Question answering using SOTA Transformer models shows brittleness to domain transfer. Soft contextual data augmentation (SCA) is a data augmentation strategy that has shown promising results on neural machine translation [1] and we apply this to out-of-domain QA.

Does soft contextual data augmentation improve performance on out-of-domain QA?

Our method: SCA with DistilBERT
- We use a pre-trained DistilBERT to output the probability distribution at a sampled token position.
- We truncate the probability distribution to the top k tokens in the vocabulary and take a linear combination of their embeddings.

Resulting soft token embedding:

\[ E_k(x_i) = \frac{\sum_{j=1}^{k} p[w_j | x_{ci}, x_{ci}] E[w_j]}{\sum_{j=1}^{k} p[w_j | x_{ci}, x_{ci}]} \]

Related work
- Other data augmentation methods for OOD QA [2]:
  - Domain sampling: determine which datasets can contribute more to OOD performance.
  - Negative sampling: include “no answer” segments and abotional option for model.
  - Back-translation: translate data to pivot language and back to target language.
  - Active learning: sample examples based on difficulty calculated by scoring functions.

Data and setup
- In-domain datasets: SQuAD (Wikipedia), Natural Questions (Google queries on Wikipedia), NewsQA (news articles).
- Out-of-domain datasets: RACE (reading exams), Duorc (movie reviews), RelationExtraction (synthetic relation questions).
- QA task: input: context paragraph and question; output: answer: span.
- Models: Pre-trained DistilBERT used as LM to generate soft tokens and as QA model to train on augmented data to perform QA task.

Experimental results
- Tuned hyperparameters to k = 5 and \( \eta = 3 \times 10^{-5} \).
- Across 3 masking strategies and varying % augmented, best model was masking 10% of tokens separately F1: 49.2, EM: 32.2, (baseline: F1: 47.72, EM: 30.65) (Fig 1).
- Improvement is specific to OOD dev sets (Fig 2), suggesting SCA improves robustness specifically.
- Augmenting only context: similar improvements.

Analysis
- Lower rate of complete misses:
  - Out of 382 dev examples, our model predicted the exact answer and baseline didn’t for 25, and 9 vice versa.
  - Counted “complete misses” (CM): incorrect answers which had no containment relation with the correct answer.
  - Out of 9 our model got wrong, 22% were CM. Out of 25 baseline got wrong, 44% were CM. Example of non-CM:

  Context: NKGD2 is encoded by KLRK1 gene which is located in the NK-gene complex (NKC) situated on chromosome 6 in mice and chromosome 12 in humans.
  Question: What is the name of the chromosome where you can find NKGD2?
  Correct answer: Chromosome 12
  Our model’s answer: Chromosome 6 in mice and chromosome 12

- Higher success rate on context-question pairs where paraphrasing is important:
  - Counted context-question pairs which had paraphrasing between context and question.
  - Out of 9 our model got wrong, 22% had paraphrasing. Out of 25 baseline got wrong, 40% had paraphrasing.
  - To analyze the effectiveness of our soft tokens, we fed the context into the pre-trained DistilBERT we used in SCA to see if paraphrased words in the question were included in the truncated soft token distribution.

Context: Griffin gives them the Archevet explains it can only work in zero gravity: K gets the idea to head to Cape Canaveral on 16th July 1963 (the day the Apollo 11 ship launched).
Question: Where must they go to attach the Archevet?
Correct answer and our model’s answer: Cape Canaveral. Baseline’s answer: No k.

Conclusion
- Soft contextual data augmentation (SCA) improves QA performance specifically on out-of-domain datasets. Qualitatively, our results suggest that SCA tends to improve performance when paraphrasing is involved between the context and the question.