What is needed for simple spatial language capabilities in VQA?

Alexander Kuhnle
Ann Copestake
Department of Computer Science and Technology
University of Cambridge
{aok25,aac10}@cam.ac.uk

Abstract datasets for spatial relations
ShapeWorld simulator to generate VQA-style data (https://github.com/AlexKuhnle/ShapeWorld)
Task: image caption agreement, with caption interpreted as question and agreement as yes/no answer.
Data: image of size 64 × 64 with 4-10 objects, accompanied by a natural language statement, plus binary agreement value.
Dataset: explicit relation
Spatial relation acts as the verb phrase, “X is to the left of Y.”
Dataset: implicit comparative
Spatial relation is realised as adjectival predication in positive/comparative form, “The left X is Y.” (of two X).
Dataset: implicit superlative
Spatial relation is realised as adjectival predication in superlative form, “The leftmost X is Y.” (of at least two X).
Spatial relations
relative x-position: ‘left/-most’ and ‘right/-most’, relative y-position: ‘above/upper/-most’ and ‘below/lower/-most’, relative z-position: ‘behind’ and ‘in front of’, relative distance: ‘closer/-est to the X than’ and ‘farther/-est from the X than’.

VQA models and experimental setup
Uniform hyperparameter setup (for fair comparison)
Hyperparameters are aligned across all models for the generic parts (non-core modules) of their architecture.
Image module: three convolution layers, 128 kernels, size 3 × 3, stride 2, followed by BN and ReLU.
Language module: 128-dim. embeddings, processed by LSTM/GRU of size 512 (128 in case of RenNet).
Classification module: 1024-dim. hidden layer followed by BN and ReLU, then 2-dim. softmax layer yielding answer distribution.
Optimization: Adam with learning rate 0.0003, batch size of 64.
Training: 200k iterations ~ 25 epochs, averaged over 3 runs.

VQA models / core modules
https://github.com/AlexKuhnle/film
- CNN / LSTM: uni-modal vision/language baseline
- CNN-LSTM: multi-modal late-fusion baseline
- SAN: Stacked Attention Network (Yang et al., 2016)
- RenNet: Relation Network (Santoro et al., 2017)
- FiLM: Feature-wise Linear Modulation (Perez et al. 2018)
- MC: Early-Fusion Multimodal Core Model (Malinowski and Doersch, 2018)

Model performance over course of training

FiLM model variations
explicit relation
implicit comparative
implicit superlative

Multimodal core model variations
explicit relation
implicit comparative
implicit superlative

CNN-LSTM model variations
explicit relation
implicit comparative
implicit superlative

ShapeWorld simulator to generate VQA-style data (https://github.com/AlexKuhnle/ShapeWorld)