Eyes Are All You Need?

CS4248-04

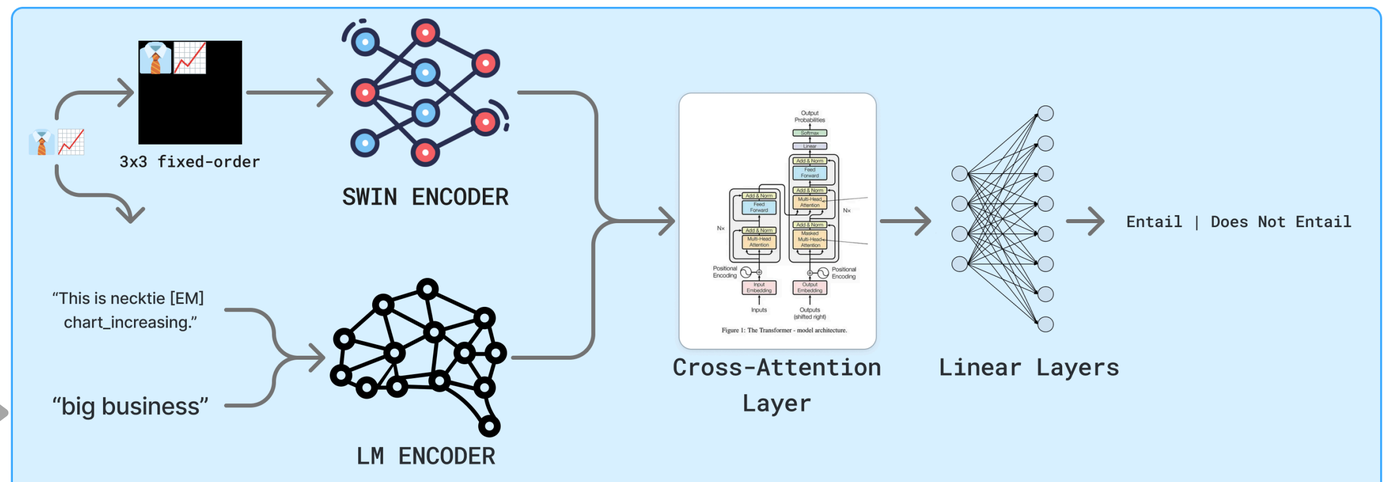
Natural Language Processing

Ryan Poon, Timothy Leow, Li ChengYue, Seth Chow, Hung Ming Kin

Does Visual Information improve performance on EmoTE task?

Basic Experiments

Augmented the original pipeline with a ViT (Visual Transformer) - the SWIN Encoder to introduce visual information into the equation. Encoded emoji sequences as a 3x3 fixed-order image



Fixed Order	BERT-base	RoBERTa-base	RoBERTa-large	BART-large	Total Improvements
Direct	✅	❌	❌	❌	1
Metaphorical	✅	❌	❌	❌	1
Semantic List	❌	✅	✅	❌	2
Reduplication	✅	✅	✅	❌	3
Single	✅	❌	❌	✅	2
Negative	Unimproved	❌	✅	✅	2
Overall	✅	❌	✅	❌	2
Total Improvements	5	2	4	2	13 / 28 (46.4%)

Ablated	BERT-base	RoBERTa-base	RoBERTa-large	BART-large	Total Improvements
Direct	✅	❌	✅	✅	3
Metaphorical	✅	❌	✅	❌	2
Semantic List	❌	✅	❌	❌	1
Reduplication	✅	✅	❌	❌	2
Single	✅	✅	❌	❌	2
Negative	❌	✅	❌	✅	2
Overall	✅	❌	❌	✅	2
Total Improvements	5	4	2	3	14/28 (50.0%)

Key Finding

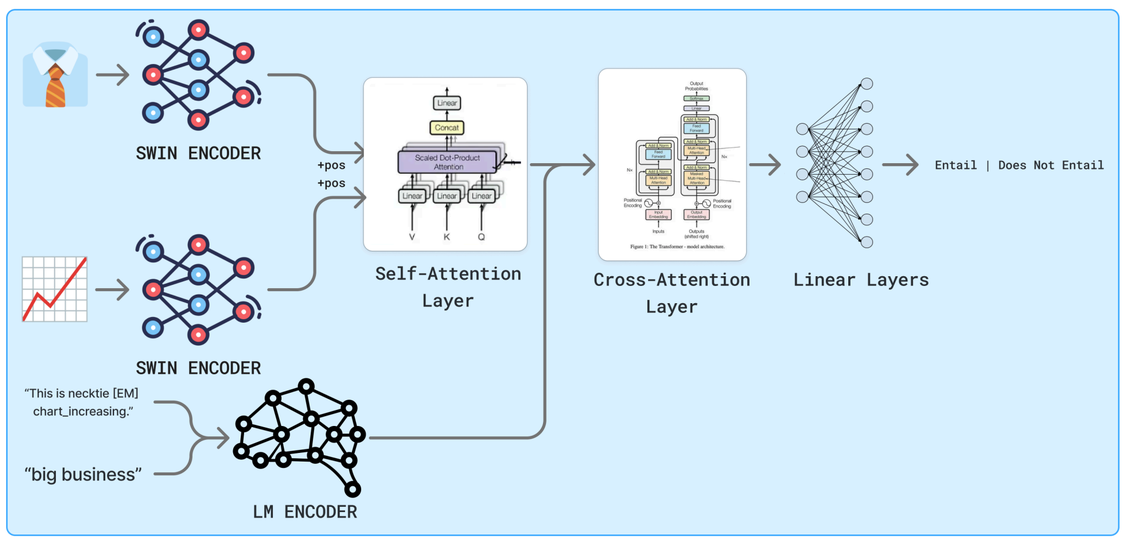
Our image representation does not preserve emoji sequences well, which may limit their ability to contribute to entailment due to the loss of positional information.

Positional Representation

Is there any way we can augment our representation to encode positional information better?

Positional Encoding Experiments

Introduced trainable positional embeddings with dynamic image inputs.

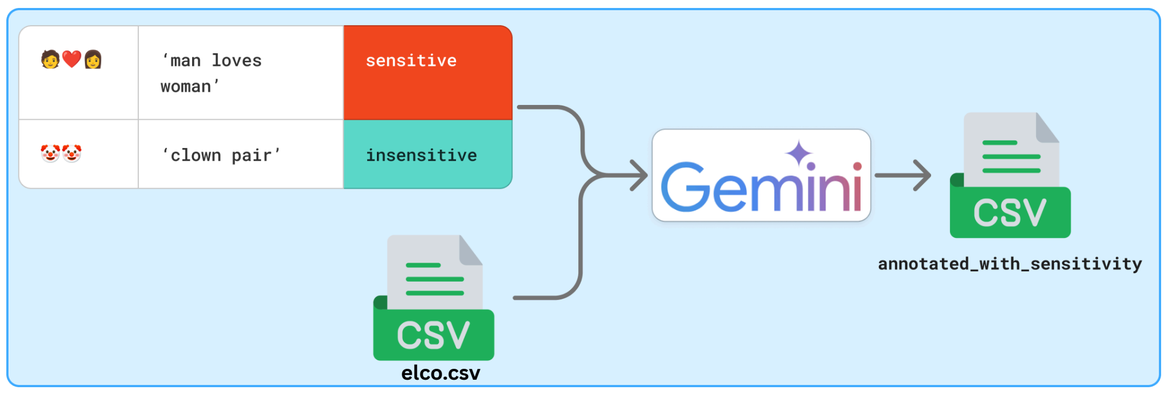


Positionally Sensitive Sequences

Which kind of sequences does our image format represent well?

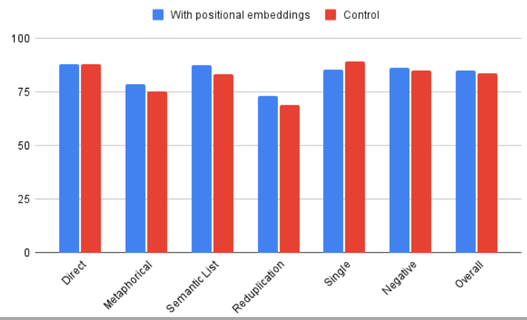
Positional Sensitivity Experiments

Annotated dataset with sensitivity, followed by further analysis on performance by sensitivity



With Positional Embeddings, against original

	BERT-base	RoBERTa-base	RoBERTa-large	BART-large	Avg
Direct	✅	✅	❌	✅	❌
Metaphorical	✅	✅	❌	❌	✅
Semantic List	✅	✅	❌	❌	✅
Reduplication	✅	✅	✅	❌	✅
Single	✅	✅	❌	✅	✅
Negative	✅	❌	✅	✅	✅
Overall	✅	✅	✅	✅	✅



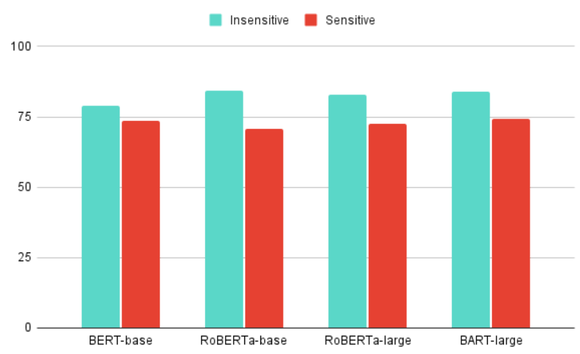
Without Positional Embeddings, against original

	BERT-base	RoBERTa-base	RoBERTa-large	BART-large	Avg
Direct	✅	✅	❌	❌	❌
Metaphorical	✅	❌	❌	❌	❌
Semantic List	❌	✅	❌	❌	❌
Reduplication	✅	✅	✅	❌	✅
Single	✅	✅	✅	✅	✅
Negative	❌	❌	❌	✅	❌
Overall	✅	✅	❌	❌	❌

Key Finding

Performance is generally better after adding learnable position embeddings, except for the samples that only contain single emoji. This is further explained by positional sensitivity experiments on the right

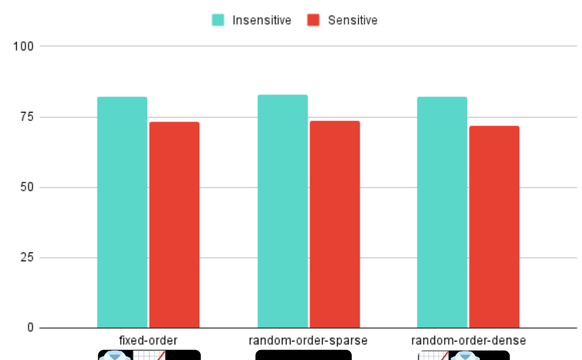
Comparing performance by sensitivities



Key Finding

Performance is higher on positionally-insensitive sequences using our representation.
This suggests our image format better represents positionally-insensitive sequences

Comparing performance across image formats



Key Finding

Performance is similar across 3x3 fixed order, 3x3 random dense, and 3x3 random sparse image formats. This suggests the model likely cannot infer sequence-related information from emoji placement.

Semantic and Positional Degradation

What are the consequences of losing both semantic and positional information in emoji sequences for entailment prediction?

Text Ablation Experiment

Ablated text input from the model.

Key Finding

Swin/CNN + GRU/LSTM achieves high accuracy on single-image inference but fails to capture sequential patterns in emoji sequences. This is due to the lack of pretrained temporal priors and insufficient data to learn complex dependencies from scratch.