Applying NLP Deep Learning Ideas to Image Classification

Gary Ren
garyren@stanford.edu
gren@microsoft.com

Background

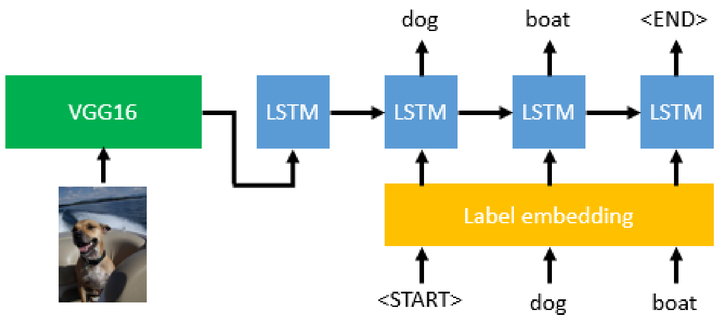
- Task: multi-label image classification
  - Identify all the different objects in an image
- Why: many real world images contain multiple objects
- How: deep learning ideas from NLP
  - Word embedding
  - Recurrent neural networks
  - Attention mechanism

Dataset

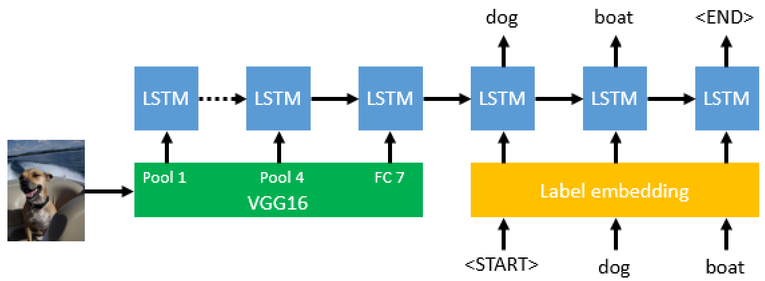
- Microsoft Common Objects in Context (MS-COCO)
  - Image classification, segmentation, and captioning dataset
  - 123k images, 80 object types
  - Wide range of objects from dogs to airplanes to pizza
  - 2.9 different objects per image on average

Models

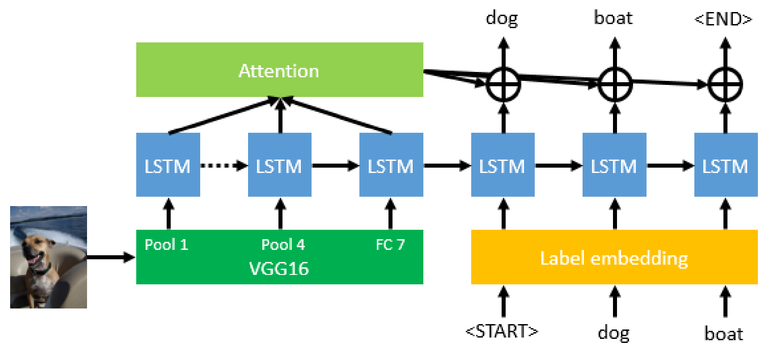
- CNN as initialization



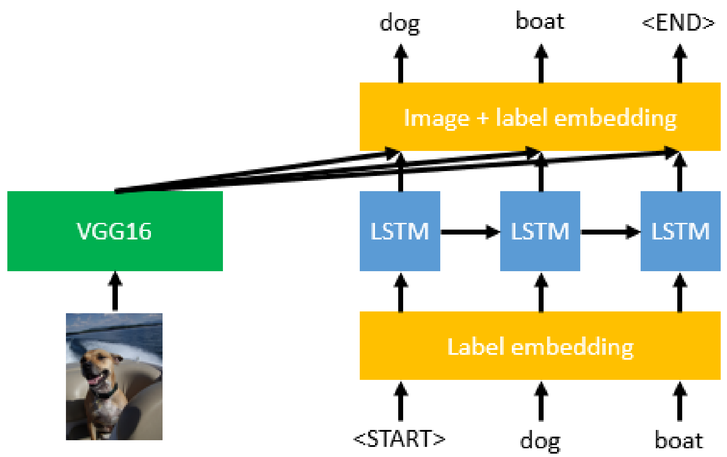
- CNN sequence to sequence



- CNN sequence to sequence w/attention



- Image/label joint embedding



Results

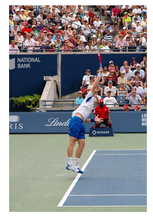
Model	Precision	Recall	F1
WARP	59.8	61.4	60.7
Softmax	60.2	62.1	61.1
Binary cross entropy	61.7	65.0	63.3
CNN initialize – hidden state	64.0	64.1	64.1
Image/label joint embedding	66.6	65.3	65.9
CNN initialize – input	67.2	66.1	66.6
CNN seq2seq attention	67.7	65.6	66.6*
CNN seq2seq	68.9	65.6	**67.2**

*still training

Analysis

- Helps to identify objects that commonly appear together
- Helps to identify objects that are visually barely discernible

laptop
keyboard
mouse
tv



person
tennis racket
sports ball
chair

Conclusion

- Applying word embedding, RNNs, and attention to multi-label image classification shows improvement over traditional methods
- Future work:
  - More hyperparameter tuning
  - Apply attention to image/label joint embedding