Attention on Attention: Architectures for Visual Question Answering

Yong Jiang

Tencent AI Lab && ShanghaiTech University

Paper

Attention on Attention: Architectures for Visual Question Answering (VQA)

Jasdeep Singh
Stanford University
jasdeep@stanford.edu

Vincent Ying
Stanford University
vhying@stanford.edu

Alex Nutkiewicz
Stanford University
alexer@stanford.edu

Task: VQA

Who is wearing glasses?

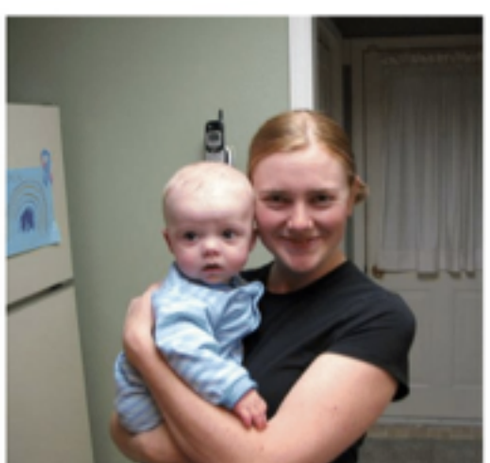
man

woman

Where is the child sitting?

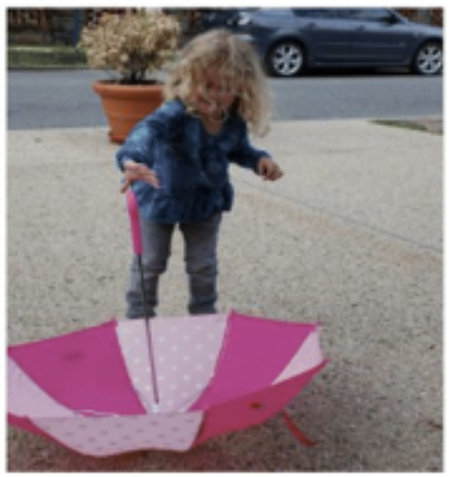
fridge

arms



Is the umbrella upside down?

yes



no

How many children are in the bed?

2

1

Network Architecture

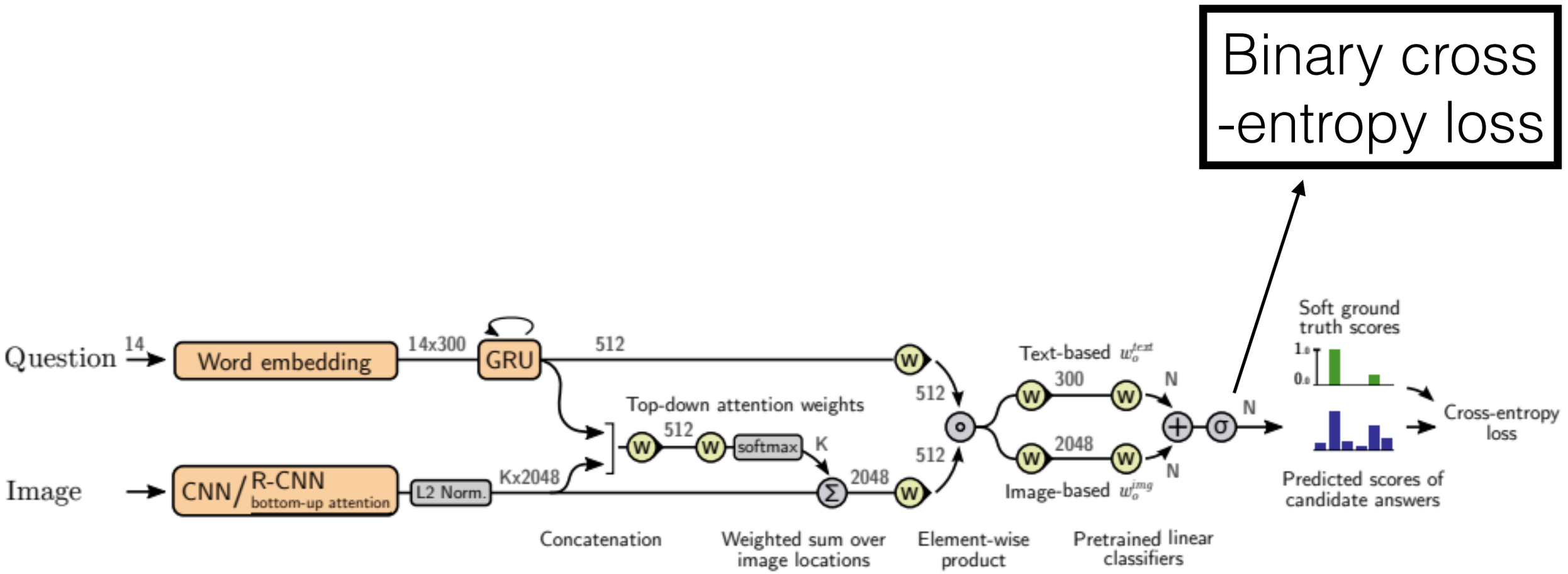
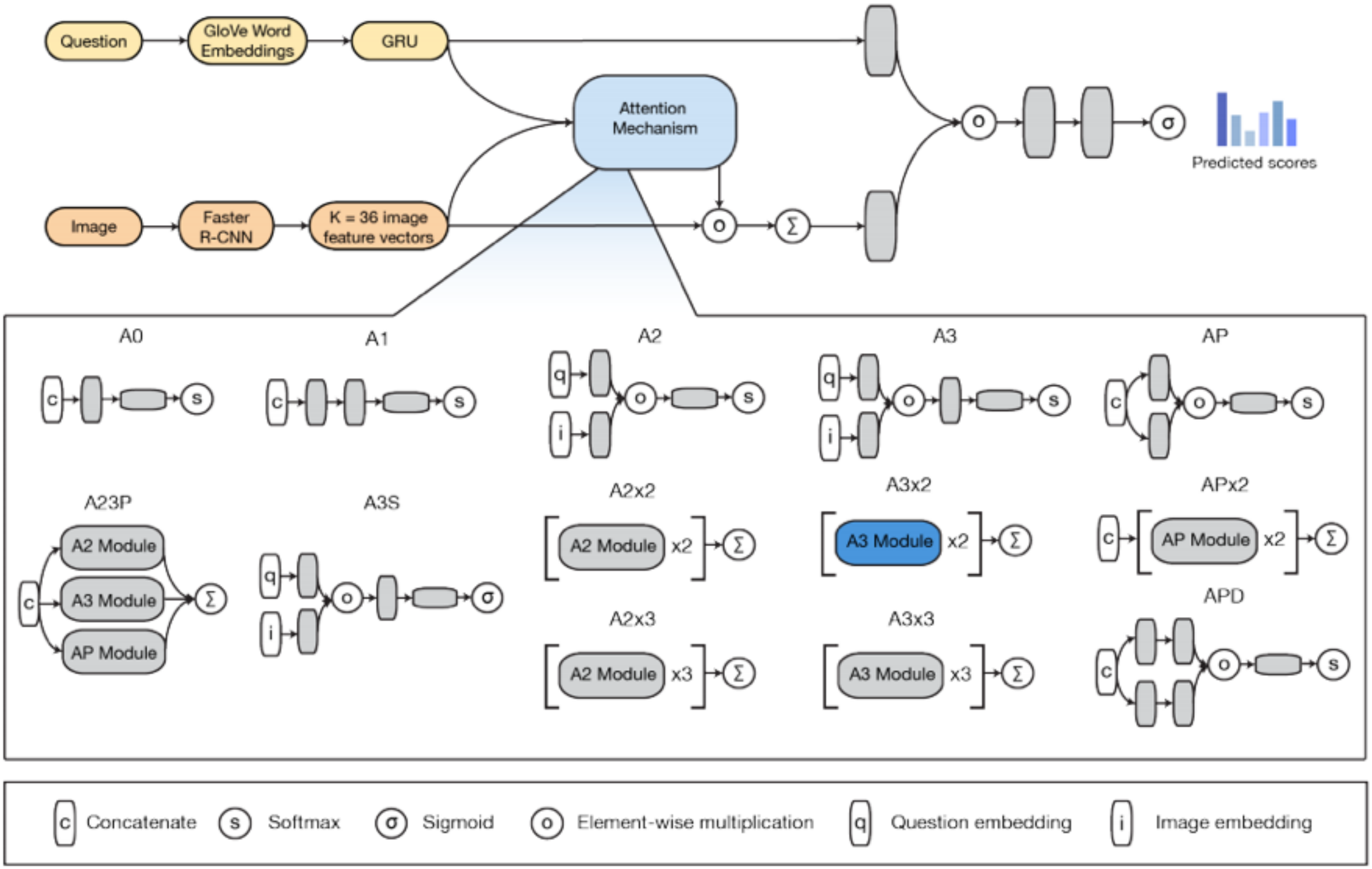


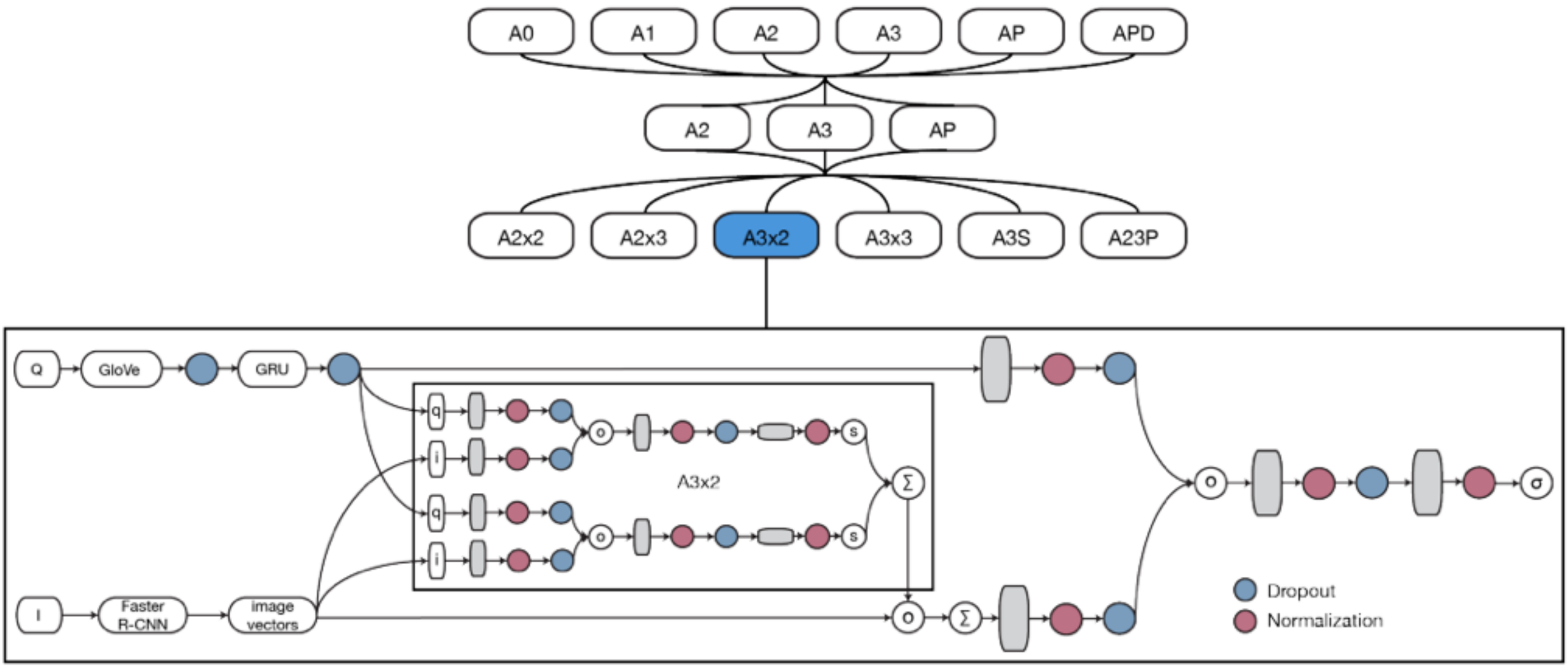
Figure 2. Overview of the proposed model. A deep neural network implements a joint embedding of the input question and image, followed by a multi-label classifier over a fixed set of candidate answers. Gray numbers indicate the dimensions of the vector representations between layers. Yellow elements use learned parameters. The elements Ⓦ represent linear layers, and Ⓦ▷ non-linear layers (gated tanh).

Winner's system in VQA competition 2017

Network Architecture



Network Searching



Hyper-parameters

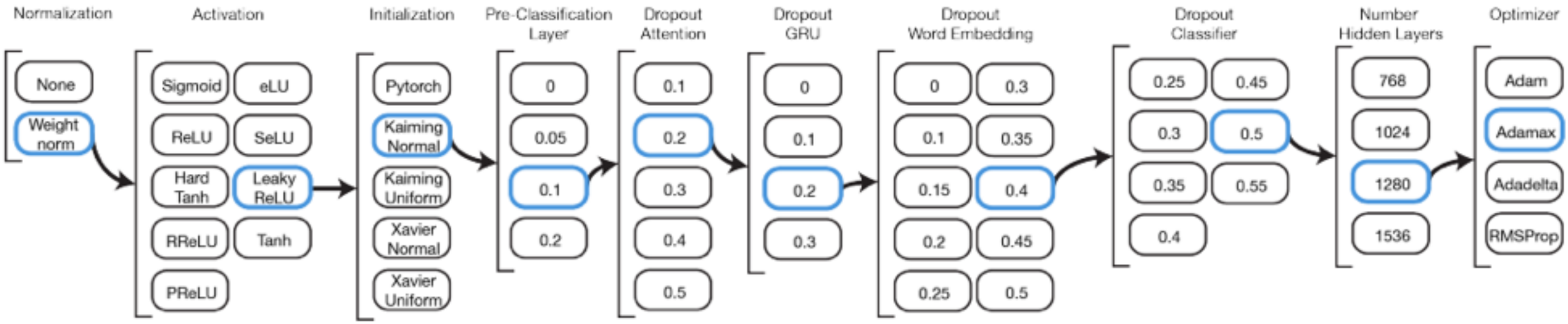


Figure 3: Hyperparameters and selected values used for experimentation. Boxes highlighted in blue had the highest performance and were selected for the final model.

Results

Table 1: Performance of Our Model vs. State-of-the-Art

MODEL	VAL PERFORMANCE SCORE
Our Model	Score **64.78** %
Teney et al. Model	Score 63.15 %