RichRep

Augmenting Cross Entropy-based supervised training with a contrastive signal

Team 31:

Rishabh Anand, Hu Yuxin, Marcus Peh,
Bhuvaneswaran Vignesh, Rahul Prasad, Thomas Cherian

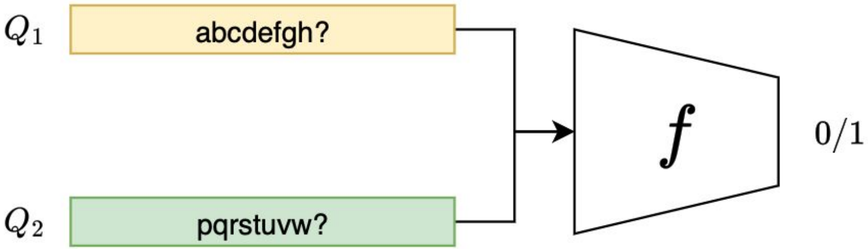
Introduction and Related Work

Duplicate Question Prediction

Quora Question Pairs

- GLUE benchmark task
- 400k training, 40k testing
- ~60% positive, ~40% negative

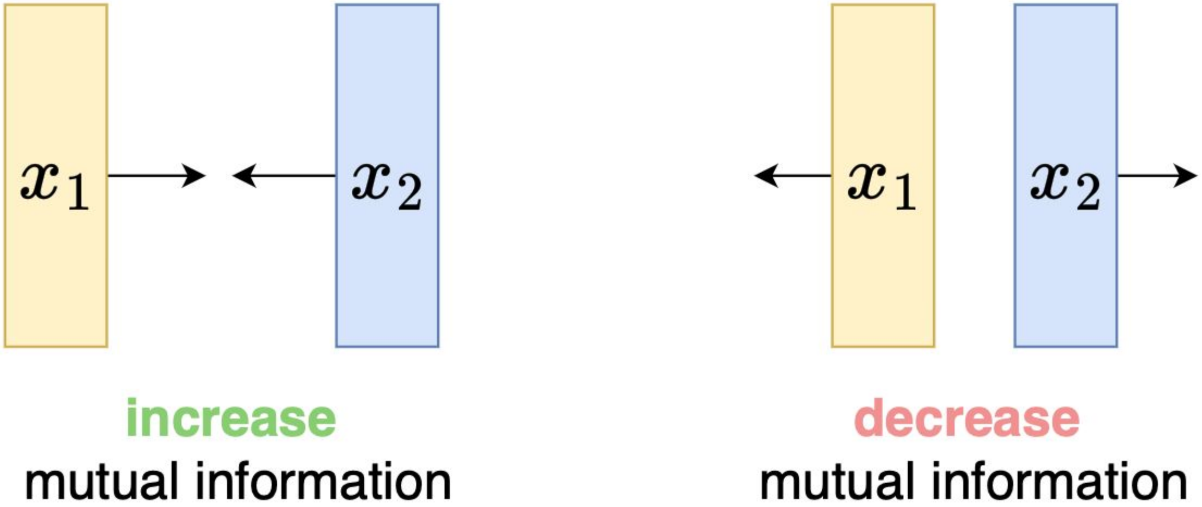
GLUE × Quora



Preliminaries

Contrastive Learning

- Remember good ol' cross-entropy?
  - It doesn't account for **pairwise relationships**
- How do we learn then?
  - What about working with with **features** instead?
  - How we address the shortcomings of supervised losses?
- Popularised by Facebook AI and Google Brain

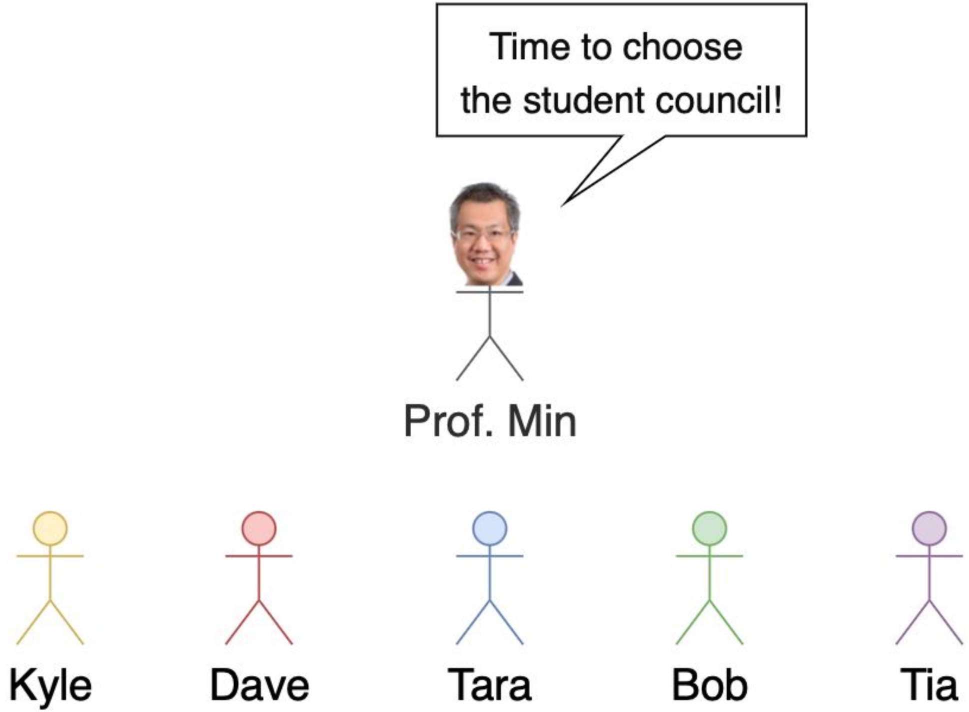


Like a **magnet**: like samples *attract*, unlike samples *repel* !!!

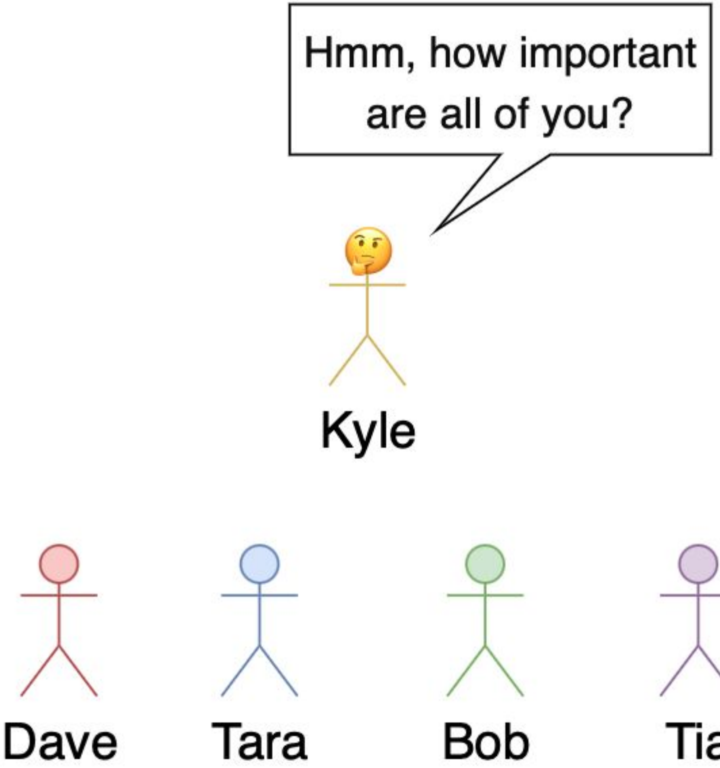
Self-Attention Block (Vaswani et al., 2017)

- Next in the evolution of sequence learners
    - Stacks of *Self-Attention* layers
    - High representation power
- Self-Attention, you say?
    - Learns global **long-context** information
    - Pairwise “voting” function among tokens

Research at Google

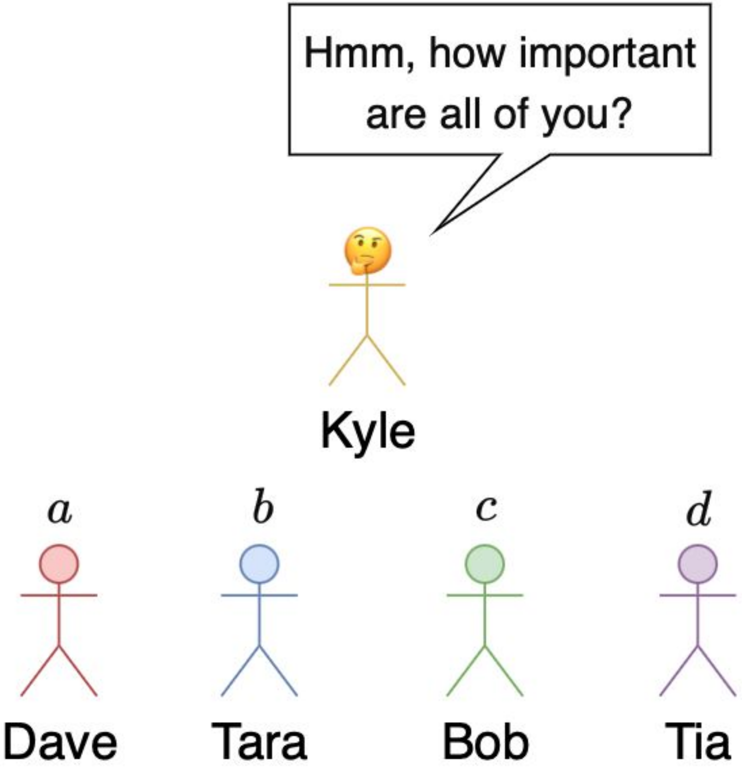


Self-Attention. Pairwise "voting" of importance



Self-Attention. Pairwise "voting" of importance

Repeat this
for everyone!



Self-Attention. Pairwise "voting" of importance

We will have 5^2 numbers

	Kyle	Dave	Tara	Bob	Tia
Kyle	·	a	b	c	d
Dave	·	·	·	·	·
Tara	·	·	·	·	·
Bob	·	·	·	·	·
Tia	·	·	·	·	·

students are the ***words***

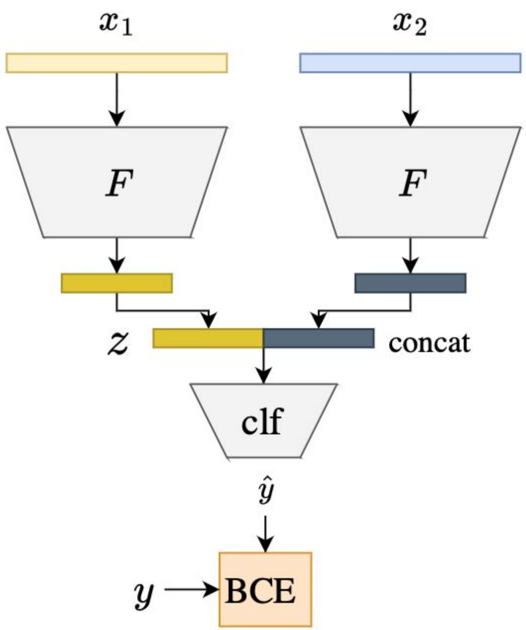
Self-Attention. Pairwise “voting” of importance

This Work

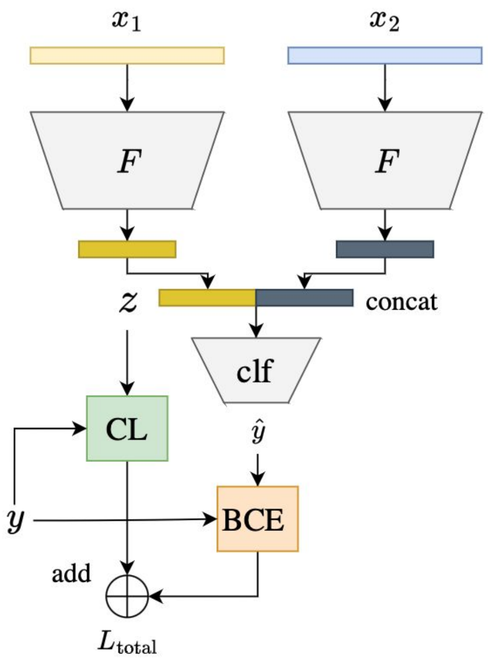
Project Aim

We show that Cross Entropy-based can be augmented with Contrastive Training!!!

RichRep Architecture & Losses



LSTM,
Self-attention block



Dual input encoder architecture popularised by **CLIP** and **DALLE**

Data Preprocessing

1. EDA did not reveal any useful strategies
    - Tried lemmatization, stopword removal, stemming, and cos-sim classification
2. **Sentence embeddings** over word embeddings
    - Given the diversity of sentences, a pretrained model (**BERT**) was appropriate
3. **Accuracy** as a metric
    - Existing literature uses accuracy when reporting results
    - Our data is mostly well-balanced

Contributions

We show,

1. adding an extra SSL signal **augments** supervised training loops
    - Seen through improved testing performance (~2% increase in accuracy)
2. SSL loss works across both **small** and **large** data regimes
    - Scalable across different training dataset sizes
3. SSL can perform the **heavy-lifting** during training if need be
    - SSL retains good performance on testing set with negligible supervised loss
    - But both must be used in conjunction

Contributions

We show,

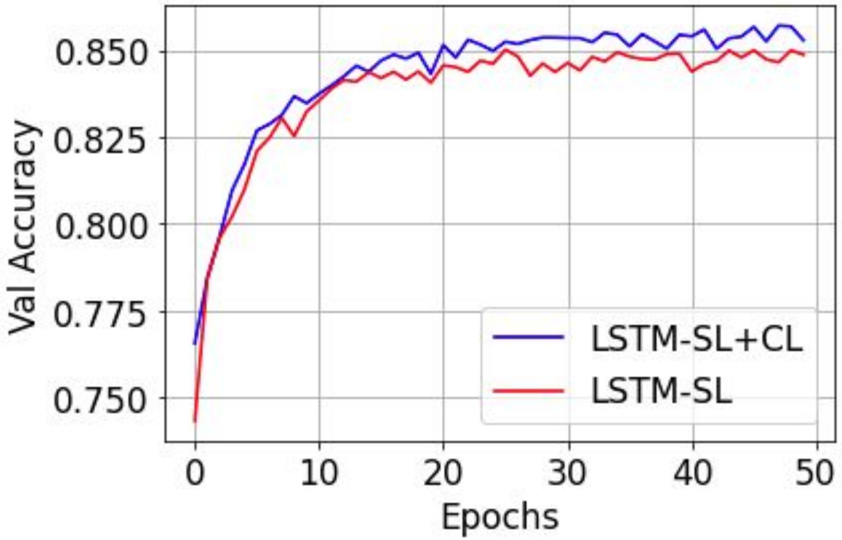
1. adding an extra SSL signal **augments** supervised training loops
    - Seen through improved performance on testing set
2. SSL loss works across both **small** and **large** data regimes
    - Scalable across different training dataset sizes
3. SSL can perform the **heavy-lifting** during training if need be
    - SSL retains good performance on testing set with negligible supervised loss
    - But both must be used in conjunction

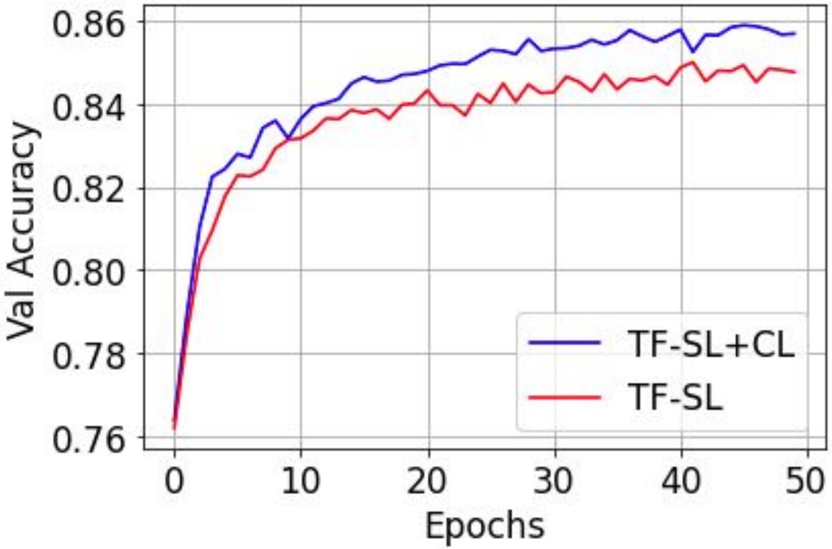
Contributions

We show that,

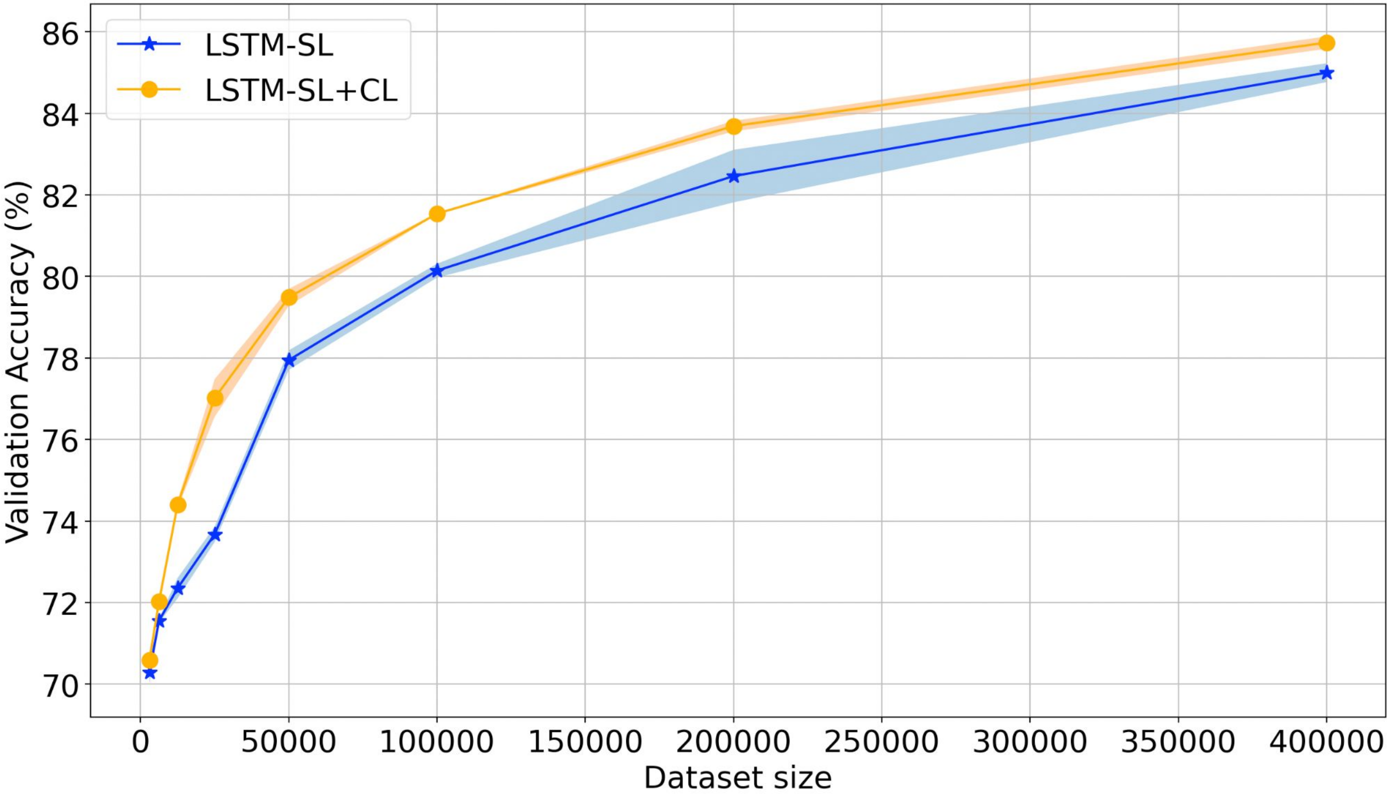
1. adding an extra SSL signal **augments** supervised training loops
    - Seen through improved performance on testing set
2. SSL loss works across both **small** and **large** data regimes
    - Scalable across different training dataset sizes
3. SSL can perform the **heavy-lifting** during training if need be
    - SSL retains good performance on testing set with negligible supervised loss
    - But both must be used in conjunction

Results and Discussion

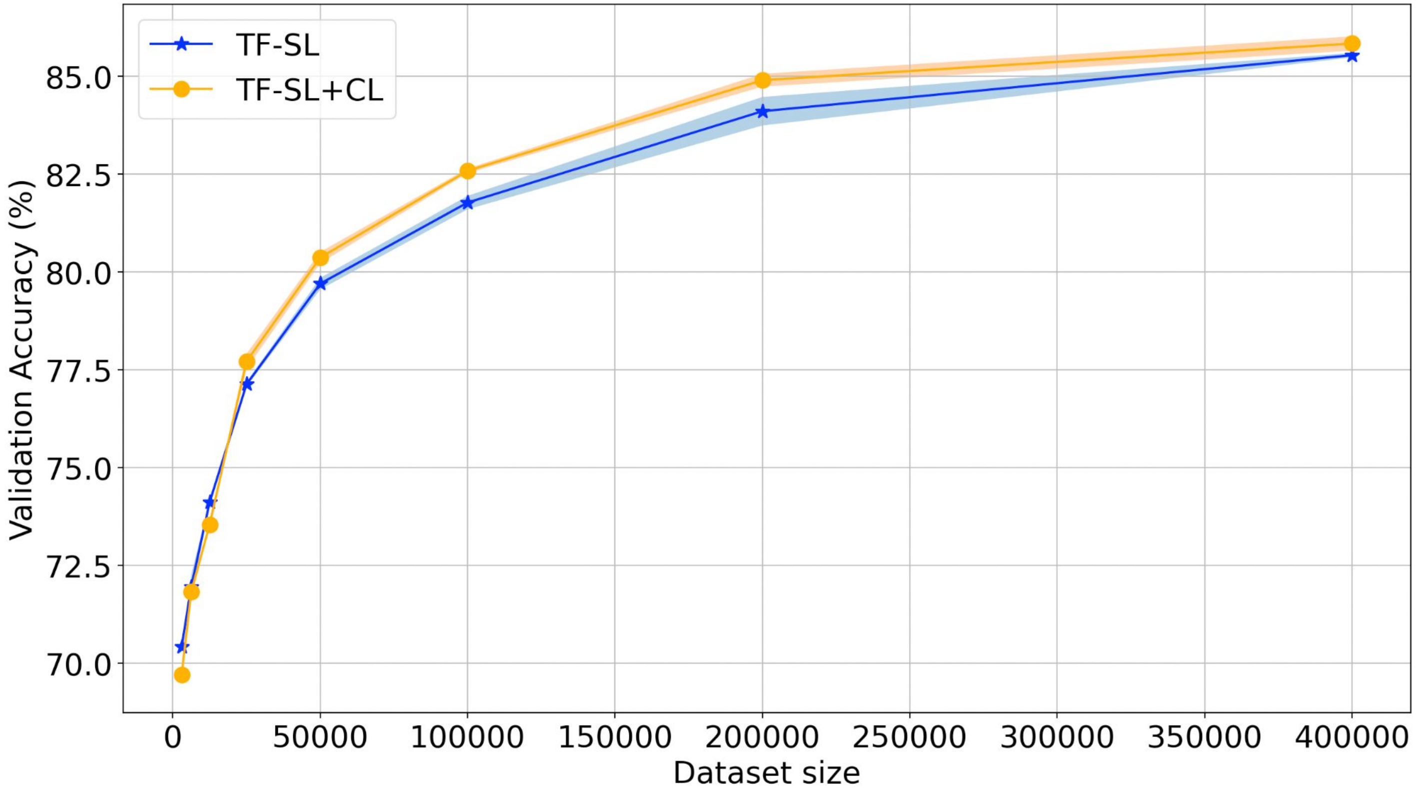




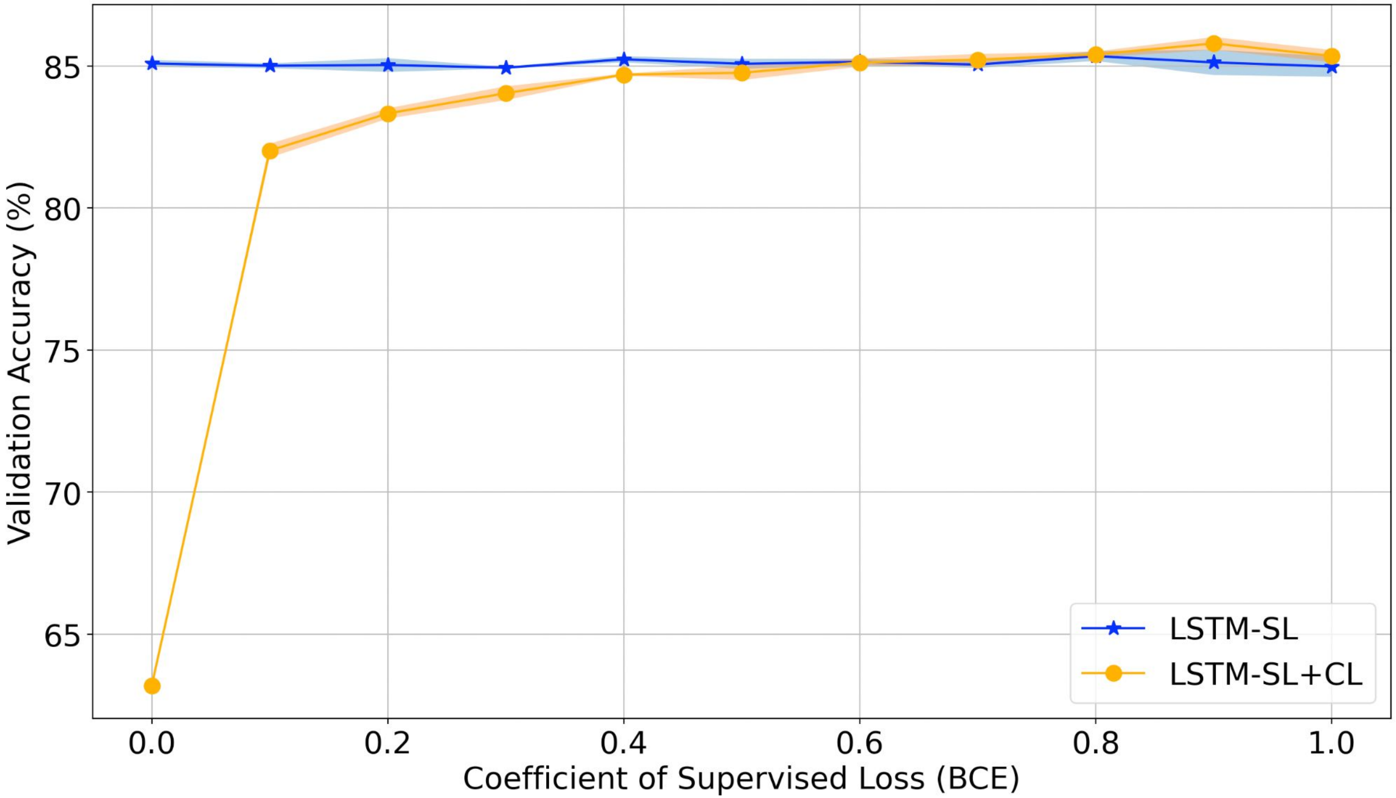
Result 1a. Adding CL loss improves performance across models



Result 2a. SSL loss works in both small and large data regimes (LSTM)



Result 2b. SSL loss works in both small and large data regimes (Transformer)



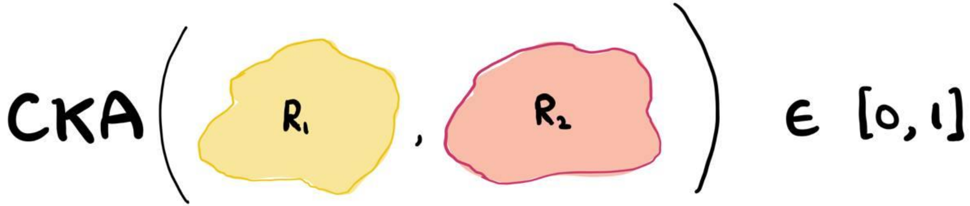
Insight 3a. Use SSL loss in conjunction with SL loss to reap benefits (LSTM)

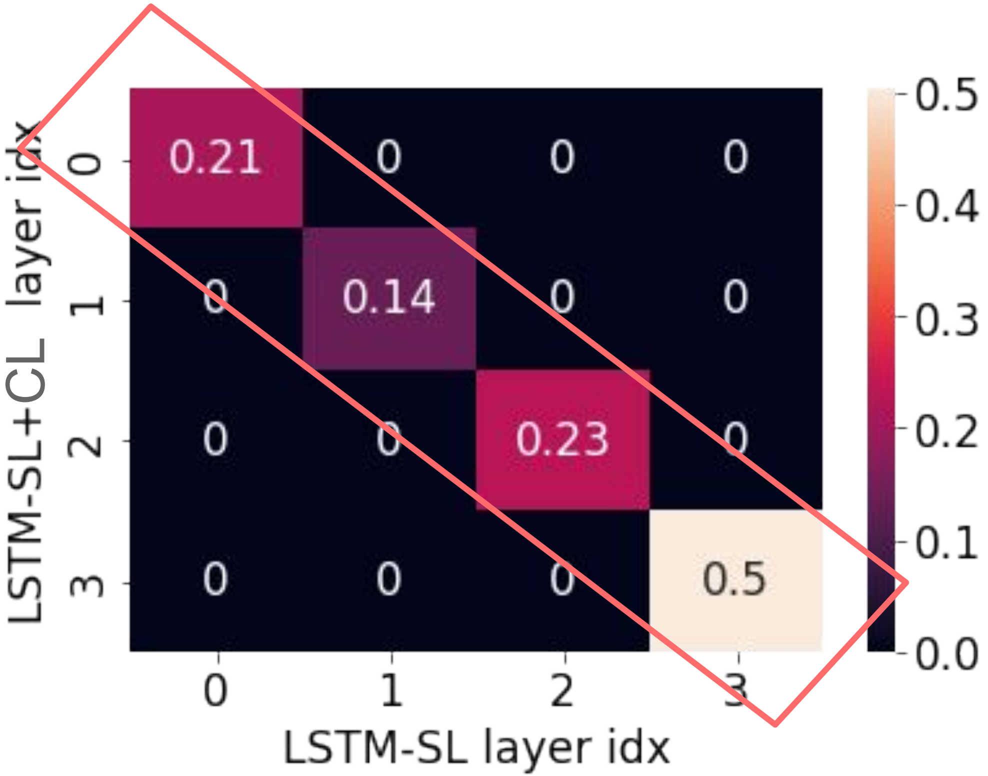
Further Analyses

(on LSTM)

Centered Kernel Alignment (CKA)

- Measure global similarity of representations of two models
- CKA Score between 0 to 1
  - Scores near 0 → models learned different things from data
  - Scores near 1 → models learned similar things from data
- Can be done on entire models or **individual layers**

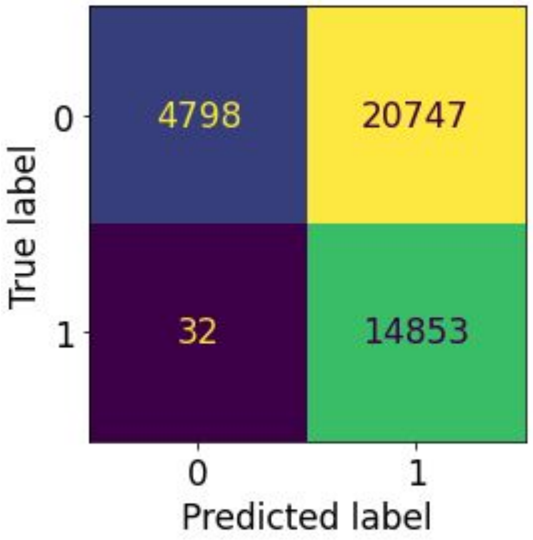




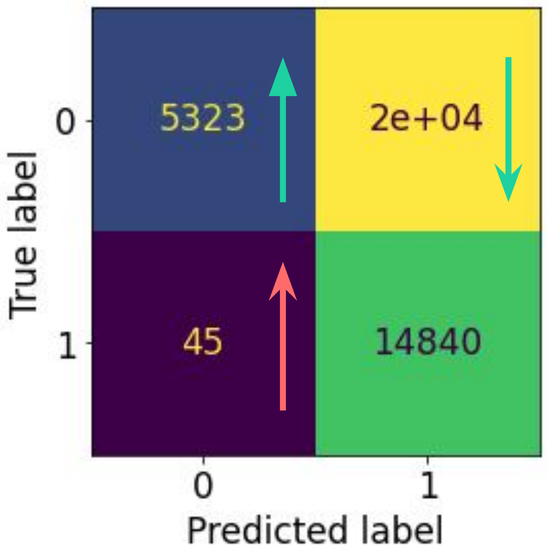
Analysis 1. SL and SL+CL models learn very differently on the same dataset

Validation accuracy breakdown:
Congruency in large components of failure

LSTM-SL



LSTM-SL+CL



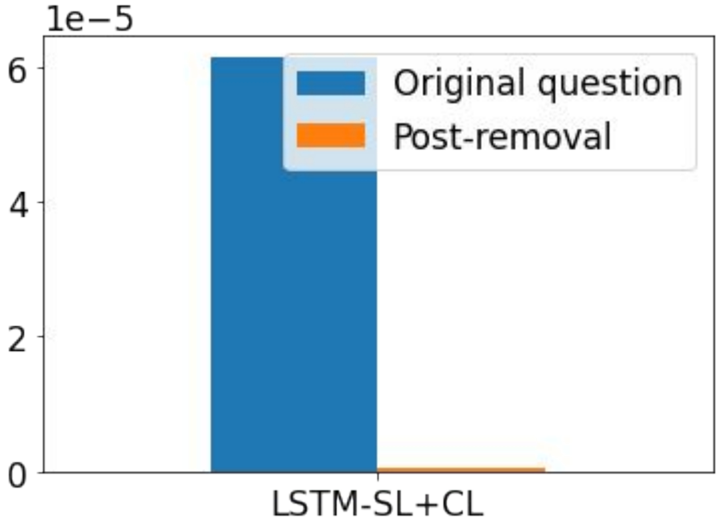
Analysis 2. SL+CL model improves *true negative* and *false positive* performance

What is it about the questions that severely affects the SL+CL model?

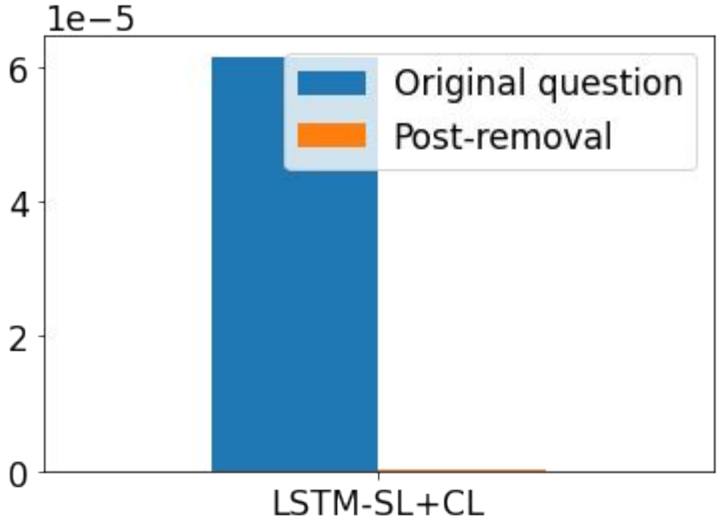
Consider **parts of speech**

Analysis 3. SL+CL is *highly sensitive* to key parts of speech.

Pre-sigmoid Logits



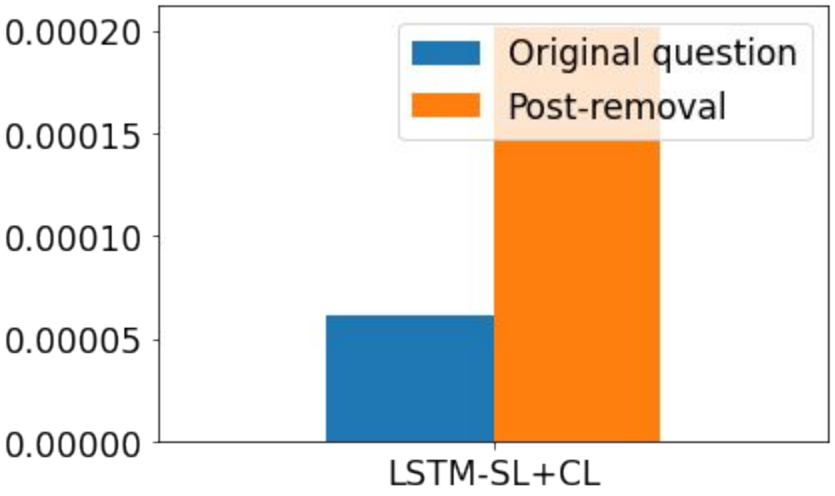
Removing **Verbs**



Removing **Nouns**

Analysis 3. SL+CL is *highly sensitive* to key parts of speech.

Pre-sigmoid Logits



Removing **Prepositions**

Analysis 3. SL+CL is *highly sensitive* to key parts of speech.

Ultimate conclusion:

You win some, you lose some

Parting Words

- **What our project is not:** a naive model comparison/showcase
- **What our project is:** a scalable framework to augment training
- CL helps learn richer representations
- Useful for **industry applications** with their rich data sources

Future Work

- Minimising the dependency on labels (via self-supervised learning)
- Incorporating **linguistics** into model architecture design
- Studying **topical** model performance (politics, education, entertainment, ...)
- Experimenting with *other* contrastive losses (eg: NXETent, NPairsLoss)

Appendix