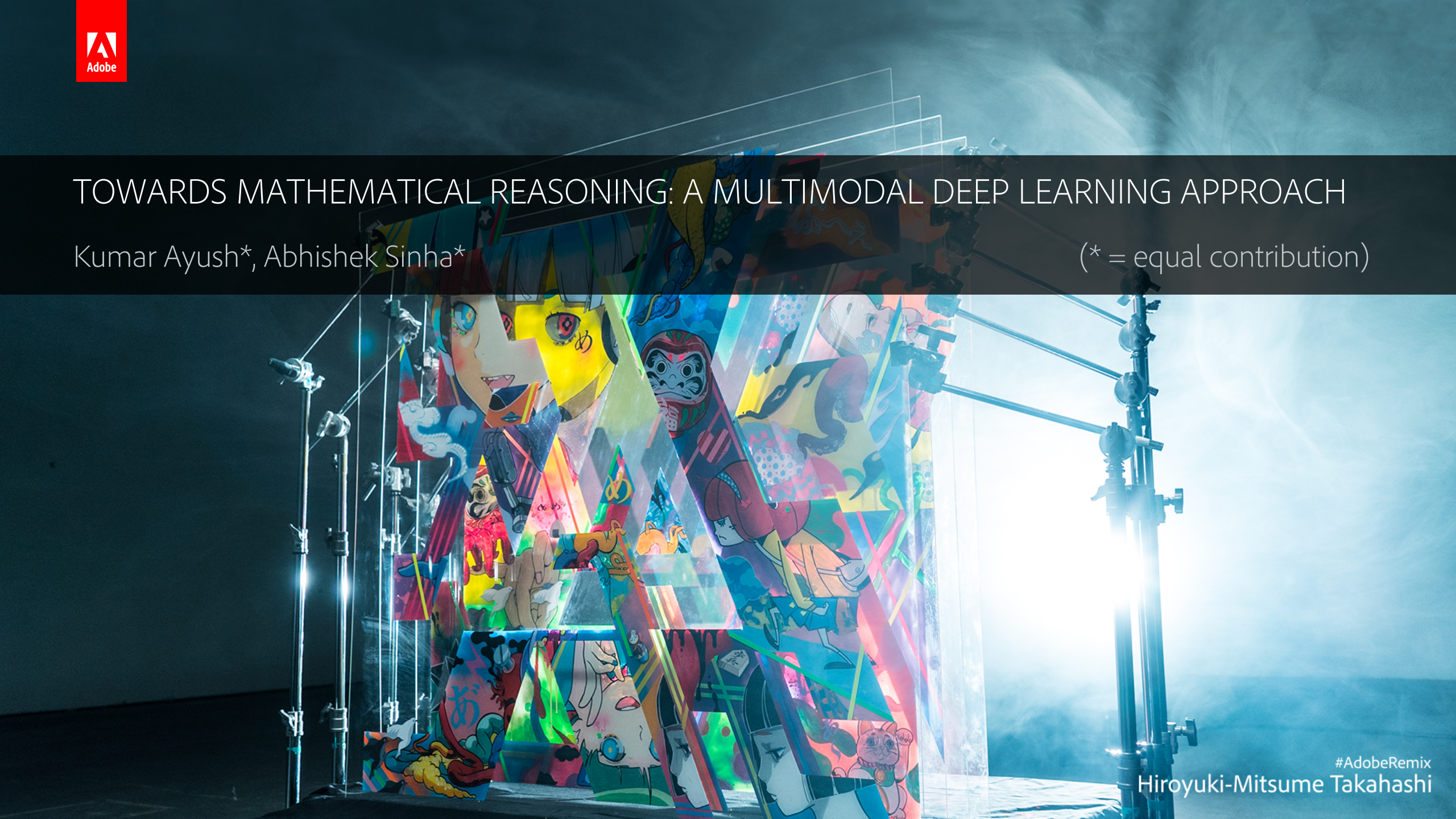
TOWARDS MATHEMATICAL REASONING: A MULTIMODAL DEEP LEARNING APPROACH

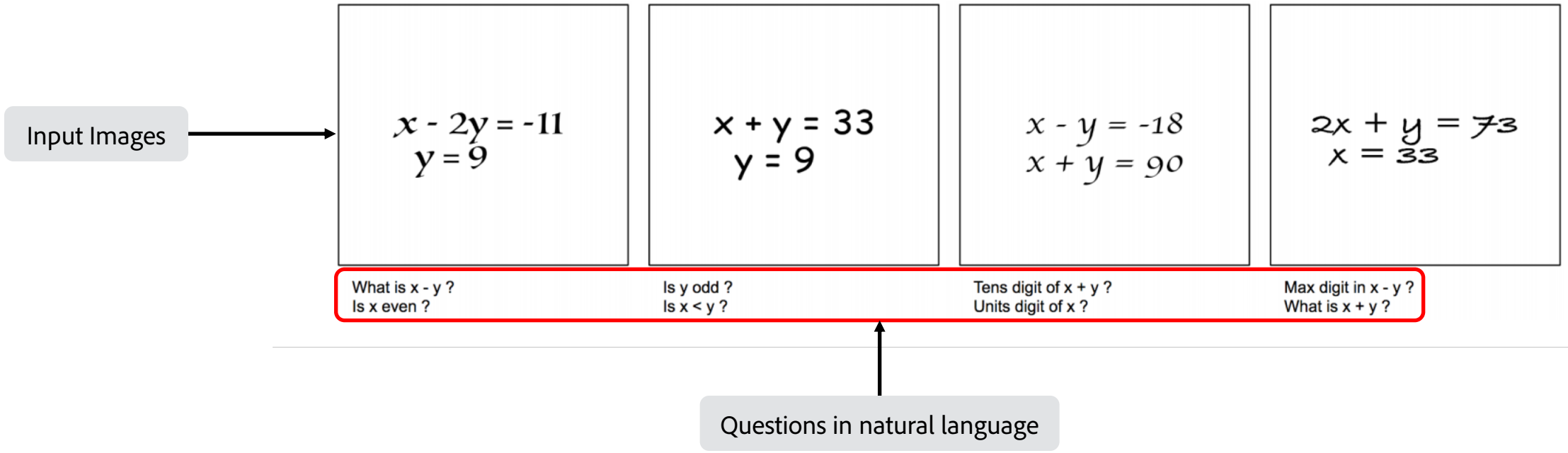
Kumar Ayush*, Abhishek Sinha*

(* = equal contribution)



Goal

- Given an **image with a simple linear algebraic equation system** and a **question in natural language** based on the variables in the equations, we propose an **end-to-end multimodal deep learning model** that produces accurate answers to the questions.



Why this problem is Interesting?

- We model this problem as **Visual Question Answering task**.

  Modelling the problem of solving simple linear equations as a VQA task makes it interesting as the system now **requires three kinds of understanding**:

  a) **visual understanding** to recognize digits, variables, operators and equal sign,

  b) **conceptual understanding** of the symbolic meanings of coefficients, constants, variables, operators and equality and

  c) **high level understanding** of the interaction between the image and the questions in order to accurately answer them

Why this problem is Interesting?

- To solve this problem,

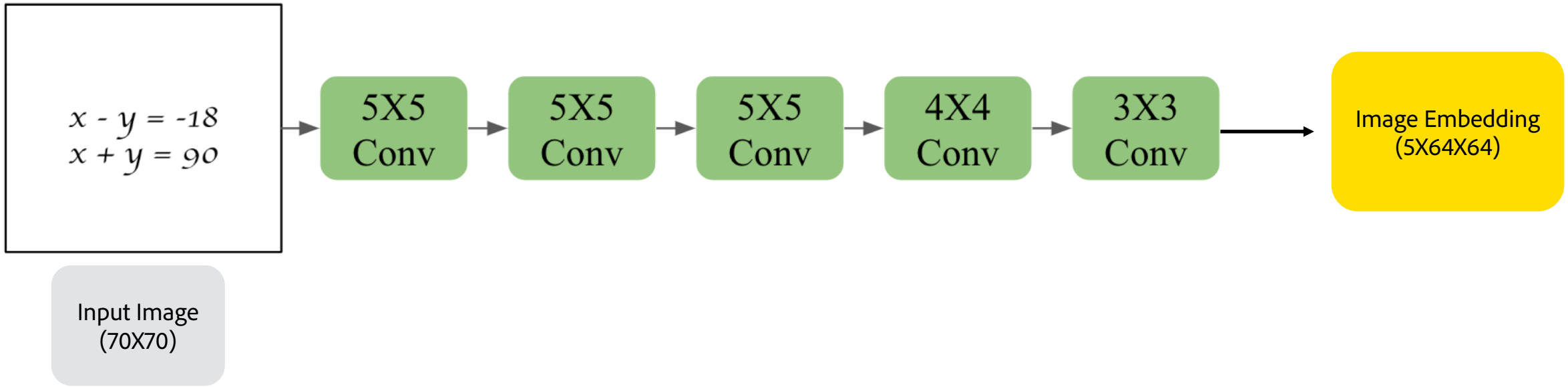
  a) **precise image and text models** are required and,

  b) most importantly, **high level interactions between these two modalities have to be carefully encoded** into the model in order to provide the correct answer.

Model Overview

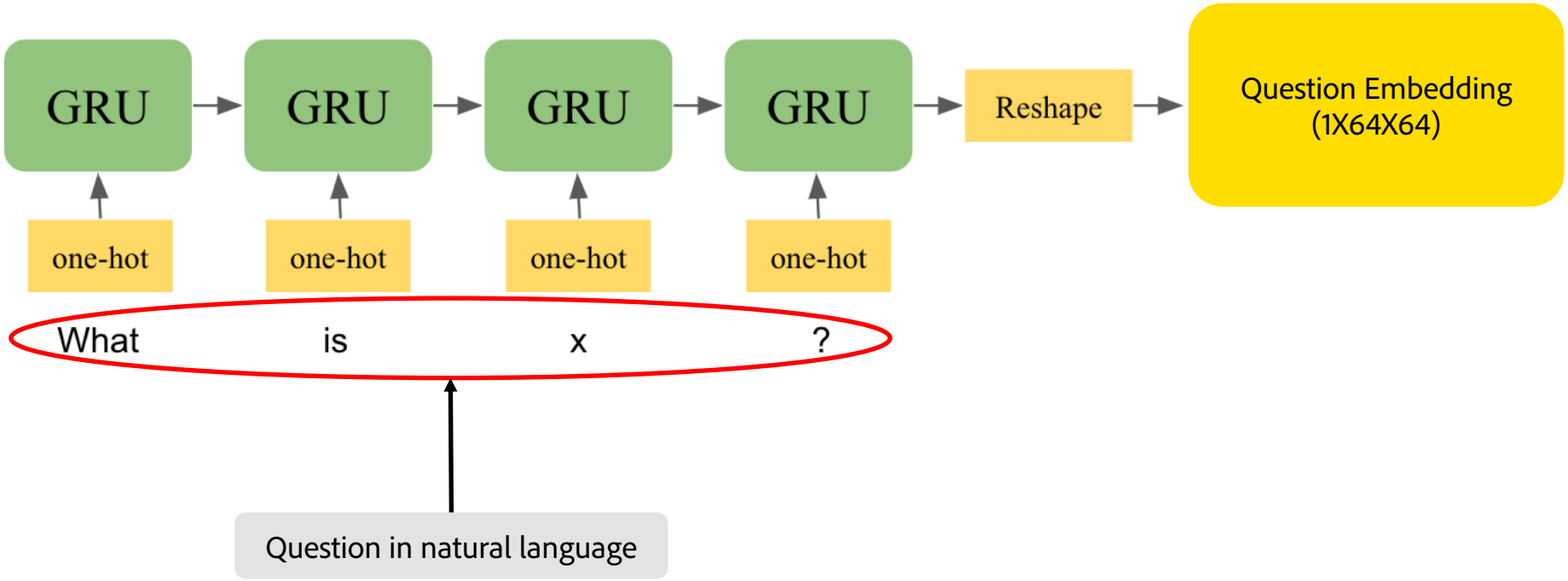
Model Overview

Image Embedding



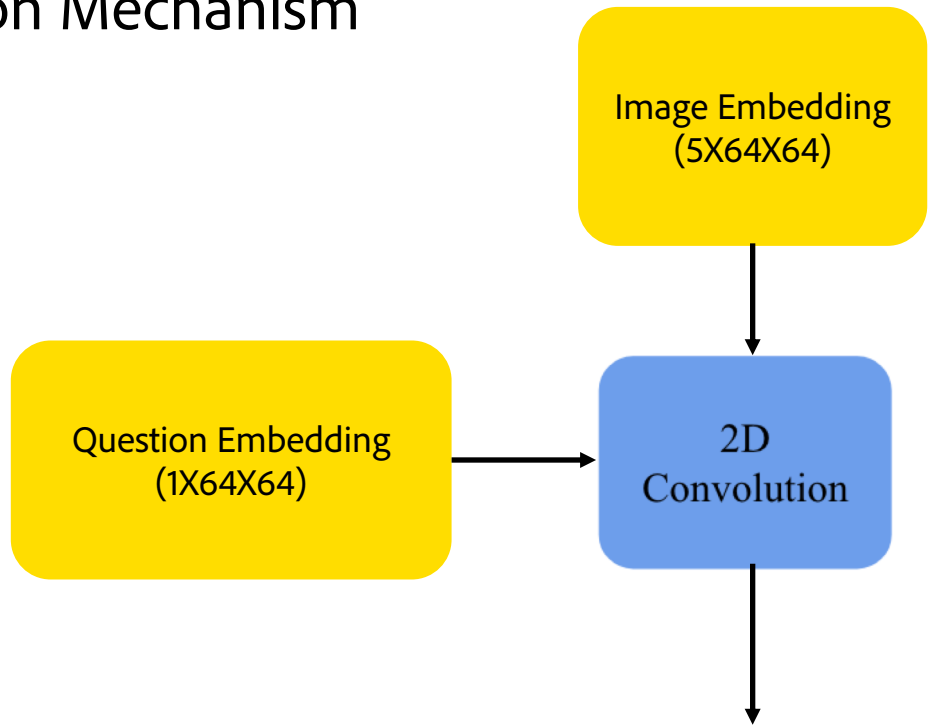
Model Overview

Question Embedding



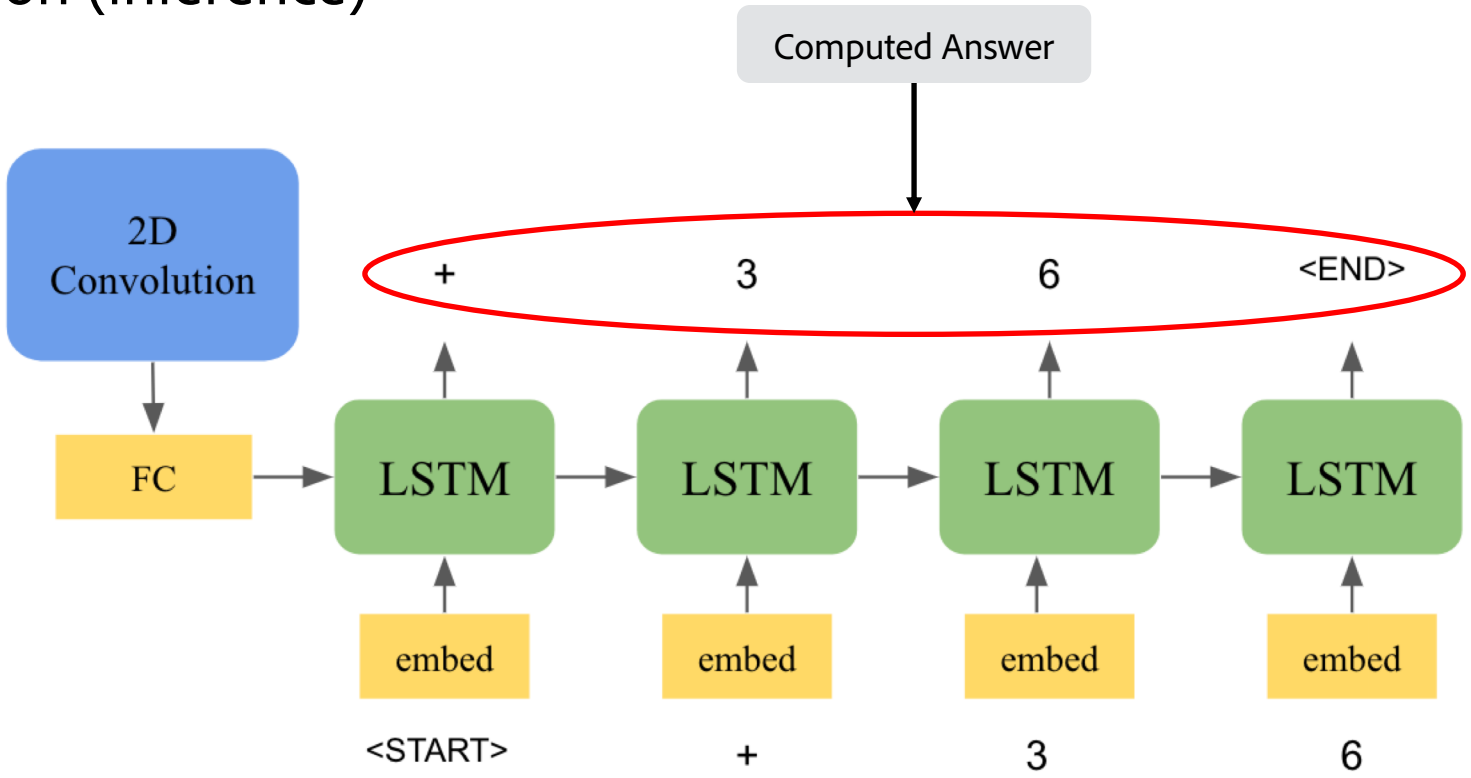
Model Overview

Multimodal Fusion Mechanism

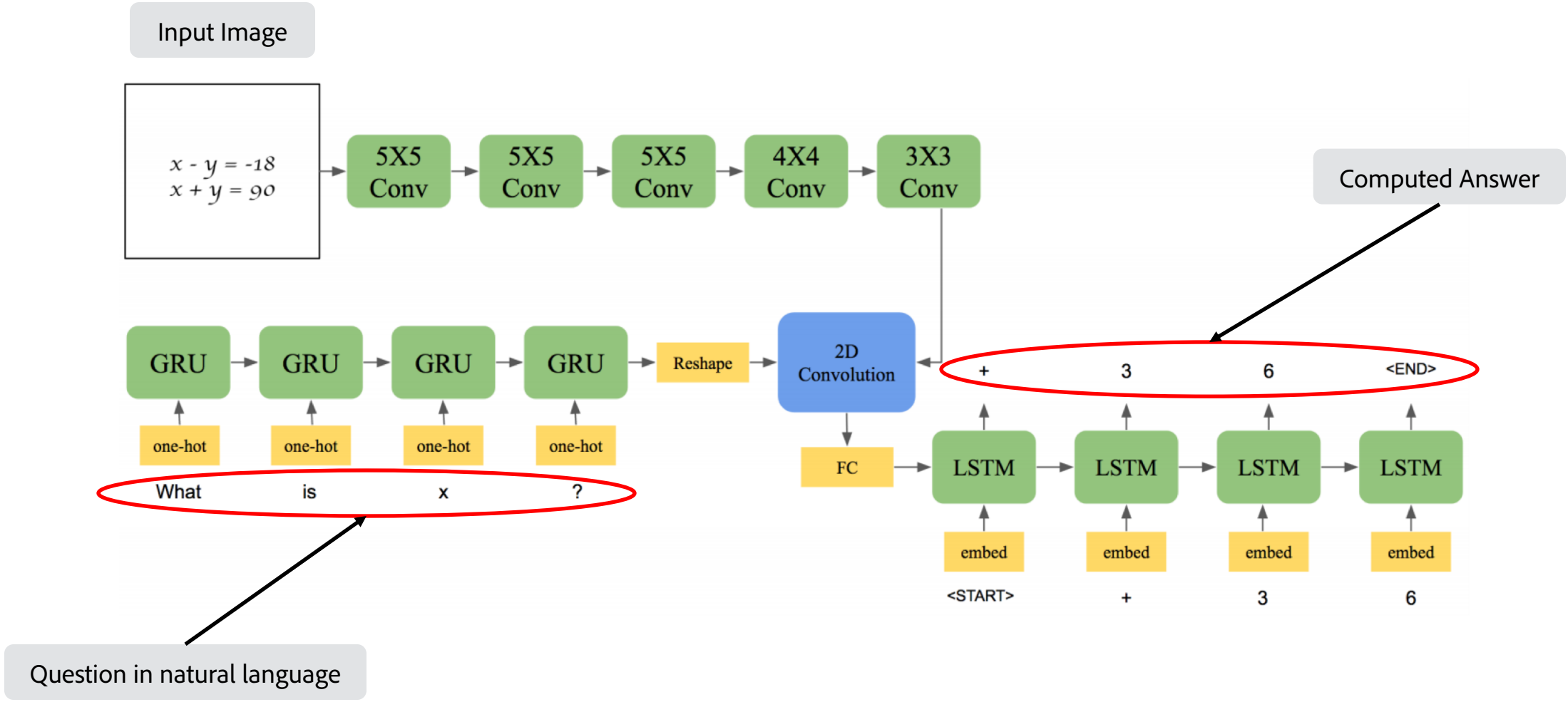


Model Overview

Answer Computation (Inference)



Model Overview (Complete Model)

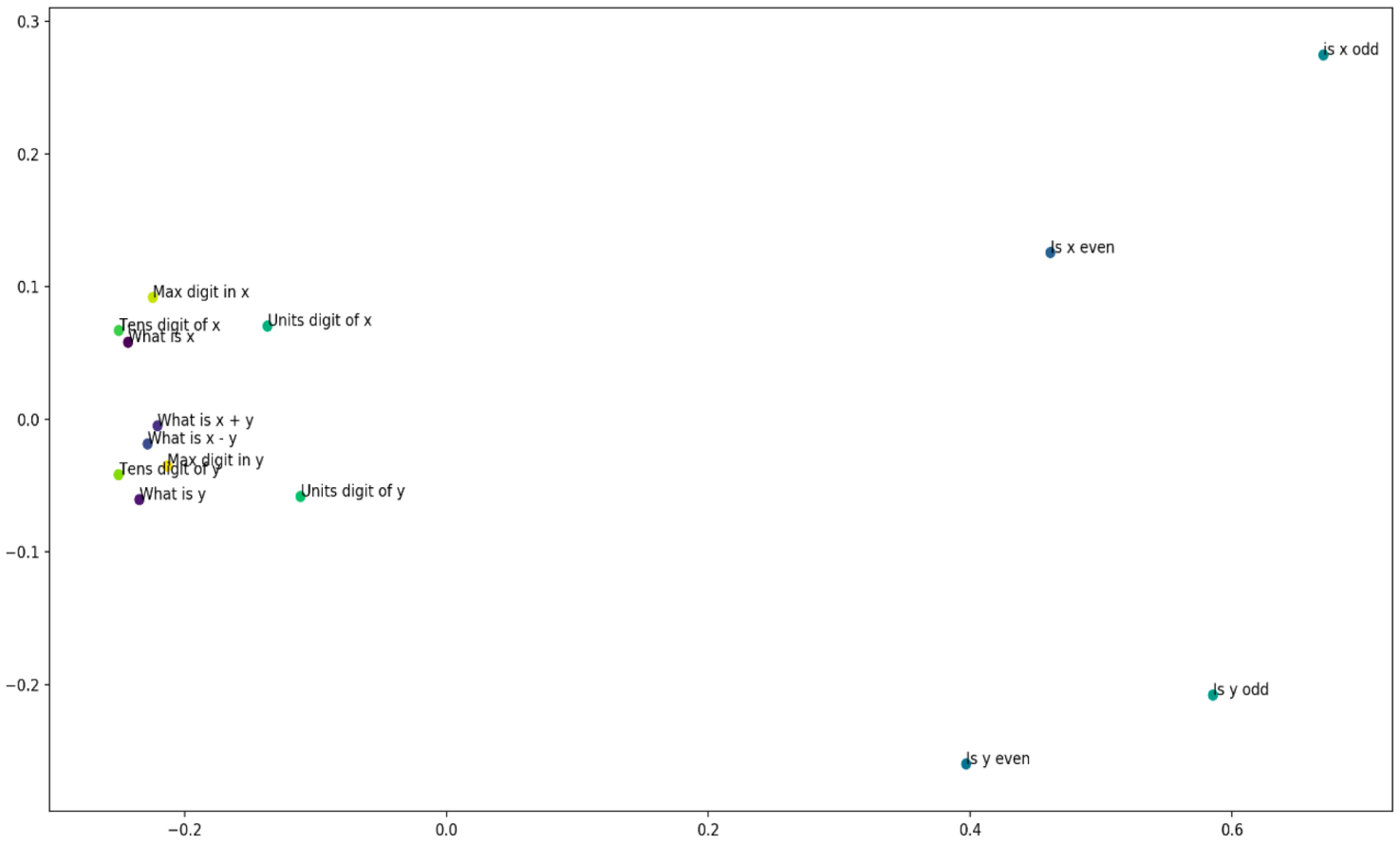


Experimental Results

Table 1. Comparison with baseline fusion schemes on test dataset. The first three columns represent percentage accuracy and the final column represents the mean error in case of numerical outputs.

	T/F	Num	All	Mean Error
Ours (LTR)	85.7	**80.3**	82.1	**5.9**
Ours (RTL)	**97**	79.7	86	11.2
Concatenation	67.6	34.9	40.8	–
Mutliplication	50.3	3.1	20.3	–
Neural VQA	50.5	3.2	20.1	–

Visualization of question embeddings using **Principal Component Analysis**



- **Questions related to a particular expression form a cluster**. For example, questions pertaining to only x like Ten's digit of x?, Max digit in x?, What is x?, etc., are lying in a distinct cluster.
- Questions with **True/False answers** are **lying far away** from questions with **numerical answers.**
- Network is able to **understand the difference between questions with T/F and numerical answers**.
- Amongst questions with numerical answers, the network is able to **learn the difference between different expressions** like only x, only y, x + y, etc.

Main Contributions

- To the best of our knowledge, this is the first work which aims at **solving simple linear equations from images in an end-to-end fashion**.

- We model this problem as a **Visual Question Answering (VQA) task** wherein we ask questions which are based on the final numerical values of the variables along with other complex questions. **We create an end-to end model for the task**.

- We compare our model with various baselines and show that our model performs better than them.

Published at 25th IEEE International Conference on Image Processing (ICIP) 2018.

TOWARDS MATHEMATICAL REASONING: A MULTIMODAL DEEP LEARNING APPROACH

Kumar Ayush and Abhishek Sinha**

Adobe Systems Inc, Noida, India

ABSTRACT

This paper presents a new direction for the visual question answering task. Given an image with a simple linear algebraic equation system and a question in natural language based on the variables in the equations, we propose an end-to-end deep learning model that produces accurate answers to questions pertaining to the value of the variables and other related questions. Modeling the problem of solving simple linear equations as a VQA task makes it interesting as the system now requires three kinds of understanding a) visual understanding to recognize digits, variables, operators and equal sign b) conceptual understanding of the symbolic meanings of coefficients, constants, variables, operators and equality and c) high level understanding of the interaction between the image and the questions in order to accurately answer them. We also create an open-source dataset for the same and compare the performance of our model with different baselines.

Index Terms— Linear equations, Deep Learning, Visual Question Answering, Mathematical Reasoning

techniques are usually used in computers to solve such linear equations. However, in case of very simple linear equations like (1) and (2)

$$2x + y = 6 \quad (1)$$

$$y = 4 \quad (2)$$

humans with symbolic understanding usually follow simple substitution steps and moving variables/constants to either side of the equality sign to solve them instead of following the aforementioned numerical methods. This paper is specifically targeted towards solving simple linear algebraic equations from images and answering questions based on the equations using deep learning.

Deep learning models have performed remarkably well in several domains such as computer vision [7, 8], natural language processing [9, 10] and speech recognition [11]. These models are able to achieve high accuracy in various tasks and hence their recent popularity. Richard Evans et al. [12] have proposed a Differentiable Inductive Logic framework, a model hybridized with neural networks, and demonstrated its working on induction tasks such as *less than* where the model

Adobe


Adobe