Hello everyone, I'm Zhaokai Wang from Team colab_buaa. Our team members include Zhaokai Wang, Renda Bao and Si Liu from Beihang University. On behalf of our team, I'd like to give a presentation of our TextCaps submission.

[next page]

Our model is based on M4C_captioner, and we mainly focus on improvement of OCR. Firstly we give an intuitive understanding of Rosetta's OCR results. Our study shows that among the words of all ground-truth captions in training set, only 4.95% appears in OCR and 73.93% only appears in vocabulary, which means about 20% of the words are unpredictable with Rosetta. We assume that those words are mostly rare words and should come from OCR. This shows that OCR sets an upper bound for accuracy and we should focus on its improvement.

[next page]

We use two methods for better text spotting. The first one is CRAFT model. It can detect text area by exploring character-level affinity. We also use affine transformation to transform irregular bounding box into rectangle for better recognition result. and here is an example.

[next page]

The second spotting method is ABCNet. It uses Bezier curve to detect oriented and curved text and achieves remarkable result. Here are two demos from the Totaltext dataset.

[next page]

After cutting out the text area, we use four-stage STR framework for recognizing. It first uses thin-plate spline to transform the original image into normalized image. Then, it extracts visual features with ResNet, and captures contextual information with BiLSTM. Finally, it predicts a sequence of characters with previous features.

[next page]

We combine new OCR tokens with original Rosetta tokens and here is the result. We can see that OCR tokens nearly trippled with our new methods.

[next page]

Here are our results on test set of TextCaps. There are 4.4% improvement of CIDEr with CRAFT, and another 3% improvement with ABCNet and affine transformation.

[next page]

We also have some ideas that may work in the future, like considering recognition confidence when making predictions. That's because we find words with higher confidence are more recognizable and more likely to be critical information that need to be included in captions.

We also find some repetition of words in predicted captions like these on the slide. Coverage mechanism may be helpful to avoid this redundancy.

[next page]

That's all. Thank you. If you have any questions, please feel free to ask me by email.