Stanford CS 329X - Human-Centered NLP

Lecture 9: Beyond Benchmarking

05/01/2023

Lecturer: Prof. Diyi Yang *Readings: Douwe Kiela [2021] Sambasivan et al. [2021]*
Scribe: Alexander Peng

Key phrases: Benchmarking in AI, issues with benchmarking, human-centered benchmarking, metrics, interactive AI systems, generative AI agents, AI-to-AI interaction, language variations, dialects, co-designing with native speakers, explainable evaluation, fine-grained benchmarks.

In this lecture, we delve into the realm of human-centered benchmarking in the field of artificial intelligence and machine learning, with a particular focus on the challenges and opportunities it presents. Benchmarking plays a pivotal role in tracking progress and refining models in AI research. It encompasses the use of **one or more datasets, associated metrics, and performance aggregation to assess and compare different models.**

1 Importance of Benchmarking in Machine Learning and AI

Benchmarking is crucial in machine learning and artificial intelligence (AI) as it provides a means to track the progress of research and development. By establishing a set of consistent and reproducible evaluation criteria, it becomes possible to compare and contrast the performance of various algorithms and models. A benchmark typically consists of one or more datasets, associated metrics, and an aggregation of performance scores.

1.1 History of Benchmarking

The concept of benchmarking has its roots in the field of computer hardware comparison, where it was used to evaluate the performance of different hardware components. With the growth of AI and machine learning, benchmarking has evolved to encompass the evaluation of algorithms and models as well. Government agencies and industry organizations have played a significant role in fostering the development of benchmarks by organizing evaluations and workshops to compare and improve system performance.

2 Issues with Benchmarking

2.1 Saturation

Many North Star benchmarks, such as ImageNet, Glue, and Squad, have been saturated by AI systems that surpass human performance. This presents a challenge in determining progress, as it

becomes increasingly difficult to differentiate between small performance improvements when AI systems already outperform humans.

2.2 Artifacts

The construction of datasets may inadvertently introduce artifacts that can be exploited by AI systems, leading to misleading performance evaluations. Adversarial attacks, in which AI models are subjected to carefully crafted input modifications, can reveal these issues and help identify weaknesses in AI performance.

2.3 Alignment

A significant challenge in benchmarking is ensuring that the evaluation captures relevant aspects of a given task. Test-time performance may not always be a good proxy for real-world performance, as AI systems may perform well on specific datasets but fail to generalize to more diverse and complex scenarios.

2.4 Overfitting

Leaderboard chasing and gaming the system have become issues in the AI research community. Models may be optimized to perform well on specific benchmarks but suffer from poor generalization or instability when exposed to small perturbations in input data.

2.5 Reproducibility and Stability

Qualitative evaluations can be subjective and difficult to replicate, leading to inconsistencies in the assessment of model performance. Additionally, models that perform well on benchmarks may not always generalize well to real-world cases, raising questions about the stability and utility of such systems.

3 Human-Centered Benchmarking

3.1 Language Variations and Dialects

Natural languages exhibit systematic variations that lead to the formation of dialects. AI performance may vary significantly based on the dialect used, as most models are optimized for standard language varieties and may struggle with non-standard forms.

3.2 Co-designing with Native Speakers

Collaborating with native speakers to understand their concerns and experiences with technology is crucial in developing benchmarks that are representative of diverse populations. This collaboration

can help ensure that AI systems are more inclusive and better suited to serve the needs of various user groups.

3.3 Performance on Dialect Benchmarks

AI performance may degrade significantly when evaluated on dialect-specific benchmarks. This highlights the need for more inclusive and representative benchmark development and the importance of considering the diverse language needs of users.

4 Metrics

4.1 Designing Good Metrics Requires Domain Expertise

Developing effective metrics for benchmarking requires domain expertise and a deep understanding of the downstream tasks and language variations. Good metrics should also highlight trade-offs between different settings and be updated and refined over time to account for advancements in the field and changing requirements.

4.2 Explainable Evaluation

Providing a fine-grained breakdown of model performance across different dimensions can help in creating more explainable evaluations. Aggregating performance across multiple metrics enables a more comprehensive understanding of the strengths and weaknesses of AI systems and aids in identifying areas for improvement.

5 Benchmarking in Interactive AI Systems

5.1 Generative AI Agents and AI-to-AI Interaction

Benchmarking interactive AI systems, such as generative AI agents and AI-to-AI interaction, presents unique challenges. Involving humans in the evaluation process is essential to ensure that the AI systems are designed and evaluated with user needs and expectations in mind.

5.2 Fine-grained Benchmarks that Capture Complex Details

Moving beyond binary classification tasks, evaluating AI performance in more interactive settings requires the development of fine-grained benchmarks that capture complex details and nuances. These benchmarks should account for the intricacies of human-AI interactions and evaluate system performance in a more holistic manner.

6 Conclusion

6.1 Reevaluating the Goals of Benchmarking in AI

As AI research and development continue to advance, it is crucial to reevaluate the goals of benchmarking in AI. This includes involving humans in the evaluation process and considering the needs of diverse populations to ensure that AI systems are inclusive and effective for a wide range of users.

6.2 Developing New Paradigms for Evaluating AI Performance in Interactive Systems

To effectively evaluate AI performance in interactive systems, new paradigms for benchmarking need to be developed. This involves creating benchmarks that account for the complexities of human-AI interactions and provide a more comprehensive assessment of system performance across a variety of dimensions.

References

Yixin Nie Divyansh Kaushik Atticus Geiger Zhengxuan Wu Bertie Vidgen et al. Douwe Kiela, Max Bartolo. Dynabench: Rethinking benchmarking in nlp. *arXiv preprint arXiv:2104.14337*, 2021.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M. Aroyo. everyone wants to do the model work, not the data work: Data cascades in high-stakes ai. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021.