Capgemini

STANDARD DATA LABELING EXPERTISE IS NOT ENOUGH

AI solutions combined with technological and human ingenuity lead to faster, more accurate outcomes

GET THE FUTURE YOU WANT

The rate at which organizations are developing AI solutions is growing exponentially. So is the need for high quality training data. **Research and Markets** predicts the AI training dataset market will reach US $3.1 billion by 2027.

This is unsurprising considering the vast range of AI use cases possible for virtually every industry. **Gartner research** has found that there's an urgency of leveraging AI for business transformation, but 50% of IT leaders struggle to move their AI projects past proof of concept to a production level of maturity – one reason being a lack of data necessary to train AI solutions.

96% of machine learning and AI projects don't succeed due to a lack of quality data *Source

Up to **40%** of any AI project effort is spent on data preparation *Source

In pursuit of the best – rather than just the fastest – data-labeling service

AI can automate key business tasks, provide intelligent analyses and predictions, and even power technological products and services that remove the human element from the equation (e.g., autonomous vehicles). Going from concept to production, however, can be a lengthy process and for models to work, they need quality training data.

Data needs to be cleaned and accurately annotated (to create ground-truth datasets), and the learning algorithms rigorously trained, validated, and tested, before AI solutions can provide expected outcomes.

Although in-house data labeling would, in theory, offer the highest quality labeling possible, for large-scale projects, it is simply not feasible or profitable for a company to have their employees label terabytes of data. Most organizations need to acquire the necessary training and testing datasets through other means.

Constantly accelerating and increasingly competitive market conditions have pushed organizations to opt for quick and dirty solutions in the rush to integrate AI into their business. However, many of hasty decisions only lead to a project that's over budget, behind schedule, and that doesn't meet expectations.

Responding to this need, a growing number of companies promise high-quality data-labeling services, with some even boasting access to hundreds of thousands of crowdsourced data annotators for every business need.

However, we have found that they rely mostly on a worldwide pool of annotators with the error rates and speeds associated with human beings rather than combining people with the right technology to accelerate the data-labeling process and ensure data security and privacy.

It takes the right balance between the two capabilities to create quality datasets fast and with high accuracy.

Looking beyond speed

There is no denying that speed is a crucial factor in the hunt for a reliable data-labeling service provider – as faster data annotation tends to mean earlier business benefits from AI – but there are four others:

Quality – high-quality labeled datasets will yield accurate AI models faster

Flexibility – the solution design should be customized to accommodate new challenges as they occur

Scalability – the annotation workforce must be matched to the demand for data

Cost – managed annotation services need to be competitively priced.

Drive enhanced outcomes through leveraging three tiers of data expertise

The only way a data-labeling service provider can guarantee speed, quality, scalability, and flexibility at a low cost is to have:

- A competent, project certified annotation workforce of approved domain experts.
- ML-assisted accelerators to cut the data-annotation time and effort.
- The capability to generate synthetic data, to fill in any data gaps, in the push to reach the highest accuracy in building algorithms.

Manual data labeling

Data needs to be labeled fast and efficiently with an attention to precision. To serve both generic and advanced annotation tasks, the right partner should have a dedicated training curriculum created for each project and supported by a choice of technical solutions to match project-specific requirements.

It should preferably have its own teams of data scientists and access to trusted, certified domain experts that can quickly support project needs. Additionally, a full audit trail and privacy assurance will put organizations at ease that their data is safe and secure.

ML-assisted labeling

Having humans in the loop who gather and prepare data for use is vital. But relying just on the human factor for data annotation will slow down a project considerably.

Data labeling can be accelerated through ML itself. By having your own ML model pre-label the data, the annotation effort can be reduced by up to 70% for single task classification.

Even if a medium level of accuracy can be attained with the help of technology, it will mean less manual work to get to the accuracy level we desire and allow annotators to shift focus from data labeling to data validation and testing.

There are three categories of ML-assisted labeling:

- **Assisted** – a data-labeling operator receives automated support to speed up decisions and labeling,
- **Pre-labeled** – a task is first run by an automated solution, then is augmented by an operator,
- **Automated** – a task is automated with a predictable and reproducible outcome; an operator will only review the output.

ML-assisted labeling could be further augmented by either pre-trained or project-specific solutions that increase the level of automation, allowing ML algorithms to process more data, faster. An experienced service provider will be able to select the best category and the right balance of human and technology for each task.

Synthetic data generation

Generally, the more high-quality data, the better, as an insufficient quantity of data will produce inaccurate and unpredictable results. Generative adversarial networks (GANs) can be used effectively in ML applications to create never-before-seen datasets.

Composed of two parts (the generator and discriminator), the generator produces artificial data by progressively learning through patterns based on real-world inputs while the discriminator discards implausible data. The goal is to get to a point where the discriminator is fooled to thinking the generated data is real – making the final output indistinguishable from naturally obtained datasets.

To get started, a good representative sample of real data is required, as the generator cannot create something from nothing. For generating quality tabular synthetic data, the bare minimum is 500 rows of data, while for images, a sample dataset of 575 images should yield impressive results.

Synthetic data when mixed with real data can improve model performance. As a proof of concept, Capgemini built a handwriting recognition AI model using 180 real images of documents created by hand. After testing our model, the average precision was 95.09%.

By adding an additional 200 synthetically generated documents created by ADA (Artificial Data Sampler), our data generation tool, the accuracy performance rate increased to 97.3%. This is a significant boost, considering that it usually takes vast amounts of data to increase accuracy beyond the 90% threshold.

In industries such as healthcare, data can be difficult to acquire and use mainly due to data privacy regulations such as GDPR. And methods to anonymize data don't always work since roughly 90% of this data can be re-identified with just four attributes. Companies can face hefty fines if the data can be traced back to individuals. This poses a challenge.

Precision is paramount in the medical field, so if there is not enough training data available, ML models will be ineffective at identifying anomalies or making correct diagnoses. In this case, synthetic data (medical records for code text classification, X-rays, magnetic resonance images, hospital discharge data, etc.) can be generated, then labeled according to preset parameters to create a truly anonymous, comprehensive training dataset.

About Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided everyday by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of over 340,000 team members in more than 50 countries. With its strong 55-year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fueled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2021 global revenues of €18 billion.

Get the Future You Want

www.capgemini.com

About the Authors

Vijay Bansal has extensive experience working in map production, geo-spatial data production, management, data labeling and annotation, and validation roles.

Marek Sowa is head of Capgemini's Intelligent Automation Offering & Innovation focused on adopting AI technologies into business services. He leverages the potential hidden in deep and machine learning to increase the speed, accuracy, and automation of processes. This helps clients to transform their business operations leveraging the combined power of AI and RPA to create working solutions that deliver real business value.

BSv MACS_CS_20220901_PT