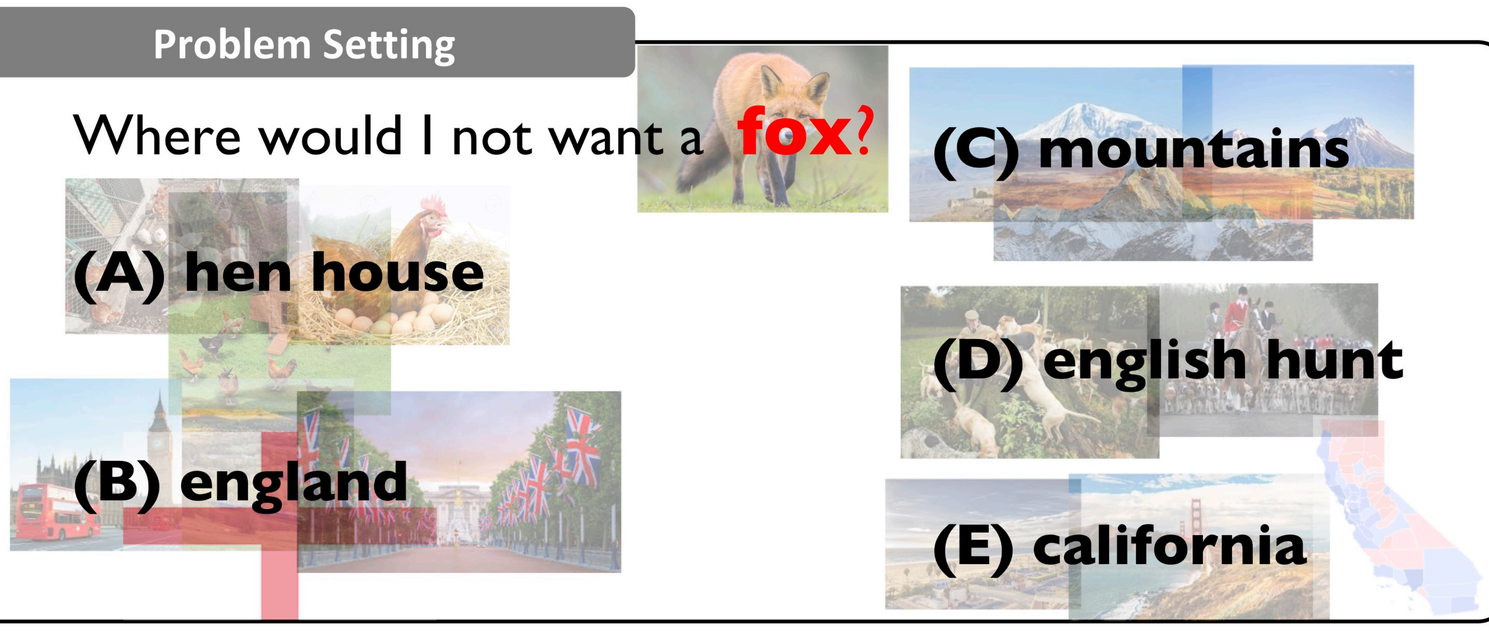
Multi-Modal Embeddings for Common Sense Reasoning

Aishwarya Kamath[1], Jonas Pfeiffer[2]
[1] New York University, [2] Technische Universität Darmstadt

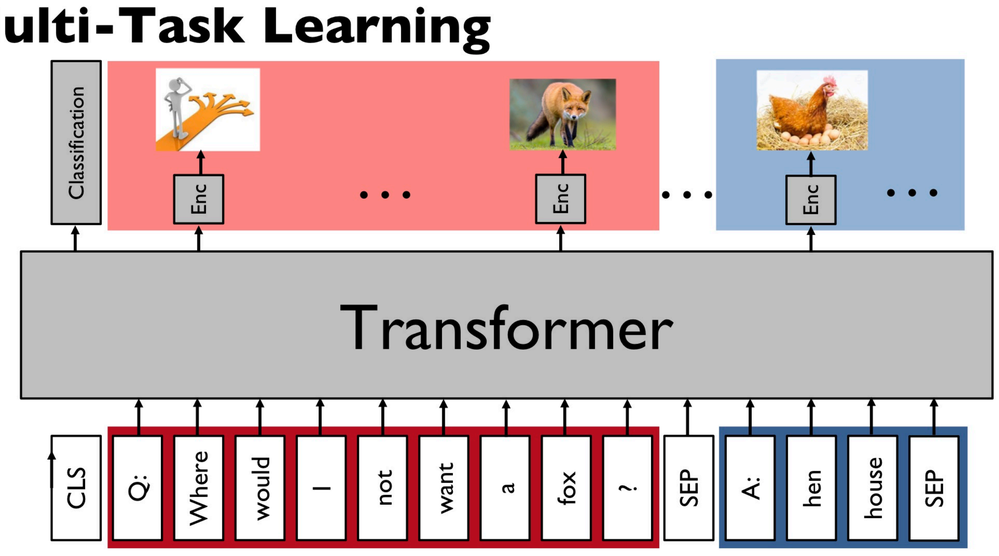
Problem Setting

Being able to accurately represent the meaning of sentences requires large amounts of **background information** that is not easily available just in the form of text.
A large amount of **common sense** is represented in the **visual world.**



Models

Multi-Task Learning



The output of the transformer at each time-step is used to predict the image representation (ResNet pooled features) of the word at that time-step.
The gradients from the image prediction module update the transformers weights to also encode the visual semantics.

Fusion

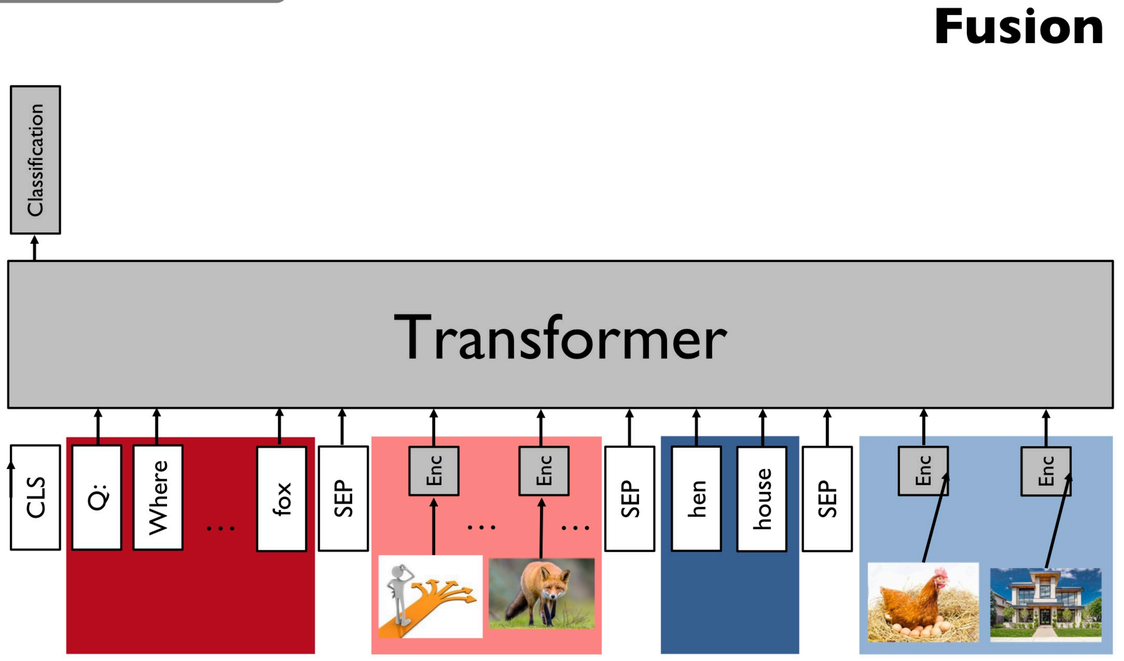


Image representations (ResNet pooled features) are passed through a linear layer (affine-transform) and then input to the transformer along with the text representation.
The model captures the multi-modal semantics by having access simultaneously to text as well as image information.

Challenges

Images that are acquired from querying google, and their subsequent pooled feature representations extracted using a ResNet are very noisy. We utilise the concreteness of words as a proxy for how useful their visual representations would be.
=>Database with concreteness scores for 40,000 English word lemmas
=>Choose a threshold for the concreteness value

Results

Current models do not perform better than text only models. This calls for more sophisticated approaches.

However anecdotal results show that the multimodal approach correctly classifies examples that (might) require visual input:

Who is likely to be excited about a crab?
fish market, pet shop, **fishmongers**, intertidal zone, obesity