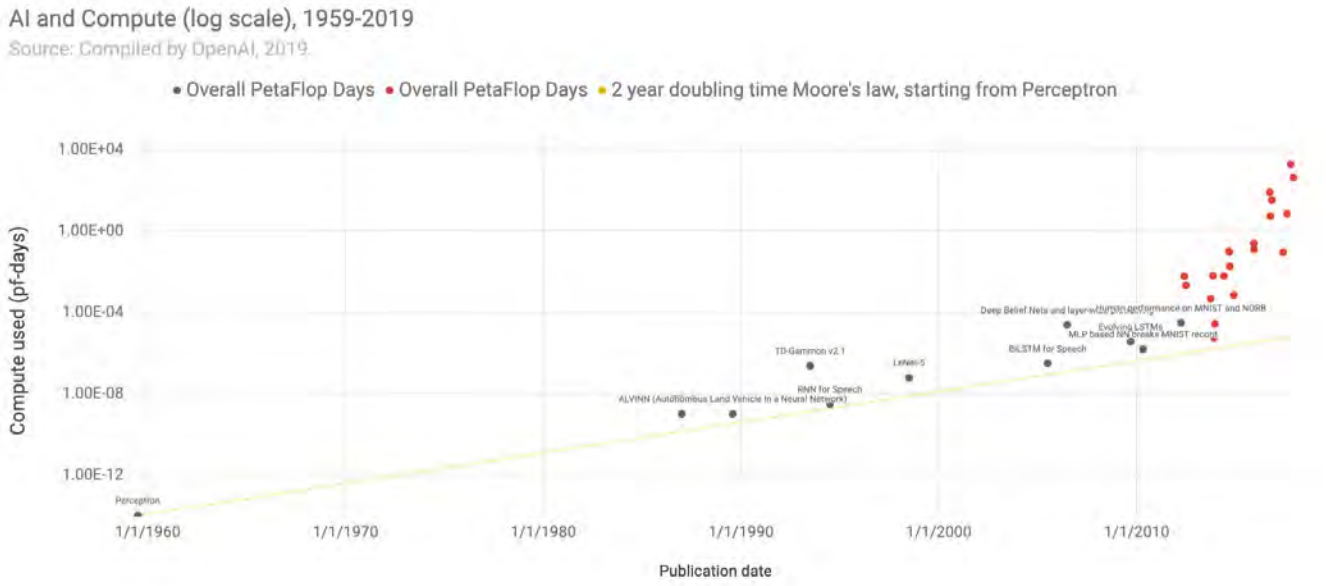

Fig. 3.14a.

Computational Capacity

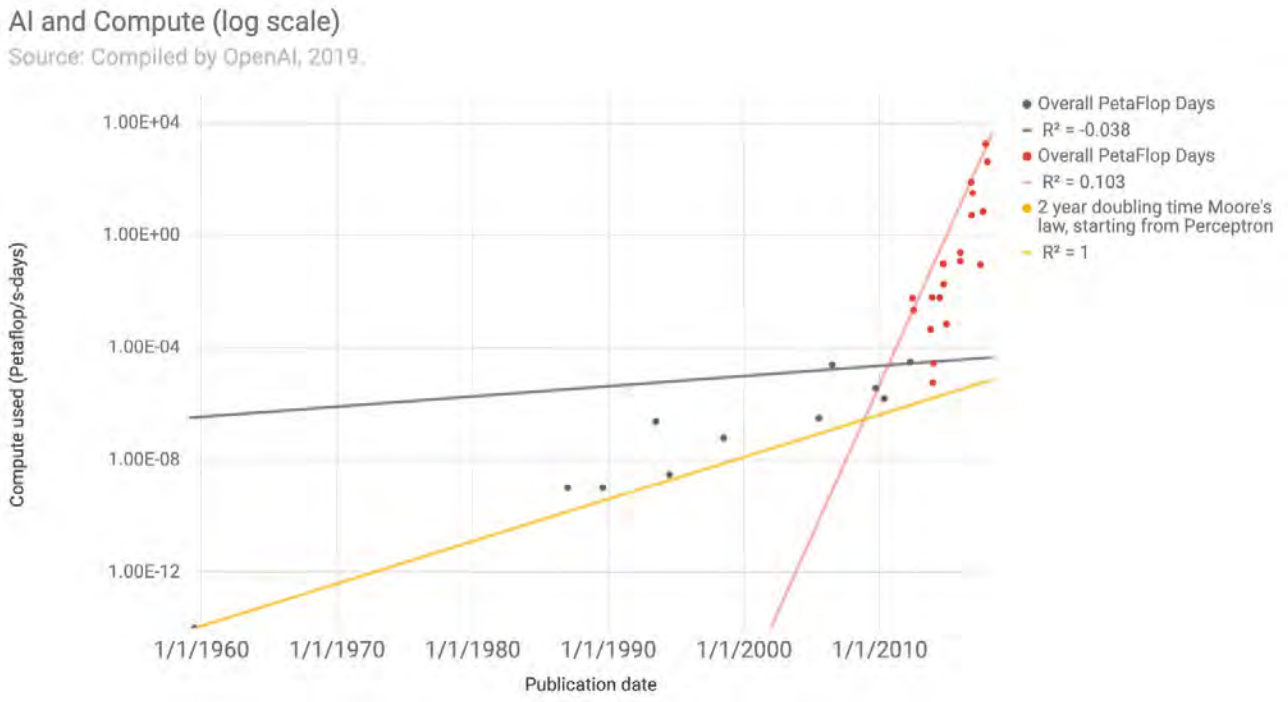

Fig. 3.14b.

Prior to 2012, AI results closely tracked Moore's Law, with compute doubling every two years. Post-2012, compute has been doubling every 3.4 months.

[Technical_Performance_Technical_Appendix]
[Access_Data]

Human-Level Performance Milestones

The inaugural 2017 AI Index report included a timeline of circumstances where AI reached or beat human-level performance. The list outlined game playing achievements, accurate medical diagnoses, and other general, but sophisticated, human tasks that AI performed at a human or superhuman level. This year, two new achievements are added to that list. It is important not to over-interpret these results. The tasks below are highly specific, and the achievements, while impressive, say nothing about the ability of the systems to generalize to other tasks.

1980

Othello

In the 1980s Kai-Fu Lee and Sanjoy Mahajan developed BILL, a Bayesian learning-based system for playing the board game Othello. In 1989, the program won the US national tournament of computer players, and beat the highest ranked US player, Brian Rose, 56—8. In 1997, a program named Logistello won every game in a six game match against the reigning Othello world champion.

1995

Checkers

In 1952, Arthur Samuels built a series of programs that played the game of checkers and improved via self-play. However, it was not until 1995 that a checkers-playing program, Chinook, beat the world champion.

1997

Chess

Some computer scientists in the 1950s predicted that a computer would defeat the human chess champion by 1967, but it was not until 1997 that IBM's DeepBlue system beat chess champion Gary Kasparov. Today, chess programs running on smartphones can play at the grandmaster level.

2011

Jeopardy!

In 2011, the IBM Watson computer system competed on the popular quiz show Jeopardy! against former winners Brad Rutter and Ken Jennings. Watson won the first place prize of $1 million.

2015

Atari Games

In 2015, a team at Google DeepMind used a reinforcement learning system to learn how to play 49 Atari games. The system was able to achieve human-level performance in a majority of the games (e.g., Breakout), though some are still significantly out of reach (e.g., Montezuma's Revenge).

2016

Object Classification in ImageNet

In 2016, the error rate of automatic labeling of ImageNet declined from 28% in 2010 to less than 3%. Human performance is about 5%.

2016

Go

In March of 2016, the AlphaGo system developed by the Google DeepMind team beat Lee Sedol, one of the world's greatest Go players, 4—1. DeepMind then released AlphaGo Master, which defeated the top ranked player, Ke Jie, in March of 2017. In October 2017, a Nature paper detailed yet another new version, AlphaGo Zero, which beat the original AlphaGo system 100—0.

2017

Skin Cancer Classification

In a 2017 Nature article, Esteva et al. describe an AI system trained on a data set of 129,450 clinical images of 2,032 different diseases and compare its diagnostic performance against 21 board-certified dermatologists. They find the AI system capable of classifying skin cancer at a level of competence comparable to the dermatologists.

2017

Speech Recognition on Switchboard

In 2017, Microsoft and IBM both achieved performance within close range of "human-parity" speech recognition in the limited Switchboard domain

2017

Poker

In January 2017, a program from CMU called Libratus defeated four to human players in a tournament of 120,000 games of two-player, heads up, no-limit Texas Hold'em. In February 2017, a program from the University of Alberta called DeepStack played a group of 11 professional players more than 3,000 games each. DeepStack won enough poker games to prove the statistical significance of its skill over the professionals.

2017

Ms. Pac-Man

Maluuba, a deep learning team acquired by Microsoft, created an AI system that learned how to reach the game's maximum point value of 999,900 on Atari 2600.

2018

Chinese - English Translation

A Microsoft machine translation system achieved human-level quality and accuracy when translating news stories from Chinese to English. The test was performed on newstest2017, a data set commonly used in machine translation competitions.

2018

Capture the Flag

A DeepMind agent reached human-level performance in a modified version of Quake III Arena Capture the Flag (a popular 3D multiplayer first-person video game). The agents showed human-like behaviours such as navigating, following, and defending. The trained agents exceeded the win-rate of strong human players both as teammates and opponents, beating several existing state-of-the art systems.

2018

DOTA 2

OpenAI Five, OpenAI's team of five neural networks, defeats amateur human teams at Dota 2 (with restrictions). OpenAI Five was trained by playing 180 years worth of games against itself every day, learning via self-play. (*OpenAI Five is not yet superhuman, as it failed to beat a professional human team*)

2018

Prostate Cancer Grading

Google developed a deep learning system that can achieve an overall accuracy of 70% when grading prostate cancer in prostatectomy specimens. The average accuracy of achieved by US board-certified general pathologists in study was 61%. Additionally, of 10 high-performing individual general pathologists who graded every sample in the validation set, the deep learning system was more accurate than 8.

2018

Alphafold

DeepMind developed Alphafold that uses vast amount of geometric sequence data to predict the 3D structure of protein at an unparalleled level of accuracy than before.

2019

Alphastar

DeepMind developed Alphastar to beat a top professional player in Starcraft II.

2019

Detect diabetic retinopathy (DR) with specialist-level accuracy

Recent study shows one of the largest clinical validation of a deep learning algorithm with significantly higher accuracy than specialists. The tradeoff for reduced false negative rate is slightly higher false positive rates with the deep learning approach.

Measurement Questions

- In recent years, we've seen machine learning based approaches demonstrate increasingly good performance on tasks as diverse as image recognition, image generation, and natural language understanding. Since many of these techniques are data-intensive or compute-intensive, there is a need for metrics that measure the *efficiency* of AI systems, as well as their raw capabilities.

- Moving from single task to multi-task evaluation for AI capabilities, how should the importance of various sub-tasks be weighted for assessing overall progress?

- How can tasks where we're making *no progress* be measured? Many measures of AI progress exist because developers can build systems which can (partially) solve the task - how can areas that are challenging for contemporary systems be assessed?