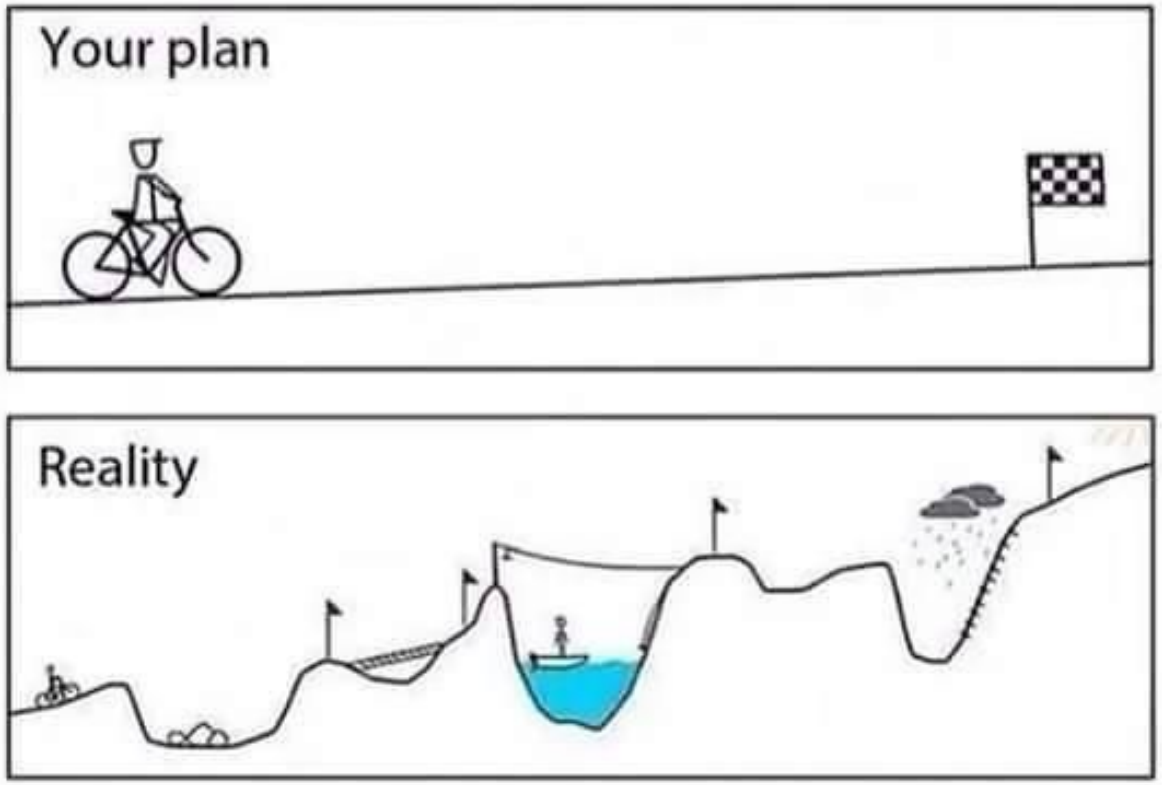
TRANSFER LEARNING FOR MACHINE READING COMPREHENSION

STUDYING KNOWLEDGE TRANSFER OVER BI-DIRECTIONAL ATTENTION FLOW NETWORKS FOR READING COMPREHENSION

AGENDA

- Background + Motivation
  - Machine reading comprehension
  - Squad dataset
  - Motivation examples
  - Transfer learning
- Methods
- Results



WHAT IS MACHINE COMPREHENSION?

Passage: Tesla later approached Morgan to ask for more funds to build a more powerful transmitter. **When asked where all the money had gone, Tesla responded by saying that he was affected by the Panic of 1901**, which he (Morgan) had caused. Morgan was shocked by the reminder of his part in the stock market crash and by Tesla's breach of contract by asking for more funds. Tesla wrote another plea to Morgan, but it was also fruitless. Morgan still owed Tesla money on the original agreement, and Tesla had been facing foreclosure even before construction of the tower began.

Question: On what did Tesla blame for the loss of the initial money?

Answer: Panic of 1901

GIVEN A TEXT PASSAGE **P**

AND A QUESTION **Q**,

PREDICT THE ANSWER **A**

USING THE INFORMATION IN **P**.

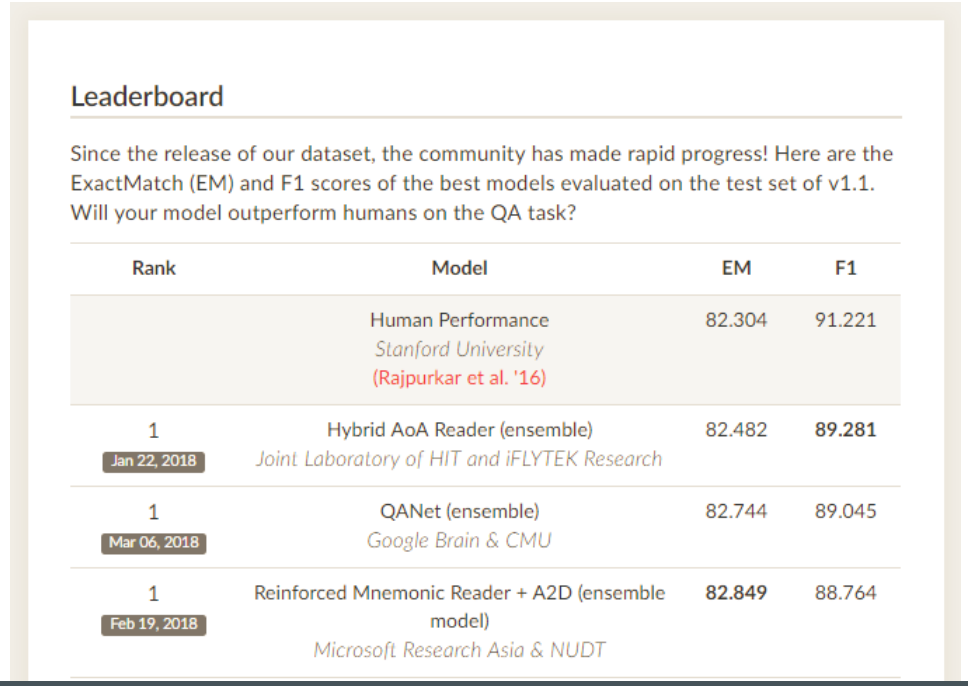
STANFORD SQUAD DATASET

100,000 + QUESTIONS

CROWDSOURCED

CORPUS: WIKIPEDIA

A IS A SUB-SPAN OF **P**



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)	82.304	91.221
1 Jan 22, 2018	Hybrid AoA Reader (ensemble) *Joint Laboratory of HIT and iFLYTEK Research*	82.482	**89.281**
1 Mar 06, 2018	QANet (ensemble) *Google Brain & CMU*	82.744	89.045
1 Feb 19, 2018	Reinforced Mnemonic Reader + A2D (ensemble model) *Microsoft Research Asia & NUDT*	**82.849**	88.764

MOTIVATION: EXAMPLES

- Question-Answering chatbot
- Car embedded bot
- Car manual + services
- *What should be the tire pressure?*

- Internet assistant
- Browser extension
- User asks questions while browsing
- Example: blog post

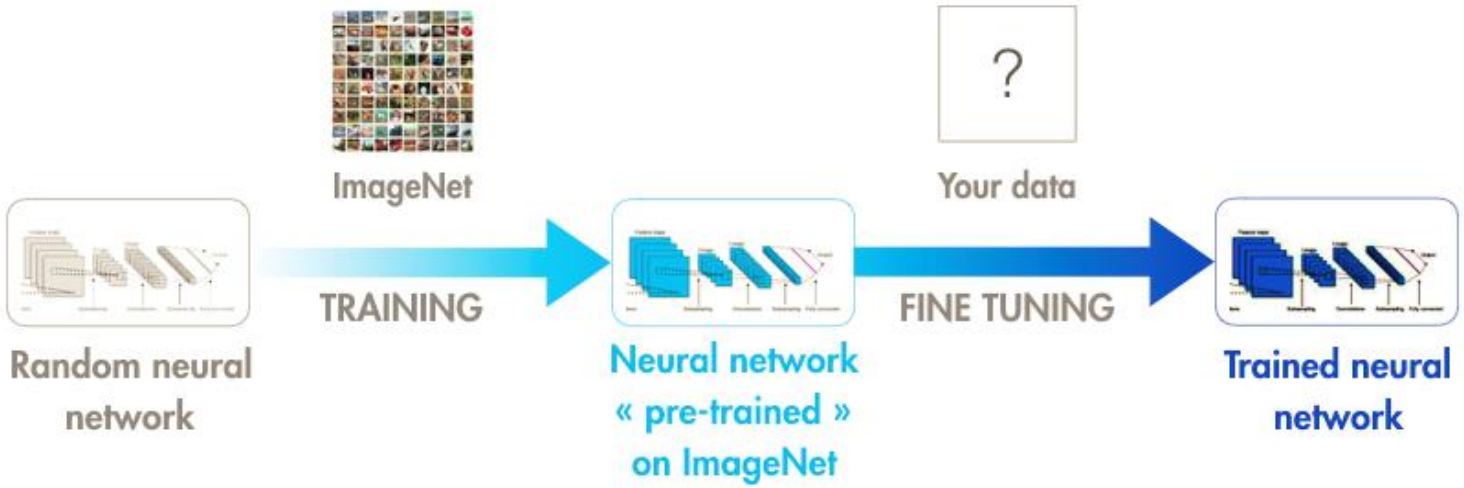
PROBLEM: OTHER DOMAINS DON'T HAVE LUXURY DATASETS LIKE SQUAD!!!

GOAL: TRANSFER KNOWLEDGE FROM SQUAD TO OTHER CORPORA

INSPIRATION: COMPUTER VISION

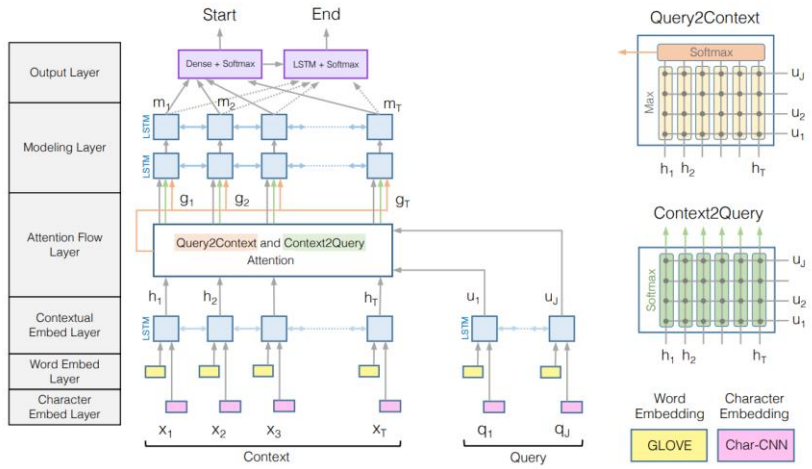
TRAIN WITH SQUAD DATASET

FINE TUNE IN OTHER DOMAIN



METHOD: STUDYING FEASIBILITY OF TRANSFER LEARNING FOR MACHINE COMPREHENSION

SOURCE DOMAIN	• SQUAD DATASET • WIKIPEDIA CONTENT
TARGET DOMAIN	• MS-MARCO DATASET • INTERNET CONTENT • WE USE BETWEEN 0 AND 500 LABELED EXAMPLES
ARCHITECTURE	• BI-DIRECTIONAL ATTENTION FLOW NETWORK (BI-DAF)
METHOD	• TRAIN BI-DAF ON SQUAD • FINE-TUNE FOR MS-MARCO EXAMPLES
HARDWARE	• NVIDIA K80 GPU • TRAIN ON SQUAD: 16.4 HS



RESULTS

MODEL	F1 (LR = 0.8)	F1 (LR = 0.5)
R-NET + SQUAD (STATE OF THE ART)	0.88764	N/A
BI-DAF + SQUAD (OUR BASE MODEL)	0.7568401 (0.773)	0.7407
BI-DAF + MS-MARCO (NO TRANSFER, 500 EXAMPLES)	0.1102	N/A
BI-DAF + SQUAD (DIRECT EVALUATION ON MS-MARCO)	0.5097	0.4833
BI-DAF + SQUAD + MS-MARCO (131 EXAMPLES)	0.54881	0.55023
BI-DAF + SQUAD + MS-MARCO (250 EXAMPLES)	0.55717	0.56829
BI-DAF + SQUAD + MS-MARCO (500 EXAMPLES)	0.56014	0.59145

CONCLUSIONS

ANALYSIS

- ACCURACY / F1: GOOD START BUT CAN BE IMPROVED!
- IS IT THE KIND OF QUESTION? NO! F1 SCORES ARE SIMILAR ACROSS QUESTION CATEGORIES
  - NUMERIC (28.4%)
  - DESCRIPTION (52.6%)
  - LOCATION (5.7%)
- LOW PERFORMANCE:
  - INFORMAL WRITING, SOCIAL NETWORKS, NEWS, SLANG!
  - WHY QUESTIONS
  - QUESTIONS WITH NO ANSWER

NEXT STEPS

- IN-PROGRESS: WEB NAVIGATION ASSISTANT CHAT BOT WITH MODEL
- STUDY NOT FREEZING EMBEDDING LAYER DURING FINE TUNING
- REPEAT FOR OTHER NON ATTENTION-BASED ARCHITECTURES SUCH AS R-NET
- REPEAT FOR ENSEMBLES USED IN STATE OF THE ART

QUESTIONS?