SQuAD

The Stanford Question Answering Dataset

What is SQuAD?

Stanford **Qu**estion **A**nswering **D**ataset (SQuAD) is a new reading comprehension dataset, consisting of questions posed by crowdworkers on a set of Wikipedia articles, where the answer to every question is a segment of text, or *span*, from the corresponding reading passage. With 100,000+ question-answer pairs on 500+ articles, SQuAD is significantly larger than previous reading comprehension datasets.

Explore SQuAD and model predictions (/SQuAD-explorer/explore/1.1/dev/)

Read the paper (Rajpurkar et al. '16) (http://arxiv.org/abs/1606.05250)

Getting Started

We've built a few resources to help you get started with the dataset.
Download a copy of the dataset (distributed under the CC BY-SA 4.0 (http://creativecommons.org/licenses/by-sa/4.0/legalcode) license):

Training Set v1.1 (30 MB) (/SQuAD-explorer/dataset/train-v1.1.json)

Dev Set v1.1 (5 MB) (/SQuAD-explorer/dataset/dev-v1.1.json)

To evaluate your models, we have also made available the evaluation script we will use for official evaluation, along with a sample prediction file that the script will take as input. To run the evaluation, use `python evaluate-v1.1.py <path_to_dev-v1.1> <path_to_predictions>`.

Evaluation Script v1.1 (https://worksheets.codalab.org/rest/bundles/0xbcd57bee090b421c982906709c8c27e1/contents/blob/)

Sample Prediction File (on Dev v1.1) (https://worksheets.codalab.org/rest/bundles/0xc83bf36cf8714819ba11802b59cb809e/contents/blob/)

Once you have a built a model that works to your expectations on the dev set, you submit it to get official scores on the dev and a hidden test set. To preserve the integrity of test results, we do not release the test set to the public. Instead, we require you to submit your model so that we can run it on the test set for you. Here's a tutorial walking you through official evaluation of your model:

Submission Tutorial (https://worksheets.codalab.org/worksheets/0x8403d867f9a3444685c344f4f0bc8d34/)

Because SQuAD is an ongoing effort, we expect the dataset to evolve.
To keep up to date with major changes to the dataset, please subscribe:

email address

Subscribe

Have Questions?

Ask us questions at our google group (https://groups.google.com/forum/#!forum/squad-stanford-qa) or at pranavsr@stanford.edu (mailto:pranavsr@stanford.edu).

Tweet ★ Star 54

Leaderboard

Since the release of our dataset, the community has made rapid progress! Here are the ExactMatch (EM) and F1 scores of the best models evaluated on the test and development sets of v1.1. Will your model outperform humans on the QA task?

Rank	Model	EM	F1
1 Mar 2017	r-net (ensemble) *Microsoft Research Asia*	76.922	84.006

2 Mar 2017	ReasoNet (ensemble) *MSR Redmond*	75.034	82.552
3 Apr 2017	Mnemonic Reader (ensemble) *NUDT & Fudan University*	73.67	81.694
3 Apr 2017	SEDT+BiDAF (ensemble) *CMU* https://arxiv.org/abs/1703.00572 (https://arxiv.org/abs/1703.00572)	73.723	81.53
3 Feb 2017	BiDAF (ensemble) *Allen Institute for AI & University of Washington* https://arxiv.org/abs/1611.01603 (https://arxiv.org/abs/1611.01603)	73.744	81.525
4 May 2017	jNet (ensemble) *USTC & National Research Council Canada & York University* https://arxiv.org/abs/1703.04617 (https://arxiv.org/abs/1703.04617)	73.01	81.517
4 Jan 2017	Multi-Perspective Matching (ensemble) *IBM Research* https://arxiv.org/abs/1612.04211 (https://arxiv.org/abs/1612.04211)	73.765	81.257
5 Apr 2017	T-gating (ensemble) *Peking University*	72.758	81.001
6 Apr 2017	r-net (single model) *Microsoft Research Asia*	72.338	80.717
7 Nov 2016	Dynamic Coattention Networks (ensemble) *Salesforce Research* https://arxiv.org/abs/1611.01604 (https://arxiv.org/abs/1611.01604)	71.625	80.383
7 Apr 2017	QFASE *NUS*	71.898	79.989
8 Apr 2017	Interactive AoA Reader (single model) *Joint Laboratory of HIT and iFLYTEK Research*	71.153	79.937
9 Mar 2017	jNet (single model) *USTC & National Research Council Canada & York University* https://arxiv.org/abs/1703.04617 (https://arxiv.org/abs/1703.04617)	70.607	79.821
9 Apr 2017	Ruminating Reader (single model) *New York University* https://arxiv.org/abs/1704.07415 (https://arxiv.org/abs/1704.07415)	70.639	79.456
10 Mar 2017	ReasoNet (single model) *MSR Redmond*	70.555	79.364
10 Mar 2017	Document Reader (single model) *Facebook AI Research*	70.733	79.353
11 Apr 2017	Mnemonic Reader (single model) *NUDT & Fudan University*	69.863	79.207
11 Dec 2016	FastQAExt *German Research Center for Artificial Intelligence* https://arxiv.org/abs/1703.04816 (https://arxiv.org/abs/1703.04816)	70.849	78.857
12 Apr 2017	Multi-Perspective Matching (single model) *IBM Research* https://arxiv.org/abs/1612.04211 (https://arxiv.org/abs/1612.04211)	70.387	78.784
13 Apr 2017	SEDT+BiDAF (single model) *CMU* https://arxiv.org/abs/1703.00572 (https://arxiv.org/abs/1703.00572)	68.478	77.971
13 Mar 2017	RaSoR (single model) *Google NY, Tel-Aviv University* https://arxiv.org/abs/1611.01436 (https://arxiv.org/abs/1611.01436)	69.642	77.696
14 Apr 2017	T-gating (single model) *Peking University*	68.132	77.569
15 Nov 2016	BiDAF (single model) *Allen Institute for AI & University of Washington* https://arxiv.org/abs/1611.01603 (https://arxiv.org/abs/1611.01603)	67.974	77.323

15 Dec 2016	FastQA *German Research Center for Artificial Intelligence* https://arxiv.org/abs/1703.04816 (https://arxiv.org/abs/1703.04816)	68.436	77.07
16 Oct 2016	Match-LSTM with Ans-Ptr (Boundary) (ensemble) *Singapore Management University* https://arxiv.org/abs/1608.07905 (https://arxiv.org/abs/1608.07905)	67.901	77.022
17 Feb 2017	Iterative Co-attention Network *Fudan University*	67.502	76.786
18 Nov 2016	Dynamic Coattention Networks (single model) *Salesforce Research* https://arxiv.org/abs/1611.01604 (https://arxiv.org/abs/1611.01604)	66.233	75.896
19 Oct 2016	Match-LSTM with Bi-Ans-Ptr (Boundary) *Singapore Management University* https://arxiv.org/abs/1608.07905 (https://arxiv.org/abs/1608.07905)	64.744	73.743
20 Feb 2017	Attentive CNN context with LSTM *NLPR, CASIA*	63.306	73.463
21 Nov 2016	Fine-Grained Gating *Carnegie Mellon University* https://arxiv.org/abs/1611.01724 (https://arxiv.org/abs/1611.01724)	62.446	73.327
21 Sep 2016	Dynamic Chunk Reader *IBM* https://arxiv.org/abs/1610.09996 (https://arxiv.org/abs/1610.09996)	62.499	70.956
22 Aug 2016	Match-LSTM with Ans-Ptr (Boundary) *Singapore Management University* https://arxiv.org/abs/1608.07905 (https://arxiv.org/abs/1608.07905)	60.474	70.695
23 Aug 2016	Match-LSTM with Ans-Ptr (Sentence) *Singapore Management University* https://arxiv.org/abs/1608.07905 (https://arxiv.org/abs/1608.07905)	54.505	67.748
	Human Performance *Stanford University* (Rajpurkar et al. '16) (http://arxiv.org/abs/1606.05250)	82.304	91.221