Supplementary material

to

What is needed for simple spatial language capabilities in VQA?

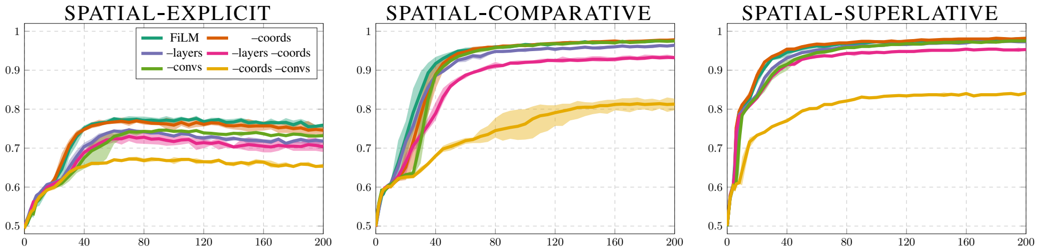

Figure 1: Accuracy performance curves over the course of training, for the FiLM model and various ablations: no coordinate map (–coords), one as opposed to four FiLM layers (–layers), fully-connected as opposed to convolutional layers (–convs).

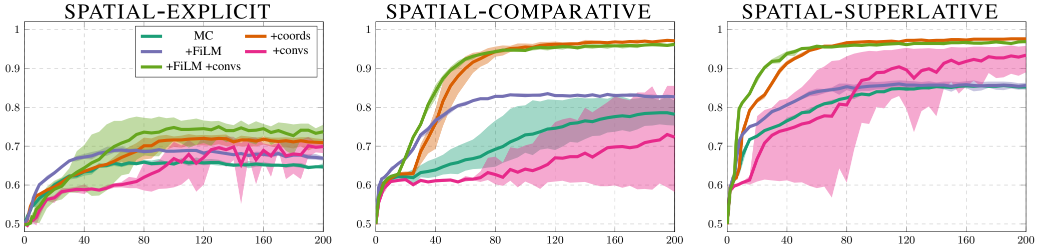

Figure 2: Accuracy performance curves over the course of training, for the MC model and various modifications: with coordinate map (+coords), FiLM fusion as opposed to concatenation (+FiLM), convolutional as opposed to fully-connected layers (+convs).

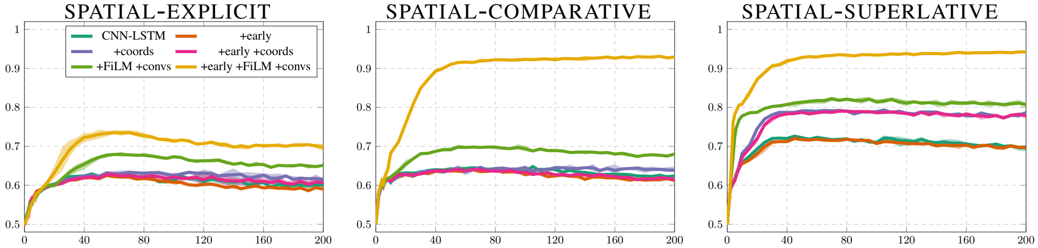

Figure 3: Accuracy performance curves over the course of training, for the CNN-LSTM model and various modifications: with early fusion (+early), with coordinate map (+coords), FiLM fusion as opposed to concatenation (+FiLM), convolutional layer (+convs).