1. The Novel Object Captioning Task

- Describe images containing novel objects (not present in the available image-caption training data) by learning from image labels or object annotations.
- Motivation: Scaling image captioning to many more visual concepts, without collecting expensive caption training data.

Training data:
- Image-caption data for 72 CCOC classes.
- Image labels for 8 CCOC classes.

Motivation:
- Caption training data) by learning from image labels or object annotations.

Beam 0:
- Initial state in the FSA.

Beam 1:
- A general algorithm for training RNNs on partially-specified sequences.

Beam 2:
- Given a dataset of partially-specified training sequences x and current model parameters θ, iterates these two steps:
  1. Estimating the complete data y by approximating \( p(x|y) \approx \int p(x|s) p(s|y) ds \), using constrained beam search (CBS), where \( \theta = (\mathbf{L}, \mathbf{S}, \mathbf{F}, \mathbf{P}) \) is an FSA that recognizes sequences that are consistent with the observed partially-specified sequence x.
  2. Learning (or updating) the model parameters by setting \( \theta \leftarrow \arg \max_{\theta'} \sum y \log p(y|\theta') \).

Beam 3:
- Corresponds to the FSA CBS decoding example with beam size 2, at time step 4. There is one destination beam for various next words.

2. Approach: Partially-Specified Sequence Supervision (PS3)

- A general algorithm for training RNNs on partially-specified sequences.
- Given a dataset of partially-specified training sequences x and current model parameters θ, iterate these two steps:
  1. Estimating the complete data y by approximating \( p(x|y) \approx \int p(x|s) p(s|y) ds \), using constrained beam search (CBS), where \( \theta = (\mathbf{L}, \mathbf{S}, \mathbf{F}, \mathbf{P}) \) is an FSA that recognizes sequences that are consistent with the observed partially-specified sequence x.
  2. Learning (or updating) the model parameters by setting \( \theta \leftarrow \arg \max_{\theta'} \sum y \log p(y|\theta') \).

Beam 3:
- For a set of pictures showing a slice of traffic.

3. Constrained Beam Search (CBS)

- CBS decoding example with beam size 2, at time step 4. There is one search beam for each state in the FSA. Beam 3 corresponds to the FSA accepting state \( s_2 \).

Step 2: Backprop errors using the entire vocab & minibatch as usual.

A minibatch of images with both complete and partially-specified captions.

4. Results and Examples