Evaluating NLU Models with Harder Generalization Tasks

Atticus Geiger

Overview

- Standard vs non-standard generalization tasks for NLU models
- Adversarial testing
- Artificial tasks

Standard Generalization Tasks

- Find a dataset for your NLU task
- Arbitrarily split your dataset into training, development, and testing sets
- Train a model on the training set and then evaluate performance on unseen testing examples
- This is the standard evaluation framework we have used in this class

Standard Generalization Tasks

- In our third homework, our NLU task was NLI on single words
- Our edge-disjoint task follows our standard evaluation framework of arbitrarily creating training and testing splits
- Our word-disjoint task breaks from this standard, presenting the new more difficult task of generalizing to unseen words

Non-Standard Generalization Tasks

- I want to encourage you to consider breaking from this standard evaluation framework
- We should try to create generalization tasks that are difficult, well motivated, and answer specific questions about model capabilities

Operationalizing an Ambitious Question

- “Can a model learn to comprehend a passage of text?”
- To answer this question, the Stanford Question Answer Dataset, an awesome resource for your projects, was crowd sourced (Rajpurkar et al. 2016)
- We might think that if a model achieves human level performance on the standard generalization task using this dataset, then the model can comprehend a passage of text

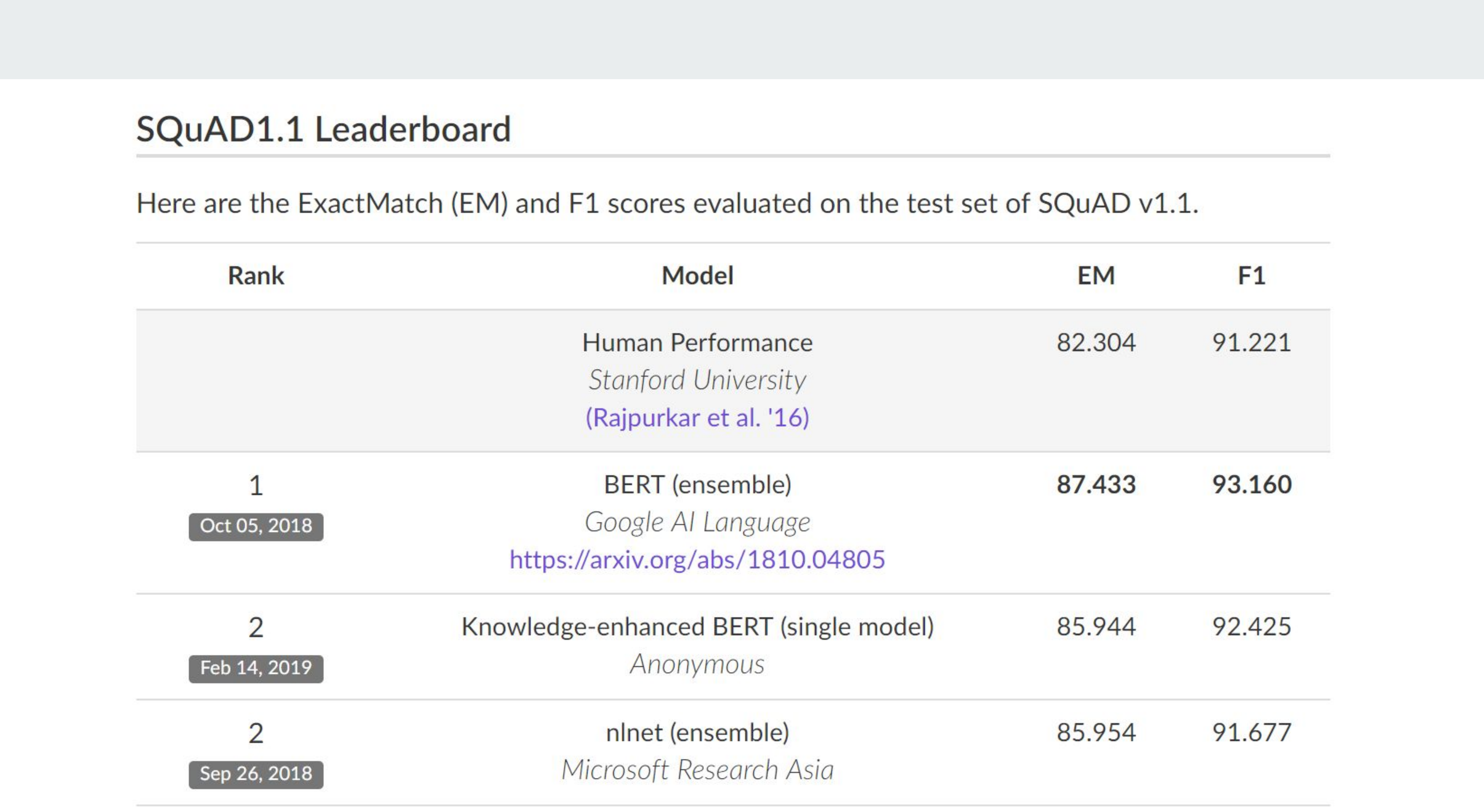
Stanford Question Answer Dataset (SQuAD)

Passage:

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by **John Elway**, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question: What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer: John Elway



SQuAD1.1 Leaderboard

Here are the ExactMatch (EM) and F1 scores evaluated on the test set of SQuAD v1.1.

Rank	Model	EM	F1
	Human Performance *Stanford University* (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) *Google AI Language* https://arxiv.org/abs/1810.04805	**87.433**	**93.160**
2 Feb 14, 2019	Knowledge-enhanced BERT (single model) *Anonymous*	85.944	92.425
2 Sep 26, 2018	nlnet (ensemble) *Microsoft Research Asia*	85.954	91.677

Question answering is solved!

- Triumphant day for AI
- Natural language understanding is essentially a done deal
- Pretty soon we will have conscious robots
- Time to go home

Adversarial Testing (Jia et al. 2017)

- Models trained on SQuAD might not understand language as deeply as we might have hoped
- Systematically perturb examples from training data to generate a test set by appending a misleading sentence
- Use this adversarial test set as your evaluation metric

Train Example

Passage:

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by **John Elway**, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question: What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer: John Elway

Model Prediction: John Elway

Adversarial Test Example

Passage:

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by **John Elway**, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. **Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.**

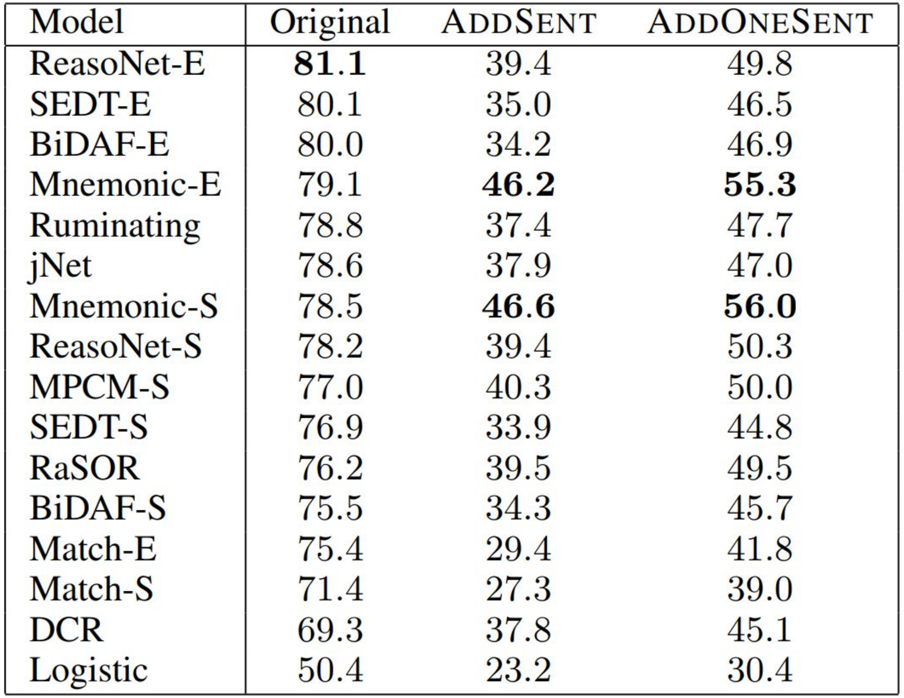
Question: What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer: John Elway

Model Prediction: Jeff Dean

Adversarial Testing

- The average performance of 16 published models trained on SQuAD drops from a 75% F1 score to a 36% F1 score



Model	Original	AddSent	AddOneSent
ReasoNet-E	**81.1**	39.4	49.8
SEDT-E	80.1	35.0	46.5
BiDAF-E	80.0	34.2	46.9
Mnemonic-E	79.1	**46.2**	**55.3**
Ruminating	78.8	37.4	47.7
jNet	78.6	37.9	47.0
Mnemonic-S	78.5	**46.6**	**56.0**
ReasoNet-S	78.2	39.4	50.3
MPCM-S	77.0	40.3	50.0
SEDT-S	76.9	33.9	44.8
RaSOR	76.2	39.5	49.5
BiDAF-S	75.5	34.3	45.7
Match-E	75.4	29.4	41.8
Match-S	71.4	27.3	39.0
DCR	69.3	37.8	45.1
Logistic	50.4	23.2	30.4

Question answering is not solved :(

- Sad day for AI
- Natural language understanding is still super hard
- Time to to get back to work

Adversarial Training

- We have found a hole in these models generalization capabilities
- A natural idea is to patch this hole by including these new examples in training, and this works perfectly well
- However, when we *prepend* the misleading sentence instead *appending* it we have a new adversarial test set our models fail on yet again

Old Adversarial Test Example

Passage:

Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by **John Elway**, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. **Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.**

Question: What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer: John Elway

Patched Model Prediction: John Elway

New Adversarial Example

Passage:

Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV. Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by **John Elway**, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager.

Question: What is the name of the quarterback who was 38 in Super Bowl XXXIII?

Answer: John Elway

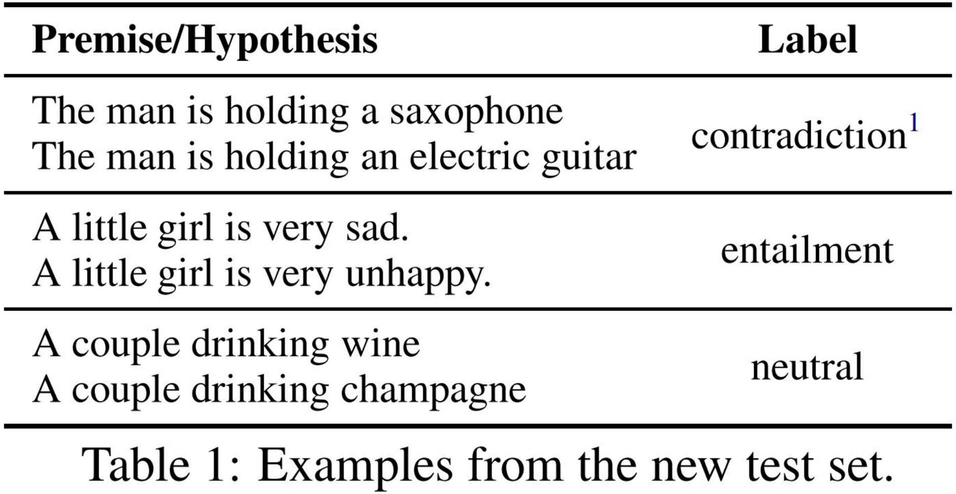
Patched Model Prediction: Jeff Dean

Adversarial Testing for NLI

- In the last couple years, there has been a growing number of more difficult generalization tasks developed for NLI
- This research has exposed the fragility of models trained on the SNLI and/or MultiNLI dataset

Breaking NLI Models with Simple Lexical Relations

Glockner et al. (2018) create an adversarial test set to expose that models have not fully learned lexical relations

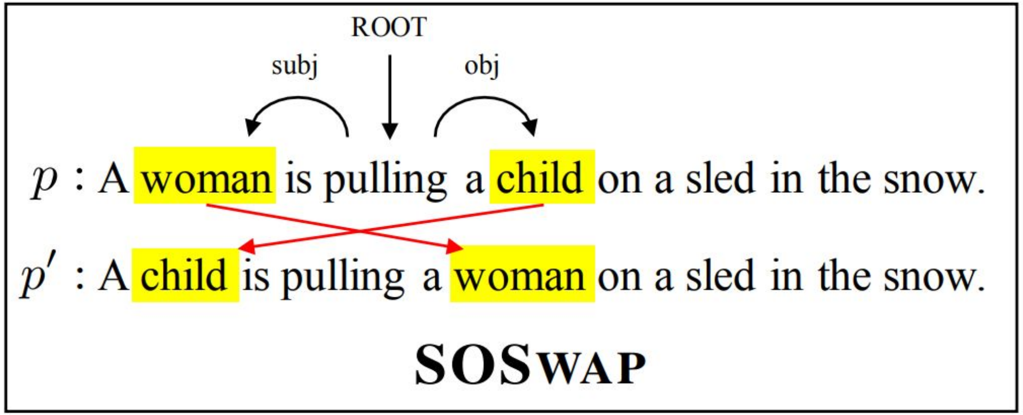


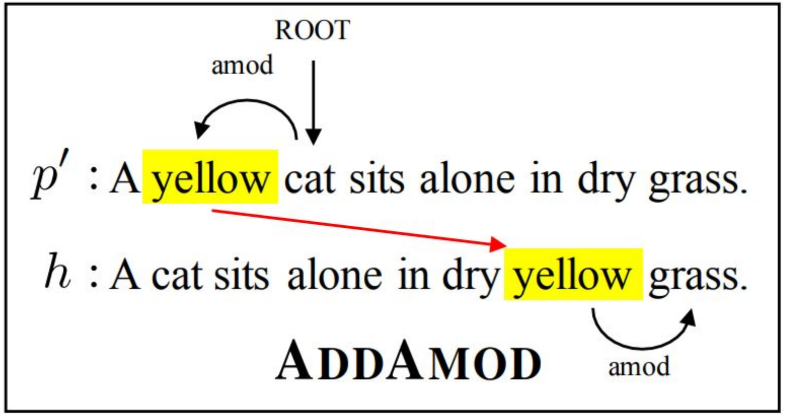
Premise/Hypothesis	Label
The man is holding a saxophone The man is holding an electric guitar	contradiction[1]
A little girl is very sad. A little girl is very unhappy.	entailment
A couple drinking wine A couple drinking champagne	neutral

Table 1: Examples from the new test set.

Evaluating Compositionality in NLI models

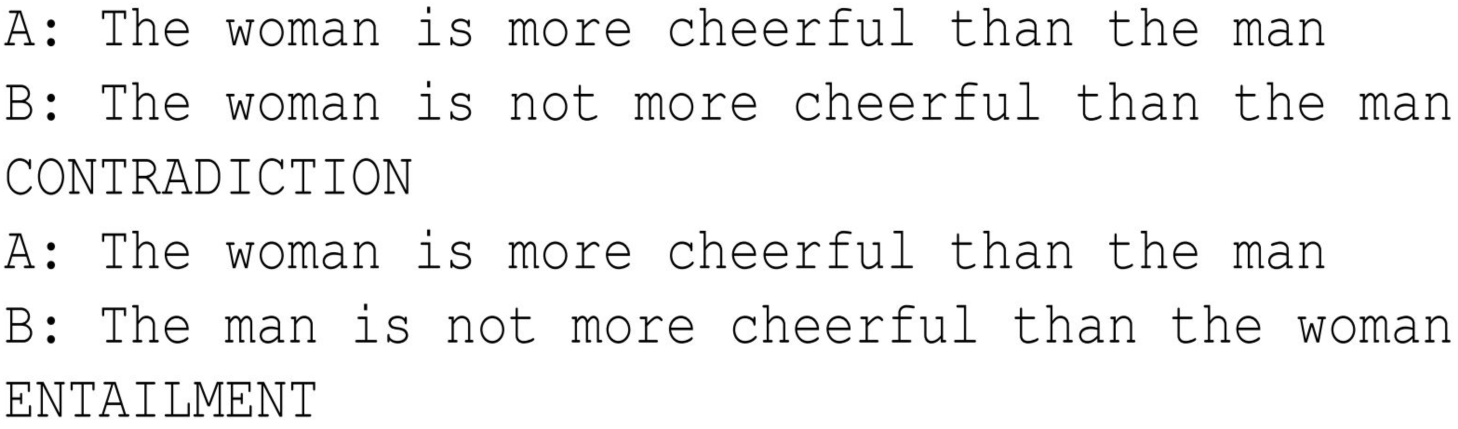
Nie and Wang et al. (2018) created adversarial testing examples to expose that models have not learned compositional semantics


ROOT
subj
obj
p : A woman is pulling a child on a sled in the snow.
p' : A child is pulling a woman on a sled in the snow.
SOSWAP


ROOT
amod
p' : A yellow cat sits alone in dry grass.
h : A cat sits alone in dry yellow grass.
ADDAMOD
amod

Evaluating Compositionality in NLI models

Dasgupta et al. (2018) expose that models fail to generalization to a particular compositional frame



A: The woman is more cheerful than the man

B: The woman is not more cheerful than the man

CONTRADICTION

A: The woman is more cheerful than the man

B: The man is not more cheerful than the woman

ENTAILMENT

Adversarial Testing for NLI

- You might wonder what these models have learned, if not lexical or compositional semantics!
- The NLP community has been hill climbing on the original SNLI test set from the moment it was released
- However, this is not the case for these new test sets
- In your projects, consider evaluating your models on these more difficult generalization tasks, where there is so much room for innovation and improvement

Artificial Generalization Tasks

- In my own research, I have constructed artificial NLI datasets
- The premises and hypotheses have the form Quantifier Adjective Noun Negation Adverb Verb Quantifier Adjective Noun
- Quantifiers can be *no*, *some*, *every*, or *not every*
- Negation and modifiers are optional
- My original intent was to stress NLI models with learning first order logical reasoning

An Example from my Dataset

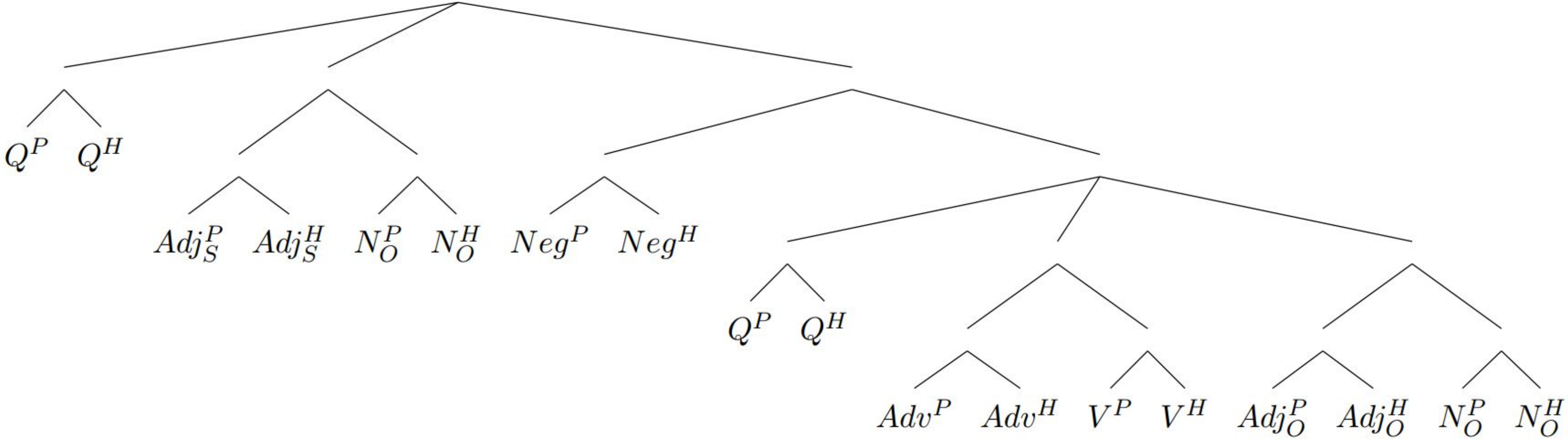
Every tall human does not kick any large rock

contradicts

No human angrily kicks some rock

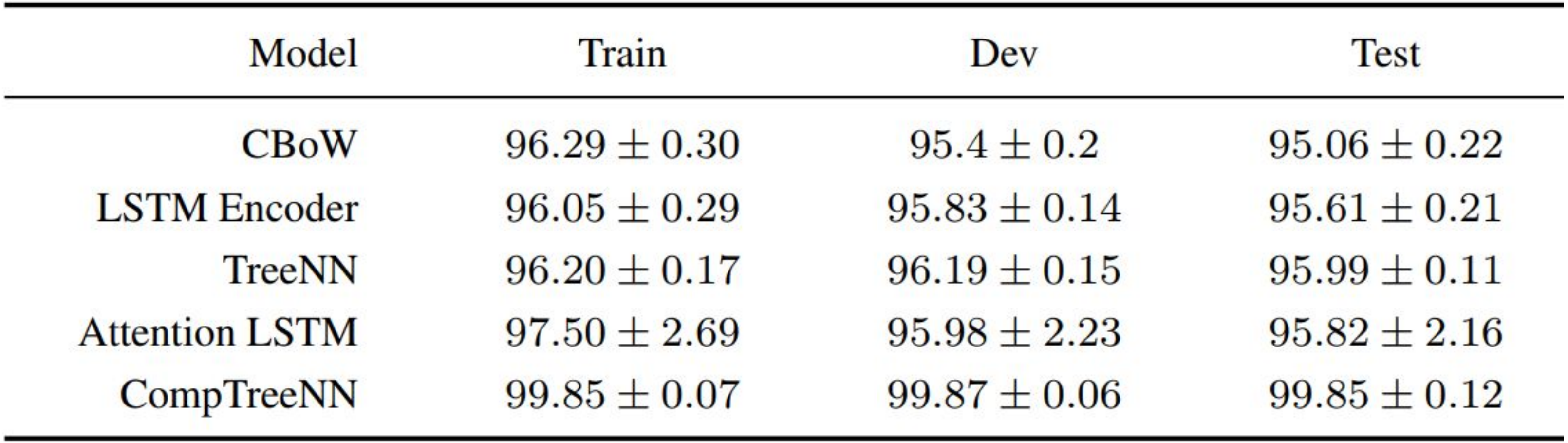
CompTreeNN Model

I tested standard neural models as well as task specific CompTreeNN model that jointly composes the premise and hypothesis



Standard Evaluation on My Data

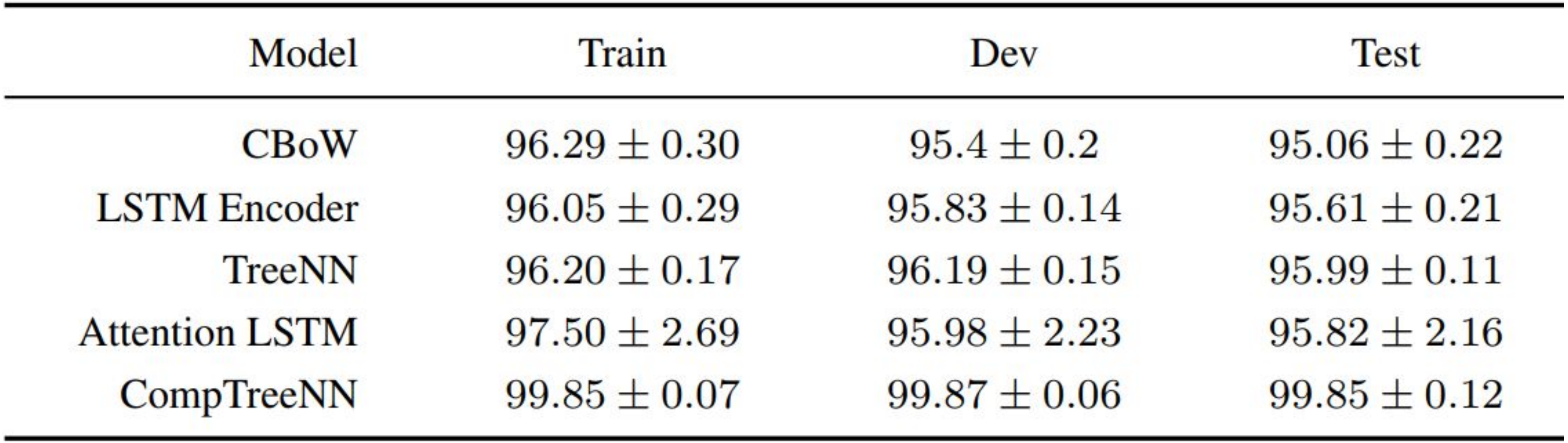
At first, I only performed a standard evaluation where I arbitrarily split my dataset into training and testing sets



Model	Train	Dev	Test
CBoW	96.29 ± 0.30	95.4 ± 0.2	95.06 ± 0.22
LSTM Encoder	96.05 ± 0.29	95.83 ± 0.14	95.61 ± 0.21
TreeNN	96.20 ± 0.17	96.19 ± 0.15	95.99 ± 0.11
Attention LSTM	97.50 ± 2.69	95.98 ± 2.23	95.82 ± 2.16
CompTreeNN	99.85 ± 0.07	99.87 ± 0.06	99.85 ± 0.12

Standard Evaluation on My Data

I discovered that standard neural models fail to encode the identity of verbs, nouns, adverbs, and adjectives while the CompTreeNN performs perfectly



Model	Train	Dev	Test
CBoW	96.29 ± 0.30	95.4 ± 0.2	95.06 ± 0.22
LSTM Encoder	96.05 ± 0.29	95.83 ± 0.14	95.61 ± 0.21
TreeNN	96.20 ± 0.17	96.19 ± 0.15	95.99 ± 0.11
Attention LSTM	97.50 ± 2.69	95.98 ± 2.23	95.82 ± 2.16
CompTreeNN	99.85 ± 0.07	99.87 ± 0.06	99.85 ± 0.12

Non-Standard Evaluation on My Data

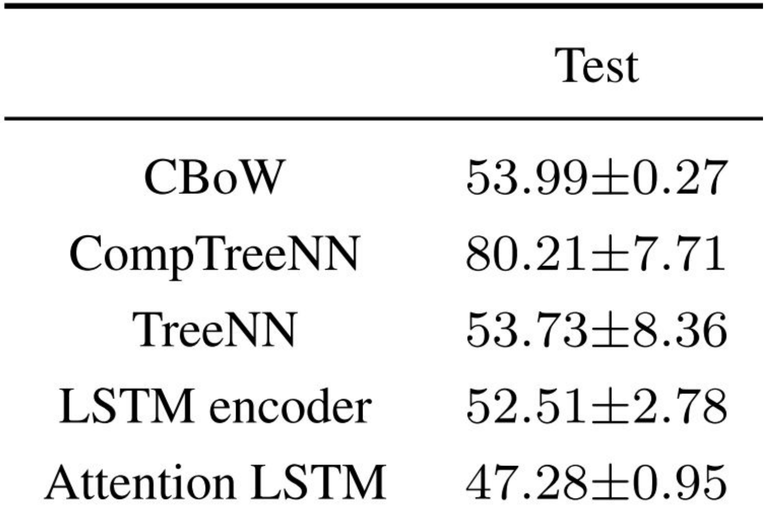
- I realized that in a standard evaluation, every possible combination of quantifiers, modifiers, and negation appear in training
- This meant a model that simply memorizes these combinations could succeed
- The standard evaluation ended up being far easier than I expected

Non-Standard Evaluation on My Data

- I decided to construct a train test split that evaluates a model's ability to perform natural logic reasoning
- I hand designed a simple baseline model that performs natural logic reasoning MacCartney and Manning (2009) or talk to Bill for more details on natural logic
- I then created a highly constrained dataset that this baseline model achieves perfect performance on

Non-Standard Evaluation on My Data

- On this task, standard models fail miserably, with only the CompTreeNN model achieving remotely good performance
- I believe this new task answers a far deeper question about these model's logical reasoning capabilities



	Test
CBoW	53.99±0.27
CompTreeNN	80.21±7.71
TreeNN	53.73±8.36
LSTM encoder	52.51±2.78
Attention LSTM	47.28±0.95

Moral of the Story

- Think deeply and carefully about what your learn from your experiments
- Often a generalization task will be far easier than you think
- Consider breaking from our standard evaluation framework to create more challenging generalization tasks that answer specific questions about model capabilities