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.

(/SQuAD-explorer/explore/1.1/dev/)

(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/0xbcd57bee090b421c982

Sample Prediction File (on Dev v1.1) (https://worksheets.codalab.org/rest/bundles/0xc83bf36cf8714819ba1

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/0x8403d867f9a3444685c

Because SQuAD is an ongoing effort, we expect the dataset to evolve.

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 set of v1.1. Will your model outperform humans on the QA task?

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16) (http://arxiv.org/abs/1606.05250)	82.304	91.221
1 Jan 22, 2018	Hybrid AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.482	**89.281**
2 Jan 05, 2018	SLQA+ (ensemble) *Alibaba iDST NLP*	82.440	88.607
2 Jan 03, 2018	r-net+ (ensemble) *Microsoft Research Asia*	**82.650**	88.493
3 Dec 17, 2017	r-net (ensemble) *Microsoft Research Asia* http://aka.ms/rnet (http://aka.ms/rnet)	82.136	88.126
3 Dec 22, 2017	AttentionReader+ (ensemble) *Tencent DPDAC NLP*	81.790	88.163
4 Nov 17, 2017	BiDAF + Self Attention + ELMo (ensemble) *Allen Institute for Artificial Intelligence*	81.003	87.432
5 Jan 12, 2018	EAZI+ (ensemble) *Yiwise NLP Group*	80.426	86.912
5 Jan 22, 2018	Hybrid AoA Reader (single model) *Joint Laboratory of HIT and iFLYTEK Research*	80.027	87.288
5 Jan 13, 2018	SLQA+ *single model*	80.436	87.021
5 Jan 04, 2018	{EAZI} (ensemble) *Yiwise NLP Group*	80.436	86.912
6 Dec 28, 2017	SLQA+ (single model) *Alibaba iDST NLP*	79.199	86.590
6 Jan 03, 2018	r-net+ (single model) *Microsoft Research Asia*	79.901	86.536
7 Dec 05, 2017	SAN (ensemble model) *Microsoft Business AI Solutions Team* https://arxiv.org/pdf/1712.03556.pdf (https://arxiv.org/pdf/1712.03556.pdf)	79.608	86.496
8 Oct 17, 2017	Interactive AoA Reader+ (ensemble) *Joint Laboratory of HIT and iFLYTEK*	79.083	86.450
9 Oct 24, 2017	FusionNet (ensemble) *Microsoft Business AI Solutions Team*	78.978	86.016

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 (https://twitter.com/share) ★ Star 148

	https://arxiv.org/abs/1711.07341 (https://arxiv.org/abs/1711.07341)		
10 Oct 22, 2017	DCN+ (ensemble) *Salesforce Research*	78.852	85.996
11 Nov 03, 2017	BiDAF + Self Attention + ELMo (single model) *Allen Institute for Artificial Intelligence*	78.580	85.833
12 Jan 02, 2018	Conductor-net (ensemble) *CMU* https://arxiv.org/abs/1710.10504 (https://arxiv.org/abs/1710.10504)	78.433	85.517
12 Nov 30, 2017	SLQA(ensemble) *Alibaba iDST NLP*	78.328	85.682
13 Jan 03, 2018	MEMEN (single model) *Zhejiang University* https://arxiv.org/abs/1707.09098 (https://arxiv.org/abs/1707.09098)	78.234	85.344
14 Jul 25, 2017	Interactive AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	77.845	85.297
15 Jan 10, 2018	MAMCN+ (single model) *Samsung Research*	77.436	85.130
16 Dec 06, 2017	AttentionReader+ (single) *Tencent DPDAC NLP*	77.342	84.925
16 Aug 21, 2017	Reinforced Mnemonic Reader (ensemble) *NUDT and Fudan University* https://arxiv.org/abs/1705.02798 (https://arxiv.org/abs/1705.02798)	77.678	84.888
17 Nov 06, 2017	Conductor-net (ensemble) *CMU* https://arxiv.org/abs/1710.10504 (https://arxiv.org/abs/1710.10504)	76.996	84.630
17 Dec 21, 2017	Jenga (ensemble) *Facebook AI Research*	77.237	84.466
17 Dec 13, 2017	RaSoR + TR + LM (single model) *Tel-Aviv University* https://arxiv.org/abs/1712.03609 (https://arxiv.org/abs/1712.03609)	77.583	84.163
17 Oct 13, 2017	r-net (single model) *Microsoft Research Asia* http://aka.ms/rnet (http://aka.ms/rnet)	76.461	84.265
17 Dec 19, 2017	FRC (single model) *in review*	76.240	84.599
17 Nov 01, 2017	SAN (single model) *Microsoft Business AI Solutions Team* https://arxiv.org/pdf/1712.03556.pdf (https://arxiv.org/pdf/1712.03556.pdf)	76.828	84.396
18 Oct 22, 2017	Conductor-net (ensemble) *CMU*	76.146	83.991
19 Sep 08, 2017	FusionNet (single model) *Microsoft Business AI Solutions team* https://arxiv.org/abs/1711.07341 (https://arxiv.org/abs/1711.07341)	75.968	83.900
19 Jul 14, 2017	smarnet (ensemble) *Eigen Technology & Zhejiang University*	75.989	83.475
19 Oct 22, 2017	Interactive AoA Reader+ (single model) *Joint Laboratory of HIT and iFLYTEK*	75.821	83.843
20 Aug 18, 2017	RaSoR + TR (single model) *Tel-Aviv University* https://arxiv.org/abs/1712.03609	75.789	83.261

	(https://arxiv.org/abs/1712.03609)		
21 Oct 23, 2017	DCN+ (single model) *Salesforce Research*	75.087	83.081
22 Jul 10, 2017	DCN+ (single model) *Salesforce Research*	74.866	82.806
22 Nov 01, 2017	Mixed model (ensemble) *Sean*	75.265	82.769
22 Oct 31, 2017	SLQA (single model) *Alibaba iDST NLP*	74.489	82.815
22 Nov 17, 2017	two-attention-self-attention (ensemble) *guotong1988*	75.223	82.716
22 May 21, 2017	MEMEN (ensemble) *Eigen Technology & Zhejiang University* https://arxiv.org/abs/1707.09098 (https://arxiv.org/abs/1707.09098)	75.370	82.658
23 Mar 09, 2017	ReasoNet (ensemble) *MSR Redmond* https://arxiv.org/abs/1609.05284 (https://arxiv.org/abs/1609.05284)	75.034	82.552
24 Jul 14, 2017	Mnemonic Reader (ensemble) *NUDT and Fudan University* https://arxiv.org/abs/1705.02798 (https://arxiv.org/abs/1705.02798)	74.268	82.371
25 Dec 23, 2017	S^3-Net (ensemble) *Kangwon National University in South Korea*	74.121	82.342
25 Oct 27, 2017	Unnamed submission by null	74.489	82.312
25 Jan 02, 2018	Conductor-net (single model) *CMU* https://arxiv.org/abs/1710.10504 (https://arxiv.org/abs/1710.10504)	74.405	82.742
26 Nov 06, 2017	Conductor-net (single) *CMU* https://arxiv.org/abs/1710.10504 (https://arxiv.org/abs/1710.10504)	73.240	81.933
26 Jul 25, 2017	Interactive AoA Reader (single model) *Joint Laboratory of HIT and iFLYTEK Research*	73.639	81.931
27 Aug 21, 2017	Reinforced Mnemonic Reader (single model) *NUDT and Fudan University* https://arxiv.org/abs/1705.02798 (https://arxiv.org/abs/1705.02798)	73.188	81.816
27 Jul 29, 2017	SEDT (ensemble model) *CMU* https://arxiv.org/abs/1703.00572 (https://arxiv.org/abs/1703.00572)	74.090	81.761
28 Dec 14, 2017	Jenga (single model) *Facebook AI Research*	73.303	81.754
28 Jul 06, 2017	SSAE (ensemble) *Tsinghua University*	74.080	81.665
29 Feb 22, 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
29 Apr 22, 2017	SEDT+BiDAF (ensemble) *CMU* https://arxiv.org/abs/1703.00572 (https://arxiv.org/abs/1703.00572)	73.723	81.530
29	Multi-Perspective Matching (ensemble)	73.765	81.257

Jan 24, 2017	*IBM Research* https://arxiv.org/abs/1612.04211 (https://arxiv.org/abs/1612.04211)		
29 May 01, 2017	jNet (ensemble) *USTC & National Research Council Canada & York University* https://arxiv.org/abs/1703.04617 (https://arxiv.org/abs/1703.04617)	73.010	81.517
30 Oct 22, 2017	Conductor-net (single) *CMU*	72.590	81.415
31 Dec 15, 2017	S^3-Net (single model) *Kangwon National University in South Korea*	71.908	81.023
31 Apr 12, 2017	T-gating (ensemble) *Peking University*	72.758	81.001
31 Sep 20, 2017	BiDAF + Self Attention (single model) *Allen Institute for Artificial Intelligence* https://arxiv.org/abs/1710.10723 (https://arxiv.org/abs/1710.10723)	72.139	81.048
31 Nov 16, 2017	two-attention-self-attention (single model) *guotong1988*	72.600	81.011
32 Nov 06, 2017	attention+self-attention (single model) *guotong1988*	71.698	80.462
33 Nov 01, 2016	Dynamic Coattention Networks (ensemble) *Salesforce Research* https://arxiv.org/abs/1611.01604 (https://arxiv.org/abs/1611.01604)	71.625	80.383
34 Jul 14, 2017	smarnet (single model) *Eigen Technology & Zhejiang University* https://arxiv.org/abs/1710.02772 (https://arxiv.org/abs/1710.02772)	71.415	80.160
34 Apr 13, 2017	QFASE *NUS*	71.898	79.989
35 Oct 27, 2017	M-NET (single) *UFL*	71.016	79.835
35 Jul 14, 2017	Mnemonic Reader (single model) *NUDT and Fudan University* https://arxiv.org/abs/1705.02798 (https://arxiv.org/abs/1705.02798)	70.995	80.146
36 Mar 24, 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
36 Apr 02, 2017	Ruminating Reader (single model) *New York University* https://arxiv.org/abs/1704.07415 (https://arxiv.org/abs/1704.07415)	70.639	79.456
37 Mar 08, 2017	ReasoNet (single model) *MSR Redmond* https://arxiv.org/abs/1609.05284 (https://arxiv.org/abs/1609.05284)	70.555	79.364
37 Mar 14, 2017	Document Reader (single model) *Facebook AI Research* https://arxiv.org/abs/1704.00051 (https://arxiv.org/abs/1704.00051)	70.733	79.353
37 Dec 28, 2016	FastQAExt *German Research Center for Artificial Intelligence* https://arxiv.org/abs/1703.04816 (https://arxiv.org/abs/1703.04816)	70.849	78.857

38 Apr 14, 2017	Multi-Perspective Matching (single model) *IBM Research* https://arxiv.org/abs/1612.04211 (https://arxiv.org/abs/1612.04211)	70.387	78.784
38 May 13, 2017	RaSoR (single model) *Google NY, Tel-Aviv University* https://arxiv.org/abs/1611.01436 (https://arxiv.org/abs/1611.01436)	70.849	78.741
39 Aug 30, 2017	SimpleBaseline (single model) *Technical University of Vienna*	69.600	78.236
40 Apr 12, 2017	SEDT+BiDAF (single model) *CMU* https://arxiv.org/abs/1703.00572 (https://arxiv.org/abs/1703.00572)	68.478	77.971
41 Jun 25, 2017	PQMN (single model) *KAIST & AIBrain & Crosscert*	68.331	77.783
42 Apr 12, 2017	T-gating (single model) *Peking University*	68.132	77.569
42 Jul 29, 2017	SEDT (single model) *CMU* https://arxiv.org/abs/1703.00572 (https://arxiv.org/abs/1703.00572)	68.163	77.527
43 Nov 28, 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
43 Jan 22, 2018	FABIR (Single Model) *in review*	67.744	77.605
44 Sep 19, 2017	AllenNLP BiDAF (single model) *Allen Institute for AI* http://allennlp.org/ (http://allennlp.org/)	67.618	77.151
44 Dec 28, 2016	FastQA *German Research Center for Artificial Intelligence* https://arxiv.org/abs/1703.04816 (https://arxiv.org/abs/1703.04816)	68.436	77.070
45 Oct 26, 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
46 Feb 05, 2017	Iterative Co-attention Network *Fudan University*	67.502	76.786
47 Nov 01, 2016	Dynamic Coattention Networks (single model) *Salesforce Research* https://arxiv.org/abs/1611.01604 (https://arxiv.org/abs/1611.01604)	66.233	75.896
47 Jan 03, 2018	newtest *single model*	66.527	75.787
48 Jan 03, 2018	baseline *single model*	64.796	74.272
49 Dec 09, 2017	Unnamed submission by ravioncodalab	64.439	73.921
49 Oct 26, 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
50 Feb 19, 2017	Attentive CNN context with LSTM *NLPR, CASIA*	63.306	73.463

50 Sep 21, 2017	OTF dict+spelling (single) *University of Montreal* https://arxiv.org/abs/1706.00286 (https://arxiv.org/abs/1706.00286)	64.083	73.056
51 Sep 21, 2017	OTF spelling (single) *University of Montreal* https://arxiv.org/abs/1706.00286 (https://arxiv.org/abs/1706.00286)	62.897	72.016
51 Nov 02, 2016	Fine-Grained Gating *Carnegie Mellon University* https://arxiv.org/abs/1611.01724 (https://arxiv.org/abs/1611.01724)	62.446	73.327
51 Sep 21, 2017	OTF spelling+lemma (single) *University of Montreal* https://arxiv.org/abs/1706.00286 (https://arxiv.org/abs/1706.00286)	62.604	71.968
52 Sep 28, 2016	Dynamic Chunk Reader *IBM* https://arxiv.org/abs/1610.09996 (https://arxiv.org/abs/1610.09996)	62.499	70.956
53 Aug 27, 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
54 Jan 05, 2018	PivRet (single model) *anonymous*	58.764	69.276
55 Aug 27, 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