Do NLP Models Know Numbers? Probing Numeracy in Embeddings

Eric Wallace*, Yizhong Wang*, Sujian Li, Sameer Singh, and Matt Gardner

Pretrained embeddings capture numerical information

- Understanding numbers is key for complex reasoning
- Most models treat numbers just like other tokens

Observation: models are already good at number questions!

. . . JaMarcus Russell completed a 91-yard touchdown pass to rookie wide receiver **Chaz Schilens**. The Texans would respond with fullback Vonta Leach getting a 1-yard touchdown run, yet the Raiders would answer with kicker Sebastian Janikowski getting a 33-yard and a 21-yard field goal. Houston would tie the game in the second quarter with kicker Kris Brown getting a **53**-yard and a 24-yard field goal. Oakland would take the lead in the third quarter with wide receiver Johnnie Lee Higgins catching a **29-yard** touchdown pass from Russell, followed up by an 80-yard punt return for a touchdown.

Q: How many yards was the longest field goal? A: 53 ✓

Q: How long was the shortest touchdown pass? A: 29-yard ✓

Q: Who caught the longest touchdown? A: Chaz Schilens ✓

Question Type	EM
Human (Test Set)	92.4
Full Validation	46.2 ← Decent
Comparative	73.6
Either-or	86.0
Superlative Questions	64.6

Great!

Idea: Probe the token embeddings

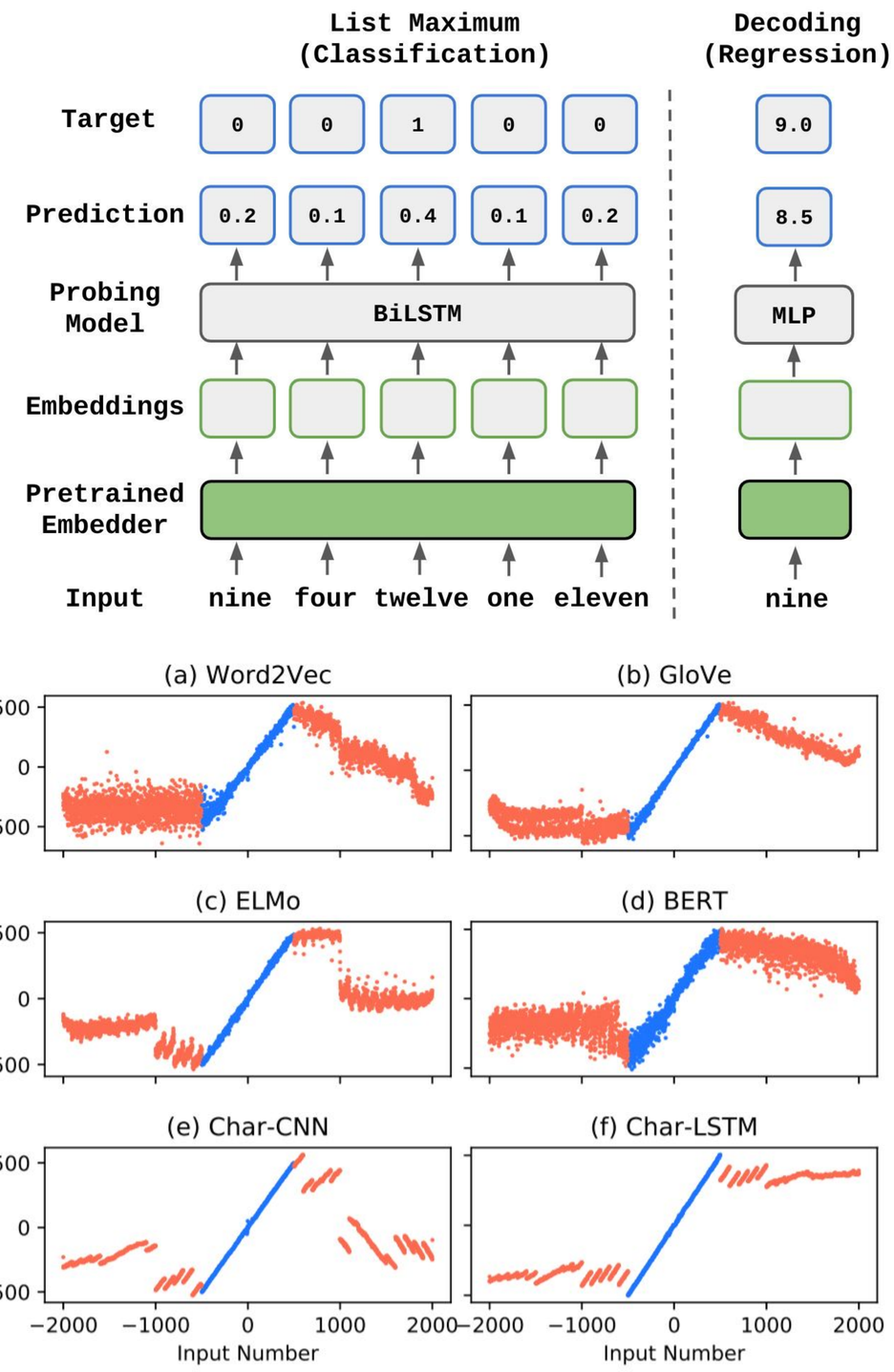
List maximum, decoding, addition

🙂 Interpolation succeeds

😢 BERT is mediocre

🙂 Floats/negatives/words

😢 Extrapolation fails



Interpolation	List Maximum (5-classes)			Decoding (RMSE)		
Integer Range	[0,99]	[0,999]	[0,9999]	[0,99]	[0,999]	[0,9999]
Random Vectors	0.16	0.23	0.21	29.86	292.88	2882.62
Word2Vec	0.90	0.78	0.71	2.34	18.77	333.47
GloVe	0.90	0.78	0.72	2.23	13.77	174.21
ELMo	0.98	0.88	0.76	2.35	13.48	62.20
BERT	0.95	0.62	0.52	3.21	29.00	431.78
Char-CNN	0.97	0.93	0.88	2.50	4.92	11.57
Char-LSTM	0.98	0.92	0.76	2.55	8.65	18.33