Progress in Dynamic Adversarial Data Collection & Adventures in Multimodal Machine Learning

Douwe Kiela

👨‍💻 Hugging Face 📚 Stanford University

---

DynaBench 🎉 ML Commons 📚 UCL 🧪 The University of North Carolina at Chapel Hill 🧪 Carnegie Mellon University 🧪 The Alan Turing Institute

Simon Fraser University 🧪 Johns Hopkins University 🧪 Stanford University 🧪 Facebook AI
Outline

1. Dynabench
   a. Overview
   b. Common Objections & Misconceptions

2. Progress in Dynamic Adversarial Data Collection
   a. Humans and Models in Loops
   b. Dynamic Adversarial Training Data

3. Adventures in Multimodal ML
   a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
   b. Foundation Models: FLAVA
News / about me..
Rethinking benchmarking in AI

Dynabench ([dynabench.org](http://dynabench.org)) is...

- A research platform.
- A community-based scientific experiment.
- An effort to challenge current benchmarking dogma and help push the boundaries of AI research.

As the name says,
Benchmark saturation in NLP
What is our goal? What is language?

Do believe the hype: we’re decent (but not great) at (some) i.i.d. problems when we have enough data and don’t care about the worst case.

Don’t believe the hype: we are FAR from truly general language understanding that encompasses all of language’s recursive, structured, generative, productive, and creative nature.
The ability to REALLY understand language

(Madry, 2018; https://adversarial-ml-tutorial.org)
What we are doing..

True intent

Distributional statistics

Text

Model

Label
What we should be doing...

Measuring not in the average case, but in the **worst case**.

**True intent**

**Speaker**

**Manifestation of intent (Message)**

**Model**

**Inferred intent**
Dynamic adversarial data collection (ANLI; Nie et al. 2019)
Dynabench goals

Dynabench is a comprehensive benchmarking platform that tackles many well-known problems in benchmarking and model evaluation.

**Saturation**
As current benchmarks quickly saturate, the field loses valuable time creating new benchmarks.

**Bias**
Inadvertent annotator artifacts and other biases can lead to overfitting.

**Alignment**
Test set performance is not always a good proxy for performance in the real-world.

**Leaderboard Culture**
Focusing too much on leaderboard rankings hinders creative solutions to AI problems.

**Reproducibility**
Self-reported results cannot be trusted.

**Accessibility**
Models that perform well on current benchmarks are often not easily accessible to the community for probing, let alone to laypeople.

**Backward Compatibility**
New benchmark or dataset cannot easily re-evaluate old models on the new data.

**Utility**
Not everyone is optimizing for the same metric. Efficiency might be traded off against accuracy.
Dynabench roles
Live demo

I was served rather the opposite of haute cuisine.

This restaurant was baad!
Summary

Language is about strong generalization. Humans expect this from other linguistic agents. But our systems might be right for the wrong reasons.

Evaluation is broken, we are haunted by Vapnik’s ghost but as a field are moving well beyond train-on-train-test-on-test in our deployments of AI (i.e., thanks to large scale pretraining and widely deployed systems).

We need to rethink this. Can we make cyclical progress and do more direct testing with humans in the loop?
Outline

1. Dynabench
   a. Overview
   b. Common Objections & Misconceptions

2. Progress in Dynamic Adversarial Data Collection
   a. Humans and Models in Loops
   b. Dynamic Adversarial Training Data

3. Adventures in Multimodal ML
   a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
   b. Foundation Models: FLAVA
Dynabench is a research platform that changes over time. It’s not about having the right answers, it’s about changing the status quo: improving benchmarking will require experimentation.
Adversarialness

Dynabench does not require adversarialness. It’s straightforward to collect data with no model in the loop, or with many models in the loop, in an adversarial, collaborative, or other setting.
Dynabench is not about NLP. It supports many modalities and has tasks in multimodality, vision and (soon) speech. Tasks are currently in English but we’d like to change that.
Cost

Model-in-the-loop data collection can be more expensive, because it requires more creativity. It’s still unclear how this cost trades off against data quality - it might be worth it. There are ways to drive the cost down.
Naturalness and distributional shift

Open question how natural the data is, or how natural it will be in the long term.
Being “fair” to the model in the loop

Dynabench aims to measure how well models would hold up if they were deployed “in the wild”. The vMER metric is “fair” in that sense for comparing models that were both put in the loop.

Test sets created with a specific model in the loop are not fair to that model. I don’t really care about that sort of fairness: a) we could simply put a diverse ensemble of models in the loop; b) we are unfair at a specific point in time, but there will be many more other models to come - this is about measuring these future models.
All rounds count

Dynabench collects data over many rounds. ALL previous rounds, including non-adversarial ones, should be used for testing. An NLI model should perform well on all NLI datasets, adversarial or non-adversarial.
Adversarial filtering

Crowdworkers are paid for every example they generate, including the ones that did not fool any model. Non-model-fooling data is generally not discarded, because it’s still useful. Different tasks/datasets filter data in different ways.
Adversarially-collected test sets

- “the constraint that a specified system must fail on the test examples makes it difficult to infer much from absolute measures of test-set performance: As long as a model makes any errors at all on any possible inputs, then we expect it to be possible to construct an adversarial test set against the model, and we expect the model to achieve zero test accuracy on that test set”
  - In Dynabench we advocate for looking at many metrics, including the time it takes a crowdworker to fool a model and how many times a crowdworker needs to try before succeeding (vMER). Not just “did you fool” but also “how easy was it to fool”.
  - The errors are not just any kind of error, they are things that humans easily get right and agree on. You need to look at the data itself! (So this reduces to the naturalness objection?)
- “We can further infer that any models that are sufficiently similar to the adversary should also perform very poorly on this test set, regardless of their ability”
  - What about the example of BERT performing below chance on WSC while deBERTa gets 96%? What does “sufficiently similar” mean and who gets to determine that?
  - What about “all models” or “a representative subset of all models”?
- “Neither of these observations would tell us anything non-trivial about the actual abilities of the models” double neg => “These observations only tell us trivial things about the ability of models”
  - That feels a bit strong?
Absolute performance numbers

- “Absolute performance numbers on adversarially-collected test sets are meaningless as measures of model capabilities”

What makes a performance number meaningless? This seems to assume (again) that we haven’t looked at the actual data. If a human can easily get the right answer and humans (mostly) agree about a given answer, and the example is natural, why should performance on that example be meaningless?

If this is about naturalness of data, are Turker-collected free-form test sets guaranteed to be more natural?
The scientific process

“significant further work is needed to avoid catastrophe. This will be difficult to achieve without a clear accounting of the abilities and limitations of current and plausible near-future systems”

Exactly this! We should, as a field, work hard to develop a clearer picture of our current capabilities and fix measurement. If we are saturating benchmarks, while we know we have all these issues, something is wrong.

In other words: be careful when you deploy a model and think about what you’re doing. We want the world to realize that evaluation is something we should take more seriously. If we can measure better, we can make better progress. This will happen by building on previous work, in cycles of progress where benchmarks “saturate” and are replaced by better once. Science will do its job if we are open to new ideas.
Outline

1. Dynabench
   a. Overview
   b. Common Objections & Misconceptions

2. Progress in Dynamic Adversarial Data Collection
   a. Humans and Models in Loops
   b. Dynamic Adversarial Training Data

3. Adventures in Multimodal ML
   a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
   b. Foundation Models: FLAVA
Broader research program

What happens when we put humans and models in loops?

Can we make faster progress? Can we make better measurements?

Can we have fewer biases and artifacts, and better robustness and alignment?

What are we still missing in our models?
What are the next challenges to solve?

How can we democratize model evaluation, help make research reproducible, learn from our mistakes as a community, and empower researchers?
Recent work out of the Dynabench team

- Kiela et al. (NAACL21). Dynabench: Rethinking Benchmarking in NLP
- Vidgen et al. (ACL21). Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection
- Potts et al. (ACL21). DynaSent: A Dynamic Benchmark for Sentiment Analysis
- Kirk et al. (2021). Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate
- Sheng & Singh et al. (NeurIPS21). Human-Adversarial Visual Question Answering
- Prasad et al. (Blackbox21). To what extent do human explanations of model behavior align with actual behavior?
- Wenzek et al. (2021). Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation
- Thrush et al. (2022). Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks
- Bartolo et al. (EMNLP21). Improving QA Model Robustness with Synthetic Adversarial Data Generation
- Kaushik et al. (ACL21). On the Efficacy of Adversarial Data Collection for Question Answering
- Bartolo et al. (2022). Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants
- Wallace et al. (2021). Analyzing Dynamic Adversarial Training Data in the Limit
Recent work out of the Dynabench team

- Kiela et al. (NAACL21). *Dynabench: Rethinking Benchmarking in NLP*
- Vidgen et al. (ACL21). *Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection*
- Potts et al. (ACL21). *DynaSent: A Dynamic Benchmark for Sentiment Analysis*
- Kirk et al. (2021). *Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate*
- Sheng & Singh et al. (NeurIPS21). *Human-Adversarial Visual Question Answering*
- Prasad et al. (Blackbox21). *To what extent do human explanations of model behavior align with actual behavior?*
- Wenzek et al. (2021). *Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation*
- Thrush et al. (2022). *Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks*
- Bartolo et al. (EMNLP21). *Improving QA Model Robustness with Synthetic Adversarial Data Generation*
- Kaushik et al. (ACL21). *On the Efficacy of Adversarial Data Collection for Question Answering*
- Bartolo et al. (2022). *Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants*
- Wallace et al. (2021). *Analyzing Dynamic Adversarial Training Data in the Limit*
Recent work out of the Dynabench team

- Kiela et al. (NAACL21). Dynabench: Rethinking Benchmarking in NLP
- Vidgen et al. (ACL21). Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection
- Potts et al. (ACL21). DynaSent: A Dynamic Benchmark for Sentiment Analysis
- Kirk et al. (2021). Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate
- Sheng & Singh et al. (NeurIPS21). Human-Adversarial Visual Question Answering
- Prasad et al. (Blackbox21). To what extent do human explanations of model behavior align with actual behavior?
- Wenzek et al. (2021). Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation
- Thrush et al. (2022). Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks
- Bartolo et al. (EMNLP21). Improving QA Model Robustness with Synthetic Adversarial Data Generation
- Kaushik et al. (ACL21). On the Efficacy of Adversarial Data Collection for Question Answering
- Bartolo et al. (2022). Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants
- Wallace et al. (2021). Analyzing Dynamic Adversarial Training Data in the Limit
Recent work out of the Dynabench team

- Kiela et al. (NAACL21). Dynabench: Rethinking Benchmarking in NLP
- Vidgen et al. (ACL21). Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection
- Potts et al. (ACL21). DynaSent: A Dynamic Benchmark for Sentiment Analysis
- Kirk et al. (2021). Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate
- Sheng & Singh et al. (NeurIPS21). Human-Adversarial Visual Question Answering
- Prasad et al. (Blackbox21). To what extent do human explanations of model behavior align with actual behavior?
- Wenzek et al. (2021). Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation
- Thrush et al. (2022). Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks

Method Papers

- Bartolo et al. (EMNLP21). Improving QA Model Robustness with Synthetic Adversarial Data Generation
- Kaushik et al. (ACL21). On the Efficacy of Adversarial Data Collection for Question Answering
- Bartolo et al. (2022). Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants
- Wallace et al. (2021). Analyzing Dynamic Adversarial Training Data in the Limit
Recent work out of the Dynabench team

- Kiela et al. (NAACL21). *Dynabench: Rethinking Benchmarking in NLP*
- Vidgen et al. (ACL21). *Learning from the Worst: Dynamically Generated Datasets Improve Online Hate Detection*
- Potts et al. (ACL21). *DynaSent: A Dynamic Benchmark for Sentiment Analysis*
- Kirk et al. (2021). *Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate*
- Sheng & Singh et al. (NeurIPS21). *Human-Adversarial Visual Question Answering*
- Prasad et al. (Blackbox21). *To what extent do human explanations of model behavior align with actual behavior?*
- Wenzek et al. (2021). *Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation*
- Thrush et al. (2022). *Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks*
- Bartolo et al. (EMNLP21). *Improving QA Model Robustness with Synthetic Adversarial Data Generation*
- Kaushik et al. (ACL21). *On the Efficacy of Adversarial Data Collection for Question Answering*
- Bartolo et al. (2022). *Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants*
- Wallace et al. (2021). *Analyzing Dynamic Adversarial Training Data in the Limit*
Humans and models in loops

- Question 1:
  - Instead of human-adversarial, how much can we improve things by just being model-adversarial using human-adversarial data?

- Question 2:
  - Can generative (adversarial) models help humans fool discriminative models?

Work by Max Bartolo et al.
Improving QA robustness with synthetic adversarial data

- Pipeline:
  1. Passage selection
  2. Answer candidate selection
  3. Question generation
  4. Filtering and re-labeling
  5. Training a new QA model

Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation

Max Bartolo†, Tristan Thrush‡, Robin Jia‡, Sebastian Riedel‡
†University College London, ‡Facebook AI Research

*Old English was not static, and its usage covered a period of 700 years, from the Anglo-Saxon settlement of Britain in the 5th century to the late 11th century... Albert Baugh dates Old English from 450 to 1150, a period of full inflections, a synthetic language. Perhaps around 85 per cent...*
## Step 2: Answer selection

<table>
<thead>
<tr>
<th>Method</th>
<th>#QA pairs</th>
<th>$\mathcal{D}_{\text{SQuAD}}$</th>
<th>$\mathcal{D}_{\text{BiDAF}}$</th>
<th>$\mathcal{D}_{\text{BERT}}$</th>
<th>$\mathcal{D}_{\text{RoBERTa}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>POS Extended</td>
<td>999,034</td>
<td>53.8</td>
<td>71.4</td>
<td>32.7</td>
<td>46.9</td>
</tr>
<tr>
<td>Noun Chunks</td>
<td>581,512</td>
<td>43.3</td>
<td>63.7</td>
<td>28.7</td>
<td>43.1</td>
</tr>
<tr>
<td>Named Entities</td>
<td>257,857</td>
<td>54.2</td>
<td>69.7</td>
<td>30.5</td>
<td>42.5</td>
</tr>
<tr>
<td>Span Extraction</td>
<td>377,774</td>
<td>64.7</td>
<td>80.1</td>
<td>37.8</td>
<td>53.9</td>
</tr>
<tr>
<td>SAL (ours)</td>
<td>566,730</td>
<td>68.2</td>
<td><strong>82.6</strong></td>
<td>43.2</td>
<td>59.3</td>
</tr>
<tr>
<td>SAL threshold (ours)</td>
<td>393,164</td>
<td><strong>68.5</strong></td>
<td>82.0</td>
<td><strong>46.0</strong></td>
<td><strong>60.3</strong></td>
</tr>
</tbody>
</table>

Table 2: Downstream test results for a RoBERTa\textsubscript{Large} QA model trained on synthetic data generated using different answer selection methods combined with a BART\textsubscript{Large} question generator (trained on SQuAD\textsubscript{10k} + $\mathcal{D}_{\text{AQA}}$).
Step 3: Question generation

<table>
<thead>
<tr>
<th>Method</th>
<th>#QA pairs</th>
<th>$D_{\text{SQuAD}}$</th>
<th>$D_{\text{BiDAF}}$</th>
<th>$D_{\text{BERT}}$</th>
<th>$D_{\text{RoBERTa}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{RSQuAD}$</td>
<td>87,599</td>
<td>73.2</td>
<td>64.3</td>
<td>31.3</td>
<td>16.1</td>
</tr>
<tr>
<td>$\text{RSQuAD+AQA}$</td>
<td>117,599</td>
<td>74.2</td>
<td>66.9</td>
<td>53.9</td>
<td>43.4</td>
</tr>
<tr>
<td>$\text{SQuAD}_{10k}$</td>
<td>87,598</td>
<td>69.2</td>
<td>52.1</td>
<td>22.4</td>
<td>13.9</td>
</tr>
<tr>
<td>$D_{\text{BiDAF}}$</td>
<td>87,598</td>
<td>67.1</td>
<td>56.5</td>
<td>33.1</td>
<td>22.0</td>
</tr>
<tr>
<td>$D_{\text{BERT}}$</td>
<td>87,598</td>
<td>67.4</td>
<td>51.1</td>
<td>30.3</td>
<td>18.8</td>
</tr>
<tr>
<td>$D_{\text{RoBERTa}}$</td>
<td>87,598</td>
<td>63.4</td>
<td>47.9</td>
<td>27.2</td>
<td>20.6</td>
</tr>
<tr>
<td>$D_{\text{AQA}}$</td>
<td>87,598</td>
<td>65.5</td>
<td>53.0</td>
<td>31.1</td>
<td>23.2</td>
</tr>
<tr>
<td>$\text{SQuAD}<em>{10k} + D</em>{\text{AQA}}$</td>
<td>87,598</td>
<td>71.9</td>
<td>58.8</td>
<td>32.9</td>
<td>19.1</td>
</tr>
</tbody>
</table>

Table 5: Downstream QA test results using generative models trained on different source data. We compare these results to baseline RoBERTa models trained on SQuAD, and on the combination of SQuAD and AdversarialQA.
### Step 4: Filtering and self-training

<table>
<thead>
<tr>
<th>Filtering Method</th>
<th>#QA pairs</th>
<th>$D_{SQuAD}$</th>
<th>$D_{BiDAF}$</th>
<th>$D_{BERT}$</th>
<th>$D_{RoBERTa}$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$EM$</td>
<td>$F_1$</td>
<td>$EM$</td>
<td>$F_1$</td>
</tr>
<tr>
<td>Answer Candidate Conf. ($thresh = 0.6$)</td>
<td>362,281</td>
<td>68.4</td>
<td>82.4</td>
<td>42.9</td>
<td>57.9</td>
</tr>
<tr>
<td>Question Generator Conf. ($thresh = 0.3$)</td>
<td>566,725</td>
<td>69.3</td>
<td>83.1</td>
<td>43.5</td>
<td>58.9</td>
</tr>
<tr>
<td>Influence Functions</td>
<td>288,636</td>
<td>68.1</td>
<td>81.9</td>
<td>43.7</td>
<td>58.6</td>
</tr>
<tr>
<td>Ensemble Roundtrip Consistency (6/6 correct)</td>
<td>250,188</td>
<td>74.2</td>
<td>86.2</td>
<td>55.1</td>
<td>67.7</td>
</tr>
<tr>
<td>Self-training (ST)</td>
<td>528,694</td>
<td>74.8</td>
<td>87.0</td>
<td>53.9</td>
<td>67.9</td>
</tr>
<tr>
<td>Answer Candidate Conf. ($thresh = 0.5$) &amp; ST</td>
<td>380,785</td>
<td><strong>75.1</strong></td>
<td><strong>87.0</strong></td>
<td><strong>56.5</strong></td>
<td><strong>70.0</strong></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Downstream QA test results for different filtering strategies, showing best hyper-parameter settings.
Findings

- Synthetic adversarial data derived from human-adversarial data **improves accuracy** and **robustness**.

<table>
<thead>
<tr>
<th>Model</th>
<th>Training Data</th>
<th>$\mathcal{D}_{BIDAF}$</th>
<th>$\mathcal{D}_{BERT}$</th>
<th>$\mathcal{D}_{ROBERTa}$</th>
<th>mvMER*</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{SQuAD}$</td>
<td>SQuAD</td>
<td>48.6 ± 1.3</td>
<td>64.2 ± 1.5</td>
<td>30.9 ± 1.3</td>
<td>43.3 ± 1.7</td>
</tr>
<tr>
<td>$R_{SQuAD+AQA}$</td>
<td>↑ + AQA</td>
<td>59.6 ± 0.5</td>
<td>73.9 ± 0.5</td>
<td>54.8 ± 0.7</td>
<td>64.8 ± 0.9</td>
</tr>
<tr>
<td>SynQA</td>
<td>↑ + SynQA$_{SQuAD}$</td>
<td>62.5 ± 0.9</td>
<td>76.0 ± 1.0</td>
<td>58.7 ± 1.4</td>
<td>68.3 ± 1.4</td>
</tr>
<tr>
<td>SynQA$_{Ext}$</td>
<td>↑ + SynQA$_{Ext}$</td>
<td><strong>62.7 ± 0.6</strong></td>
<td><strong>76.2 ± 0.5</strong></td>
<td><strong>59.0 ± 0.7</strong></td>
<td><strong>68.9 ± 0.5</strong></td>
</tr>
</tbody>
</table>

**MRQA in-domain**

<table>
<thead>
<tr>
<th>Model</th>
<th>SQuAD</th>
<th>NewsQA</th>
<th>TriviaQA</th>
<th>SearchQA</th>
<th>HotpotQA</th>
<th>NQ</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{SQuAD}$</td>
<td>84.1 ± 1.3</td>
<td>90.4 ± 1.3</td>
<td>41.0 ± 1.2</td>
<td>57.5 ± 1.6</td>
<td>60.2 ± 0.3</td>
<td>69.0 ± 0.3</td>
<td>16.0 ± 1.8</td>
</tr>
<tr>
<td>$R_{SQuAD+AQA}$</td>
<td>84.4 ± 1.0</td>
<td>90.2 ± 1.1</td>
<td>41.7 ± 1.6</td>
<td>58.0 ± 1.7</td>
<td><strong>62.7 ± 0.4</strong></td>
<td><strong>70.8 ± 0.3</strong></td>
<td>20.6 ± 2.9</td>
</tr>
<tr>
<td>SynQA</td>
<td>88.8 ± 0.3</td>
<td><strong>94.3 ± 0.2</strong></td>
<td>42.9 ± 1.6</td>
<td>60.0 ± 1.4</td>
<td>62.3 ± 1.1</td>
<td>70.2 ± 1.3</td>
<td>23.7 ± 1.3</td>
</tr>
<tr>
<td>SynQA$_{Ext}$</td>
<td><strong>89.0 ± 0.3</strong></td>
<td><strong>94.3 ± 0.2</strong></td>
<td><strong>46.2 ± 0.9</strong></td>
<td><strong>63.1 ± 0.8</strong></td>
<td><strong>58.1 ± 1.8</strong></td>
<td><strong>65.5 ± 1.0</strong></td>
<td><strong>28.7 ± 3.2</strong></td>
</tr>
</tbody>
</table>

**MRQA out-of-domain**

<table>
<thead>
<tr>
<th>Model</th>
<th>BioASQ</th>
<th>DROP</th>
<th>DuoRC</th>
<th>RACE</th>
<th>RelationExt.</th>
<th>TextbookQA</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R_{SQuAD}$</td>
<td>53.2 ± 1.1</td>
<td>68.6 ± 1.4</td>
<td>39.8 ± 2.6</td>
<td>52.7 ± 2.2</td>
<td>49.3 ± 0.7</td>
<td>60.3 ± 0.8</td>
<td>35.1 ± 1.0</td>
</tr>
<tr>
<td>$R_{SQuAD+AQA}$</td>
<td>54.6 ± 1.2</td>
<td>69.4 ± 0.8</td>
<td>59.8 ± 1.3</td>
<td>68.4 ± 1.5</td>
<td><strong>51.8 ± 1.1</strong></td>
<td><strong>62.2 ± 0.9</strong></td>
<td>38.4 ± 9.9</td>
</tr>
<tr>
<td>SynQA</td>
<td><strong>55.1 ± 1.5</strong></td>
<td>68.7 ± 1.2</td>
<td>64.3 ± 1.5</td>
<td>72.5 ± 1.7</td>
<td>51.7 ± 1.3</td>
<td>62.1 ± 0.9</td>
<td><strong>40.2 ± 1.2</strong></td>
</tr>
<tr>
<td>SynQA$_{Ext}$</td>
<td>54.9 ± 1.3</td>
<td>68.5 ± 0.9</td>
<td><strong>64.9 ± 1.1</strong></td>
<td><strong>73.0 ± 1.0</strong></td>
<td>48.8 ± 1.2</td>
<td>58.0 ± 1.0</td>
<td>38.6 ± 9.4</td>
</tr>
</tbody>
</table>

SynQA models are much harder to fool (i.e. more robust)

SynQA outperforms alternatives
Empowering crowdworkers with generative assistants

- We know now that generative models trained on adversarial data can help make models more robust.
- Can we use those models to help humans fool models as “generative adversarial assistants”? ModelS in the loop!
  a. Adversarial data is expensive - can it be made cheaper?
  b. Adversarial data can be noisy - can it be made higher quality?

Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants

Max Bartolo* Tristan Thrush† Sebastian Riedel†*
Pontus Stenetorp* Robin Jia†† Douwe Kiela†
*UCL †USC †Facebook AI Research
A hole is classified by its par, meaning the number of strokes a skilled golfer should require to complete play of the hole. The minimum par of any hole is 3 because par always includes a stroke for the tee shot and two putts. Pars of 4 and 5 strokes are ubiquitous on golf courses; more rarely, a few courses feature par-6 and even par-7 holes. Strokes other than the tee shot and putts are expected to be made from the fairway; for example, a skilled golfer expects to reach the green on a par-4 hole in two strokes—one from the...
Standard (SDC) vs Adversarial (ADC) Data Collection

- Earlier finding: “Across a variety of [Question Answering] models and datasets, we find that models trained on adversarial data usually perform better on other adversarial datasets but worse on a diverse collection of out-of-domain evaluation sets.” (Divyansh Kaushik et al. ACL 2021)

On the Efficacy of Adversarial Data Collection for Question Answering: Results from a Large-Scale Randomized Study

Divyansh Kaushik†, Douwe Kiela‡, Zachary C. Lipton†, Wen-tau Yih‡
New finding:
(the preliminary take-away on smallish data is – be careful with setup)

<table>
<thead>
<tr>
<th>Adversary-in-the-loop?</th>
<th>Median time per example</th>
<th>Validated model error rate</th>
<th>Time per model-fooling ex</th>
<th>Domain generalization</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>t (s)</td>
<td>vMER (%)</td>
<td>t/vMFE (s)</td>
<td>SQuAD$<em>{dev}$ D$</em>{BIDAF}$ D$<em>{BERT}$ D$</em>{RoBERTa}$ MRQA</td>
</tr>
<tr>
<td>□</td>
<td>56.3$^{23.6}$</td>
<td>0.63</td>
<td>11274</td>
<td>45.4 14.7 9.2 8.8 25.2</td>
</tr>
<tr>
<td>✓</td>
<td>61.2$^{27.4}$</td>
<td>1.62</td>
<td>4863</td>
<td>82.0 44.4 29.2 22.4 53.8</td>
</tr>
</tbody>
</table>

Standard (SDC) vs Adversarial (ADC) Data Collection
Using a Generator-in-the-loop makes standard data collection more faster and much higher quality.

Sampling strategies:
  a. Likelihood: sample candidates according to the generative model’s overall likelihood.
  b. Adversarial: sample questions according to the lowest F1 score queried against a QA model.
  c. Uncertainty: select generated questions in order of the least span selection confidence when queried against a QA model.

<table>
<thead>
<tr>
<th>Sampling Strategy</th>
<th>t (s)</th>
<th>vMER (%)</th>
<th>t/vMFE (s)</th>
<th>SQuAD\textsubscript{dev}</th>
<th>D\textsubscript{BiDAF}</th>
<th>D\textsubscript{BERT}</th>
<th>D\textsubscript{RoBERTa}</th>
<th>MRQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>40.2_24.8</td>
<td>0.69</td>
<td>6331</td>
<td>53.6</td>
<td>15.9</td>
<td>11.0</td>
<td>9.9</td>
<td>31.4</td>
</tr>
<tr>
<td>Adversarial</td>
<td>56.7_23.8</td>
<td>3.22</td>
<td>2277</td>
<td>80.1</td>
<td>39.1</td>
<td>21.1</td>
<td>18.7</td>
<td>49.5</td>
</tr>
<tr>
<td>Uncertainty</td>
<td>56.9_25.1</td>
<td>2.93</td>
<td>2643</td>
<td>80.1</td>
<td>40.1</td>
<td>24.3</td>
<td>22.6</td>
<td>51.1</td>
</tr>
</tbody>
</table>
Using a Generator-in-the-loop makes adversarial data collection as fast as standard data collection, with higher quality and better domain generalization.

Generative Annotation Assistant (GAA) trained on SQuad, AQA or Combined.

<table>
<thead>
<tr>
<th>GAA Training</th>
<th>Sampling</th>
<th>t (s)</th>
<th>vMER (%)</th>
<th>t/vMFE (s)</th>
<th>SQuAD_{dev}</th>
<th>D_{BiDAF}</th>
<th>D_{BERT}</th>
<th>D_{RoBERTa}</th>
<th>MRQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>Likelihood</td>
<td>66.2</td>
<td>2.40</td>
<td>3489</td>
<td>81.2</td>
<td>44.2</td>
<td>27.8</td>
<td>21.3</td>
<td>52.3</td>
</tr>
<tr>
<td>SQuAD</td>
<td>Adversarial</td>
<td>63.3</td>
<td>2.87</td>
<td>2831</td>
<td>80.2</td>
<td>41.7</td>
<td>28.8</td>
<td>20.9</td>
<td>49.3</td>
</tr>
<tr>
<td>SQuAD</td>
<td>Uncertainty</td>
<td>65.7</td>
<td>2.34</td>
<td>3505</td>
<td>82.6</td>
<td>45.1</td>
<td>29.0</td>
<td>23.0</td>
<td>52.4</td>
</tr>
<tr>
<td>AdversarialQA</td>
<td>Likelihood</td>
<td>59.0</td>
<td>2.63</td>
<td>3034</td>
<td>79.9</td>
<td>40.8</td>
<td>30.2</td>
<td>24.9</td>
<td>52.6</td>
</tr>
<tr>
<td>AdversarialQA</td>
<td>Adversarial</td>
<td>64.7</td>
<td>3.95</td>
<td>2077</td>
<td>75.7</td>
<td>38.7</td>
<td>28.8</td>
<td>23.1</td>
<td>50.3</td>
</tr>
<tr>
<td>AdversarialQA</td>
<td>Uncertainty</td>
<td>66.7</td>
<td>3.79</td>
<td>2305</td>
<td>78.3</td>
<td>41.9</td>
<td>29.4</td>
<td>22.9</td>
<td>51.0</td>
</tr>
<tr>
<td>Combined</td>
<td>Likelihood</td>
<td>52.7</td>
<td>2.51</td>
<td>2827</td>
<td>79.6</td>
<td>40.7</td>
<td>29.9</td>
<td>24.2</td>
<td>53.3</td>
</tr>
<tr>
<td>Combined</td>
<td>Adversarial</td>
<td>71.0</td>
<td>2.76</td>
<td>3450</td>
<td>78.7</td>
<td>39.8</td>
<td>26.6</td>
<td>22.0</td>
<td>49.6</td>
</tr>
<tr>
<td>Combined</td>
<td>Uncertainty</td>
<td>66.7</td>
<td>3.08</td>
<td>2854</td>
<td>81.0</td>
<td>44.0</td>
<td>26.4</td>
<td>22.2</td>
<td>52.7</td>
</tr>
</tbody>
</table>
Improving ADC further

- If you do “answer prompting” where you don’t force annotators to pick the answer but suggest one, ADC gets even faster and much higher quality.
- Starting point, traditional data collection: vMER=0.63 with t=56.3
- End point, ADC with GAA: vMER=6.08 with t=43.8

<table>
<thead>
<tr>
<th>GAA Training</th>
<th>Sampling</th>
<th>t (s)</th>
<th>vMER (%)</th>
<th>t/vME (s)</th>
<th>SQuAD_{dev}</th>
<th>D_{BiDAF}</th>
<th>D_{BERT}</th>
<th>D_{RoBERTa}</th>
<th>MRQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>AdversarialQA</td>
<td>Likelihood</td>
<td>49.9</td>
<td>6.08</td>
<td>1086</td>
<td>78.2</td>
<td>44.0</td>
<td>33.7</td>
<td>26.2</td>
<td>52.0</td>
</tr>
<tr>
<td>AdversarialQA</td>
<td>Adversarial</td>
<td>43.8</td>
<td>2.22</td>
<td>2587</td>
<td>79.9</td>
<td>44.2</td>
<td>30.6</td>
<td>23.6</td>
<td>52.1</td>
</tr>
<tr>
<td>AdversarialQA</td>
<td>Uncertainty</td>
<td>50.9</td>
<td>4.04</td>
<td>1667</td>
<td>80.4</td>
<td>42.8</td>
<td>28.8</td>
<td>22.1</td>
<td>51.1</td>
</tr>
<tr>
<td>Combined</td>
<td>Likelihood</td>
<td>49.0</td>
<td>2.72</td>
<td>2510</td>
<td>79.6</td>
<td>42.7</td>
<td>31.1</td>
<td>23.8</td>
<td>50.2</td>
</tr>
<tr>
<td>Combined</td>
<td>Adversarial</td>
<td>65.2</td>
<td>4.41</td>
<td>2042</td>
<td>80.2</td>
<td>44.7</td>
<td>31.5</td>
<td>24.8</td>
<td>53.0</td>
</tr>
<tr>
<td>Combined</td>
<td>Uncertainty</td>
<td>54.1</td>
<td>2.94</td>
<td>2740</td>
<td>81.1</td>
<td>44.8</td>
<td>27.9</td>
<td>23.8</td>
<td>51.2</td>
</tr>
</tbody>
</table>
A “new paradigm”? 

- ModelS in LoopS: 
  a. **Yes**, we can collect much higher quality data than static data using this method.
  b. **Yes**, we can collect higher quality data than regular human-and-model-in-the-loop.
  c. **Yes**, we can do so at a cost that is much lower than human-and-model-in-the-loop matching standard data collection.
Recent work out of the Dynabench team

- Kiela et al. (NAACL21). Dynabench: Rethinking Benchmarking in NLP
- Vidgen et al. (ACL21). Learning from the Worst: Dynamically GeneratedDatasets to Improve Online Hate Detection
- Potts et al. (ACL21). DynaSent: A Dynamic Benchmark for Sentiment Analysis
- Kirk et al. (2021). Hatemoji: A Test Suite and Dataset for Benchmarking and Detecting Emoji-based Hate
- Sheng & Singh et al. (NeurIPS21). Human-Adversarial Visual Question Answering
- Prasad et al. (Blackbox21). To what extent do human explanations of model behavior align with actual behavior?
- Wenzek et al. (2021). Findings of the WMT 2021 Shared Task on Large-Scale Multilingual Machine Translation
- Thrush et al. (2022). Dynatask: A Platform for Creating Dynamic AI Benchmark Tasks
- Bartolo et al. (EMNLP21). Improving QA Model Robustness with Synthetic Adversarial Data Generation
- Kaushik et al. (ACL21). On the Efficacy of Adversarial Data Collection for Question Answering
- Bartolo et al. (2022). Models in the Loop: Aiding Crowdworkers with Generative Annotation Assistants
- Wallace et al. (2021). Analyzing Dynamic Adversarial Training Data in the Limit
Dynamic adversarial data collection in the limit
Experimental setup

- Starting point: Roberta trained on “All NLI” (MNLI+SNLI+FEVER)
- We hand-construct an expert-curated test set covering a wide range of NLI phenomena.
- We do DADC for 20 rounds (ANLI only did 3).
- We select 10 contexts so that:
  a. We can afford collecting many rounds of data
  b. We have some hope of achieving saturation
  c. We have a broad range of phenomena
  d. We can create a wide-coverage test set

Work by Eric Wallace et al.
Findings: A virtuous cycle

Promising results when exploring Dynamic Adversarial Data Collection in the limit:
Findings: Diversity is key

- DADC data is more diverse, more complex and has fewer artifacts.
- DADC models get stronger over time.

<table>
<thead>
<tr>
<th></th>
<th>No Model</th>
<th>Static Model</th>
<th>Dynamic Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Diversity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unique Unigrams</td>
<td>4.0k</td>
<td>4.2k</td>
<td>4.3k</td>
</tr>
<tr>
<td>Unique Bigrams</td>
<td>23.3k</td>
<td>24.8k</td>
<td>25.6k</td>
</tr>
<tr>
<td>Inter-example Sim.</td>
<td>41.2</td>
<td>41.9</td>
<td>39.5</td>
</tr>
<tr>
<td><strong>Complexity</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntax</td>
<td>2.0</td>
<td>2.1</td>
<td>2.3</td>
</tr>
<tr>
<td>Reading Level</td>
<td>4.9</td>
<td>5.4</td>
<td>5.9</td>
</tr>
<tr>
<td>Length</td>
<td>10.1</td>
<td>10.9</td>
<td>12.1</td>
</tr>
<tr>
<td><strong>Artifacts</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hypo-only Acc %</td>
<td>75.4</td>
<td>69.3</td>
<td>69.7</td>
</tr>
<tr>
<td>Overlap Entail %</td>
<td>54.2</td>
<td>49.2</td>
<td>47.3</td>
</tr>
</tbody>
</table>
Method take-aways

- ADC looks like a good alternative for traditional crowdworker data collection.
- This is a nice side benefit, considering that the original goal was evaluation.
- Human-and-model-in-the-loop / human feedback holds a lot of promise (see OpenAI’s recent papers on this as well, or “red teaming”)
- Further work needed on many questions, including:
  a. How (un)natural is adversarial data and how much does that matter?
  b. How does dynamic adversarial data collection relate to active learning and continual learning?
  c. Can we incorporate knowledge about the model in the loop in our optimization procedures?
  d. Exploring ensembles in the loop, different scoring functions, etc.
What comes next?

We’ve opened everything up:

- Fully open source (MIT licensed)
- Dynatask: Anyone can add tasks
- Keep growing the community
- Keep pushing the boundaries
- Exploring synergies with 🤗?
Teaming up with ML Commons and DataPerf

MLCommons aims to answer the needs of the nascent machine learning industry through open, collaborative engineering in three areas:

**Benchmarking**

- Benchmarks provide consistent measurements of accuracy, speed, and efficiency. Consistent measurements enable engineers to design reliable products and services, and enable researchers to compare innovations and choose the best ideas to drive the solutions of tomorrow.

**Datasets**

- Datasets are the raw materials for all of machine learning. Models are only as good as the data they are trained on. Academics and entrepreneurs in particular depend on public datasets to create new technologies and new companies.

**Best Practices**

- Best Practices empower researchers and engineers to more easily exchange models, reproduce experiments, and build applications that leverages machine learning. Improving best practices accelerates progress in, and grows the market for, machine learning.

---

“Everyone wants to do the model work, not the data work”:
Data Cascades in High-Stakes AI

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, Lora Aroyo
[nithyasamba,kapania,highfill,akrong,paritosh,loras]@google.com
Google Research
Mountain View, CA
Outline

1. Dynabench
   a. Overview
   b. Common Objections & Misconceptions

2. Progress in Dynamic Adversarial Data Collection
   a. Humans and Models in Loops
   b. Dynamic Adversarial Training Data

3. Adventures in Multimodal ML
   a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
   b. Foundation Models: FLAVA
Vision and Language Tasks & Datasets

Citations as of 4/4/22:

- VQA/VQA2 - 3409/1227
- Visual Genome - 2776
- COCO - 1240 (22724)
- Flickr30k - 908
- VisDial - 715
- NLVR2 - 189

Power law distribution with VQA as the dominant task.

Source: Mogadala et al 2021
Visual Question Answering

VQA is plateauing and arguably/almost saturated

Source: Barbosa-Silva et al 2022
What do we want?

- Ideally, evaluation sets are:
  - High-quality and without error
  - Not too expensive
  - Not too easy
  - Discriminative between models
  - Realistic and representative of practical use-cases
  - Straightforwardly measured

- Multimodal evaluation sets, in addition, ideally are:
  - Not dominated by a specific modality
  - Actually measuring multimodal rather than unimodal performance
    (cf “making the V in VQA matter”)
Multimodal Evaluation

The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes

Douwe Kiela, Hamed Firooz, Aravind Mohan,
Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, Davide Testuggine

Human-Adversarial Visual Question Answering

Sasha Sheng†, Amanpreet Singh‡, Vedanuj Goswami†, Jose Alberto Lopez Magana†
Wojciech Galuba†, Devi Parikh†, Douwe Kiela‡
† Facebook AI Research; ‡ Tecnológico de Monterrey

Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality

Tristan Thrush†, Ryan Jiang‡, Max Bartolo‡, Amanpreet Singh§, Adina Williams†, Douwe Kiela§, Candace Ross††
† Hugging Face; ‡ Facebook AI Research; † University of Waterloo; † University College London
Hateful Memes

Motivated by the shortcomings of other V&L datasets: we need something that is harder, more realistic, and requires true multimodal reasoning and understanding.

“Mean meme” examples for illustrative purposes – not actually in the dataset
Hateful Memes

Highly trained annotators, so: decent quality but small and expensive

Key concept: benign confounders

A “challenge set” for the community to do zero-shot/finetuning from pretrained
Hateful Memes

Findings in the paper:

- Big gap with human performance.
- Region features (as opposed to grid) seem to help.
- Earlier fusion is better than middle, Is better than late.
- Multimodal pretraining doesn’t really work.

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Validation</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Acc.</td>
<td>AUROC</td>
</tr>
<tr>
<td>Human</td>
<td>-</td>
<td>-</td>
<td>84.70</td>
</tr>
<tr>
<td>Unimodal</td>
<td>Image-Grid</td>
<td>50.67</td>
<td>52.33</td>
</tr>
<tr>
<td></td>
<td>Image-Region</td>
<td>52.53</td>
<td>57.24</td>
</tr>
<tr>
<td></td>
<td>Text BERT</td>
<td>58.27</td>
<td>65.05</td>
</tr>
<tr>
<td>Multimodal (Unimodal Pretraining)</td>
<td>Late Fusion</td>
<td>59.39</td>
<td>65.07</td>
</tr>
<tr>
<td></td>
<td>Concat BERT</td>
<td>59.32</td>
<td>65.88</td>
</tr>
<tr>
<td></td>
<td>MMBT-Grid</td>
<td>59.59</td>
<td>66.73</td>
</tr>
<tr>
<td></td>
<td>MMBT-Region</td>
<td>64.75</td>
<td>72.62</td>
</tr>
<tr>
<td></td>
<td>ViLBERT</td>
<td>63.16</td>
<td>72.17</td>
</tr>
<tr>
<td></td>
<td>Visual BERT</td>
<td>65.01</td>
<td>74.14</td>
</tr>
<tr>
<td>Multimodal (Multimodal Pretraining)</td>
<td>ViLBERT CC</td>
<td>66.10</td>
<td>73.02</td>
</tr>
<tr>
<td></td>
<td>Visual BERT COCO</td>
<td>65.93</td>
<td>74.14</td>
</tr>
</tbody>
</table>
Hateful Memes Competition

After the paper came a $100k competition on an unseen test set:

<table>
<thead>
<tr>
<th>Type</th>
<th>Model</th>
<th>Unseen Dev Acc.</th>
<th>AUROC</th>
<th>Unseen Test Acc.</th>
<th>AUROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unimodal</td>
<td>Image-Region</td>
<td>61.48</td>
<td>53.54</td>
<td>60.28±0.18</td>
<td>54.64±0.80</td>
</tr>
<tr>
<td></td>
<td>Text BERT</td>
<td>60.37</td>
<td>60.88</td>
<td>63.60±0.54</td>
<td>62.65±0.40</td>
</tr>
<tr>
<td>Multimodal</td>
<td>Late Fusion</td>
<td>61.11</td>
<td>61.00</td>
<td>64.06±0.02</td>
<td>64.44±1.60</td>
</tr>
<tr>
<td></td>
<td>Concat BERT</td>
<td>64.81</td>
<td>65.42</td>
<td>65.90±0.82</td>
<td>66.28±0.66</td>
</tr>
<tr>
<td></td>
<td>MMBT-Grid</td>
<td>67.78</td>
<td>65.47</td>
<td>66.85±1.61</td>
<td>67.24±2.53</td>
</tr>
<tr>
<td></td>
<td>MMBT-Region</td>
<td>70.04</td>
<td>71.54</td>
<td>70.10±1.39</td>
<td>72.21±0.20</td>
</tr>
<tr>
<td></td>
<td>ViLBERT</td>
<td>69.26</td>
<td>72.73</td>
<td>70.86±0.70</td>
<td>73.39±1.32</td>
</tr>
<tr>
<td></td>
<td>Visual BERT</td>
<td>69.67</td>
<td>71.10</td>
<td>71.30±0.68</td>
<td>73.23±1.04</td>
</tr>
<tr>
<td>Multimodal</td>
<td>ViLBERT CC</td>
<td>70.37</td>
<td>70.78</td>
<td>70.03±1.07</td>
<td>72.78±0.50</td>
</tr>
<tr>
<td></td>
<td>Visual BERT COCO</td>
<td>70.77</td>
<td>73.70</td>
<td>69.95±1.06</td>
<td>74.59±1.56</td>
</tr>
</tbody>
</table>

Winner characteristics: frameworks matter, SOTA pretrained models, ensembles, entities, faces and external knowledge. STILL FAR FROM SOLVED.
Adversarial VQA

HM is not perfect and everybody loves VQA, can we improve VQA itself?

First multimodal approach to human-and-model-in-the-loop, dynamic adversarial data collection:
Adversarial VQA

Is VQA as a task really arguably/almost saturated?

No. Not even close (with simple questions):

<table>
<thead>
<tr>
<th>Model</th>
<th>VQA test-dev</th>
<th>AdVQA test</th>
<th>VQA val</th>
<th>AdVQA val</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human performance</td>
<td>80.78</td>
<td>91.18</td>
<td>84.73</td>
<td>87.53</td>
</tr>
<tr>
<td>Majority answer (overall)</td>
<td></td>
<td>13.38</td>
<td>24.67</td>
<td>11.65</td>
</tr>
<tr>
<td>Majority answer (per answer type)</td>
<td></td>
<td>27.39</td>
<td>31.01</td>
<td>29.24</td>
</tr>
<tr>
<td>Model in loop</td>
<td>73.56</td>
<td>10.33</td>
<td>73.51</td>
<td>10.24</td>
</tr>
<tr>
<td>Multimodal (unimodal pretrain)</td>
<td>MoViE+MCAN [42]</td>
<td>71.36</td>
<td>26.64</td>
<td>71.31</td>
</tr>
<tr>
<td></td>
<td>MMBT [28]</td>
<td>58.00</td>
<td>26.70</td>
<td>57.32</td>
</tr>
<tr>
<td></td>
<td>UniT [22]</td>
<td>64.36</td>
<td>28.15</td>
<td>64.32</td>
</tr>
<tr>
<td>Multimodal (multimodal pretrain)</td>
<td>VisualBERT [33]</td>
<td>70.37</td>
<td>28.70</td>
<td>70.05</td>
</tr>
<tr>
<td></td>
<td>ViLBERT [39]</td>
<td>69.42</td>
<td>27.36</td>
<td>69.27</td>
</tr>
<tr>
<td></td>
<td>ViLT [30]</td>
<td>64.52</td>
<td>27.11</td>
<td>65.43</td>
</tr>
<tr>
<td></td>
<td>UNITER Base [10]</td>
<td>71.87</td>
<td>25.16</td>
<td>70.50</td>
</tr>
<tr>
<td></td>
<td>UNITER Large [10]</td>
<td>73.57</td>
<td>26.94</td>
<td>72.71</td>
</tr>
<tr>
<td></td>
<td>VILLA Base [16]</td>
<td>70.94</td>
<td>25.14</td>
<td>69.50</td>
</tr>
<tr>
<td></td>
<td>VILLA Large [16]</td>
<td>72.29</td>
<td>25.79</td>
<td>71.40</td>
</tr>
<tr>
<td>Multimodal (unimodal pretrain + OCR)</td>
<td>M4C (TextVQA+STVQA) [23]</td>
<td>32.89</td>
<td>28.86</td>
<td>31.44</td>
</tr>
<tr>
<td></td>
<td>M4C (VQA v2 train set) [23]</td>
<td>67.66</td>
<td>33.52</td>
<td>66.21</td>
</tr>
</tbody>
</table>

Q: How many cats are in the image?  
A: 2  
Model: 2, 2, 2

Q: What brand is the tv?  
A: LG  
Model: sony, samsung, samsung

Q: Does the cat look happy?  
A: no  
Model: no, no, no

Q: How many cartoon drawings are present on the cat’s tie?  
A: 4  
Model: 1, 1, 2

Q: What kind of floor is the man sitting on?  
A: wood  
Model: wood, wood, wood

Q: Did someone else take this picture?  
A: no  
Model: yes, yes, yes
Adversarial VQA

Table 4: The category-wise performance of VQA models. The state-of-the-art VQA models perform very close to the majority class prior, illustrating the challenge and difficulty of AdVQA.

<table>
<thead>
<tr>
<th>Model</th>
<th>Question Type</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>yes/no</td>
<td>numbers</td>
<td>others</td>
</tr>
<tr>
<td>Majority Class</td>
<td>62.28</td>
<td>31.11</td>
<td>9.29</td>
</tr>
<tr>
<td>ResNet-152</td>
<td>62.81</td>
<td>0.18</td>
<td>0.51</td>
</tr>
<tr>
<td>BERT</td>
<td>67.58</td>
<td>26.87</td>
<td>9.25</td>
</tr>
<tr>
<td>VisualBERT</td>
<td>55.51</td>
<td>32.29</td>
<td>17.66</td>
</tr>
<tr>
<td>ViLBERT</td>
<td>55.58</td>
<td>29.49</td>
<td>16.67</td>
</tr>
<tr>
<td>MoViE+MCAN*</td>
<td>52.74</td>
<td>33.62</td>
<td>14.56</td>
</tr>
<tr>
<td>M4C (VQA2)</td>
<td>56.67</td>
<td>38.04</td>
<td>22.73</td>
</tr>
</tbody>
</table>

Table 5: The category-wise distribution of answers. Compared to VQA, AdVQA contains more “number” based and lesser “yes/no” questions supporting the prior work’s observations around failure of VQA models to count and read text.

<table>
<thead>
<tr>
<th>Question Type</th>
<th>VQA test-dev</th>
<th>AdVQA test</th>
<th>VQA test</th>
<th>AdVQA val</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes/no</td>
<td>38.36</td>
<td>17.89</td>
<td>37.70</td>
<td>17.90</td>
</tr>
<tr>
<td>number</td>
<td>12.31</td>
<td>41.91</td>
<td>11.48</td>
<td>31.80</td>
</tr>
<tr>
<td>others</td>
<td>49.33</td>
<td>40.20</td>
<td>50.82</td>
<td>50.30</td>
</tr>
</tbody>
</table>
AdVQA & AVQA

More information: adversarialvqa.org

Adversarial VQA

What is Adversarial VQA?

Adversarial VQA is a new VQA benchmark that is collected with Human-And-Model-in-the-Loop for evaluating the robustness of state-of-the-art VQA systems.

- 2 datasets: AdVQA (in-domain) and AVQA (out-of-domain)
- Collected in single round or multiple rounds
- 81,253 images (COCO/Conceptual Captions 3M/Fakeddit/VCR)
- 1.9 human-verified adversarial questions on average per image
- 10 ground truth human-written answers per verified question

Dataset

Details on downloading the latest dataset may be found on the download webpage.

August 2021: Full release (v1.0)

AdVQA (In-domain)
- Collected in single round
- 41,807 COCO images (only for val/test)
- 46,807 questions
- 468,070 human-written answers

AVQA (Out-of-domain)
- Collected with 3 rounds
- 40,637 images from Conceptual Captions/Fakeddit/VCR (for train/val/test)
- 104,410 verified questions, 73,075 unverified questions
- 1,044,100 human-written answers for verified questions, 73,075 VQA model answers for unverified questions

STILL FAR FROM SOLVED.
Winoground

CLIP (re)triggered interest in multimodality

Learning Transferable Visual Models From Natural Language Supervision

Alec Radford^1 Jong Wook Kim^1 Chris Hallacy^1 Aditya Ramesh^1 Gabriel Goh^1 Sandhini Agarwal^1 Girish Sastry^1 Amanda Askell^1 Pamela Mishkin^1 Jack Clark^1 Gretchen Krueger^1 Ilya Sutskever^1
Winoground

But how good is CLIP really?

Some relevant ideas/findings from NLP:

- Winograd schemas
  “The [trophy] doesn't fit in the [suitcase] because it is too [large/small]”
- Word order may not matter all that much

Masked Language Modeling and the Distributional Hypothesis:
Order Word Matters Pre-training for Little

Koustuv Sinha†† Robin Jia† Dieuwke Hupkes† Joelle Pineau††
Adina Williams† Douwe Kiela†
Winoground

- Examples written by linguist experts
- Using Getty Images API
- Simple way to measure by comparing scores
- In some cases, very difficult and requiring world knowledge
Winoground Findings

- SOTA models often perform below chance (again).
- VinVL/UNITER/ViLLA perform best, probably because they’re trained with image-text matching (ITM) loss.
- Paper has a breakdown by category, and shows that these models probably fall back to a weak unimodal prior.

<table>
<thead>
<tr>
<th>Model</th>
<th>Text</th>
<th>Image</th>
<th>Group</th>
</tr>
</thead>
<tbody>
<tr>
<td>MTurk Human</td>
<td>89.50</td>
<td>88.50</td>
<td>85.50</td>
</tr>
<tr>
<td>Random Chance</td>
<td>25.00</td>
<td>25.00</td>
<td>16.67</td>
</tr>
<tr>
<td>VinVL</td>
<td>37.75</td>
<td>17.75</td>
<td>14.50</td>
</tr>
<tr>
<td>UNITER_{large}</td>
<td>38.00</td>
<td>14.00</td>
<td>10.50</td>
</tr>
<tr>
<td>UNITER_{base}</td>
<td>32.25</td>
<td>13.25</td>
<td>10.00</td>
</tr>
<tr>
<td>ViLLA_{large}</td>
<td>37.00</td>
<td>13.25</td>
<td>11.00</td>
</tr>
<tr>
<td>ViLLA_{base}</td>
<td>30.00</td>
<td>12.00</td>
<td>8.00</td>
</tr>
<tr>
<td>VisualBERT_{base}</td>
<td>15.50</td>
<td>2.50</td>
<td>1.50</td>
</tr>
<tr>
<td>ViLT (ViT-B/32)</td>
<td>34.75</td>
<td>14.00</td>
<td>9.25</td>
</tr>
<tr>
<td>LXMERT</td>
<td>19.25</td>
<td>7.00</td>
<td>4.00</td>
</tr>
<tr>
<td>ViLBERT_{base}</td>
<td>23.75</td>
<td>7.25</td>
<td>4.75</td>
</tr>
<tr>
<td>UniIT_{frozen}</td>
<td>19.50</td>
<td>6.25</td>
<td>4.00</td>
</tr>
<tr>
<td>CLIP (ViT-B/32)</td>
<td>30.75</td>
<td>10.50</td>
<td>8.00</td>
</tr>
<tr>
<td>VSE++COCO (ResNet)</td>
<td>22.75</td>
<td>8.00</td>
<td>4.00</td>
</tr>
<tr>
<td>VSE++COCO (VGG)</td>
<td>18.75</td>
<td>5.50</td>
<td>3.50</td>
</tr>
<tr>
<td>VSE++Flickr30k (ResNet)</td>
<td>20.00</td>
<td>5.00</td>
<td>2.75</td>
</tr>
<tr>
<td>VSE++Flickr30k (VGG)</td>
<td>19.75</td>
<td>6.25</td>
<td>4.50</td>
</tr>
<tr>
<td>VSRN_COCO</td>
<td>17.50</td>
<td>7.00</td>
<td>3.75</td>
</tr>
<tr>
<td>VSRN_{Flickr30k}</td>
<td>20.00</td>
<td>5.00</td>
<td>3.50</td>
</tr>
</tbody>
</table>

STILL FAR FROM SOLVED.
1. Dynabench
   a. Overview
   b. Common Objections & Misconceptions

2. Progress in Dynamic Adversarial Data Collection
   a. Humans and Models in Loops
   b. Dynamic Adversarial Training Data

3. Adventures in Multimodal ML
   a. Evaluation: Hateful Memes, Adversarial VQA, Winoground
   b. Foundation Models: FLAVA
Building pretrained multimodal models - why?

- Many tasks are multimodal: the internet and the world are multimodal
- Modalities can complement each other and share knowledge and resources
- Sharing parameters and improved sample efficiency
- Architectures are overly domain specific (slowly changing with Transformers taking over everything) so we may require N models for N tasks
- Modality-agnostic large language models => foundation models.

On the Opportunities and Risks of Foundation Models

Challenges

● Paired cross-modal data is not abundantly available
● Data from prior work has not been made public
● Joint learning across modalities is hard
● Pretraining techniques are domain specific
● Unclear how to leverage unimodal data
● Compute
FLAVA

- Holistic approach to multimodality
- One foundation model spanning V&L, CV and NLP
- Impressive performance on 35 tasks across NLP, CV and V&L domains.
- Jointly pretrained on:
  - unimodal text data (CCNews + BookCorpus)
  - unimodal image data (ImageNet)
  - public paired image-text data (70M)
- All data/models are publicly released
The problem to solve

multimodal and unimodal pretraining data

image-text pairs

unpaired images

unpaired text

**FLAVA** for multi-domain joint pretraining (global contrastive, MMM, MIM, MLM, ...)

visual recognition (e.g. ImageNet)

language understanding (e.g. GLUE)

multimodal reasoning (e.g. VQA)
How does FLAVA work?
How does FLAVA work?

This cat was wonderful! He was making his daily cleaning on an ancient grave as to say “I am the boss here!”
How does FLAVA work?

This cat was wonderful! He was making his daily cleaning on an ancient grave as to say "I am the boss here!"
The PMD dataset

- 70M image-text pairs from public sources
How well does it work?

• On average, over 35 tasks, FLAVA obtains impressive performance
### How well does it work?

<table>
<thead>
<tr>
<th>Experimental setting</th>
<th>vision-only tasks</th>
<th>vision-and-language tasks</th>
<th>language-only tasks (GLUE benchmark)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ImageNet accuracy</td>
<td>VQAv2 accuracy</td>
<td>SNLI-VE accuracy</td>
</tr>
<tr>
<td>FLAVA</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>one pretrained model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>shared between tasks</td>
<td>75.5</td>
<td>72.8</td>
<td>79.0</td>
</tr>
<tr>
<td>UniT</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>one model shared between tasks</td>
<td>-</td>
<td>67.0</td>
<td>73.1</td>
</tr>
<tr>
<td>VisualBERT (Li et. al.)</td>
<td>separately fine-tuned on each task</td>
<td>-</td>
<td>70.8</td>
</tr>
<tr>
<td>CLIP (Radford et. al.)</td>
<td>80.2</td>
<td>55.3</td>
<td>73.5</td>
</tr>
<tr>
<td>BERT (Devlin et. al.)</td>
<td>separately fine-tuned on each task</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
What’s next?

● Always work in progress
● Challenges we addressed:
  ○ Data => PMD
  ○ Architecture => Transformers
  ○ Joint training => FLAVA
  ○ Requires heavy compute
● Things to explore:
  ○ Fully sharing (almost) all parameters
  ○ Training bigger models on more data
  ○ Training on all the modalities
How about closing the loop?

- We’re still working on the FLAVA open source release.
  Preliminary results on Winoground (WG) and AdVQA:

<table>
<thead>
<tr>
<th></th>
<th>WG-Text</th>
<th>WG-Image</th>
<th>WG-Group</th>
<th>AdVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best</td>
<td>37.75</td>
<td>17.75</td>
<td>14.50</td>
<td>33.67</td>
</tr>
<tr>
<td>FLAVA</td>
<td>32.25</td>
<td>19.75</td>
<td>14.00</td>
<td>36.02</td>
</tr>
</tbody>
</table>

- I don’t want to underhype but.. There is more work to be done!
Thanks multimodal collaborators
Thanks Dynabench collaborators

Who is on the team?

Everyone! People who have contributed to Dynabench so far include: Douwe Kiela, Divyansh Kaushik, Max Bartolo, Adina Williams, Yixin Nie, Grusha Prasad, Pratik Ringshia, Amanpreet Singh, Robin Jia, Sebastian Riedel, Tristan Thrush, Atticus Geiger, Chris Potts, Pontus Stenetorp, Mohit Bansal, Bertie Vidgen, Zeerak Talat, Zhiyi Ma, Ledell Wu, Sonia Kris, Zen Wu, Kawin Ethayarajh, Alberto Lopez, Sasha Sheng, Eric Wallace, Pedro Rodriguez, Rebecca Qian, Somya Jain, Guillaume Wenzek, Sahir Gomez, Anmol Gupta, Hannah Rose Kirk, Zoe Papakipos, Kok Rui Wong, Ishita Dasgupta, Anand Rajaram, Fatima Zahra Chriha, and others.
Thanks for listening

Questions?