Siamese network for binary VQA

Presenter: Sneha Mehta

May 4, 2016

ECE 6554, Advanced Computer Vision, Spring 2016

Virginia Polytechnic Institute and State University

Blacksburg VA

Instructor: Dr. Devi Parikh

Acknowledgements

- Dr. Devi Parikh and Dr. Dhruv Batra

- Graduate students, CVMLP lab, Yash Goyal and Jiasen Lu

Outline

- Background and Motivation
- Problem statement
- Approach
- Results
- Challenges
- Conclusions and Future work

Background and Motivation

- Binary Visual Question answering
  - Visual verification of concepts inquired in the question

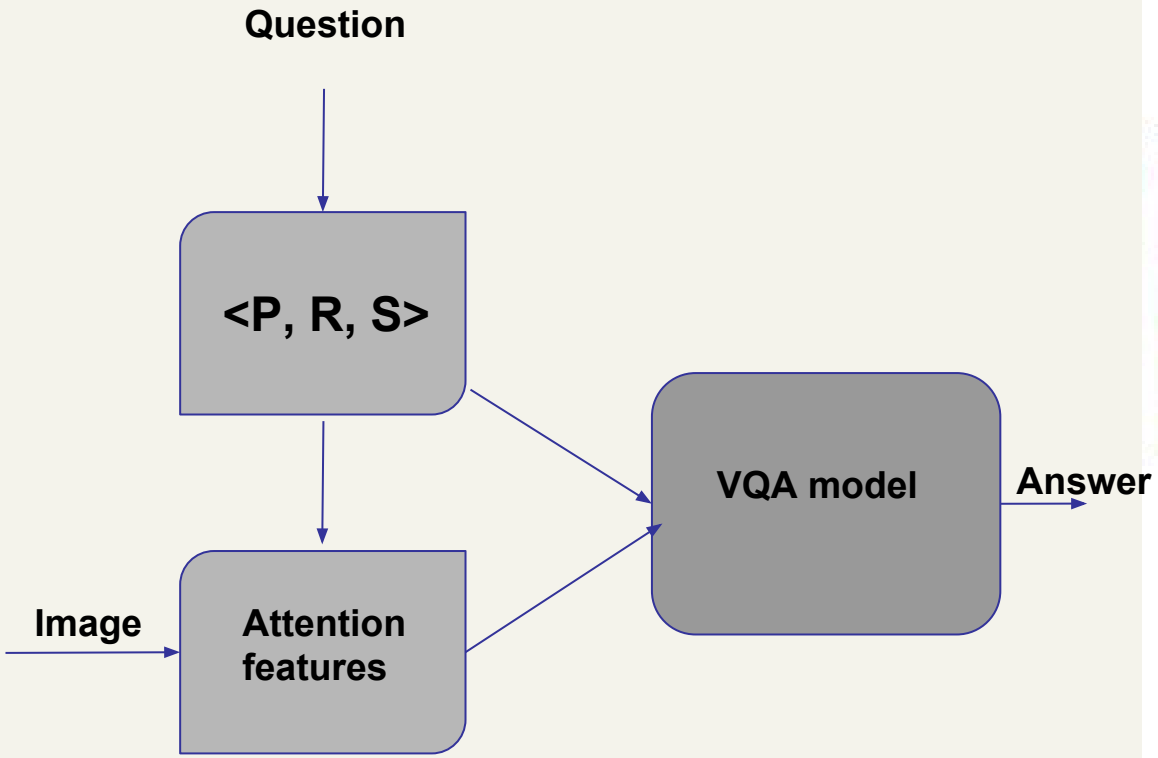
    [Yin Yang: Balancing and Answering binary visual questions]

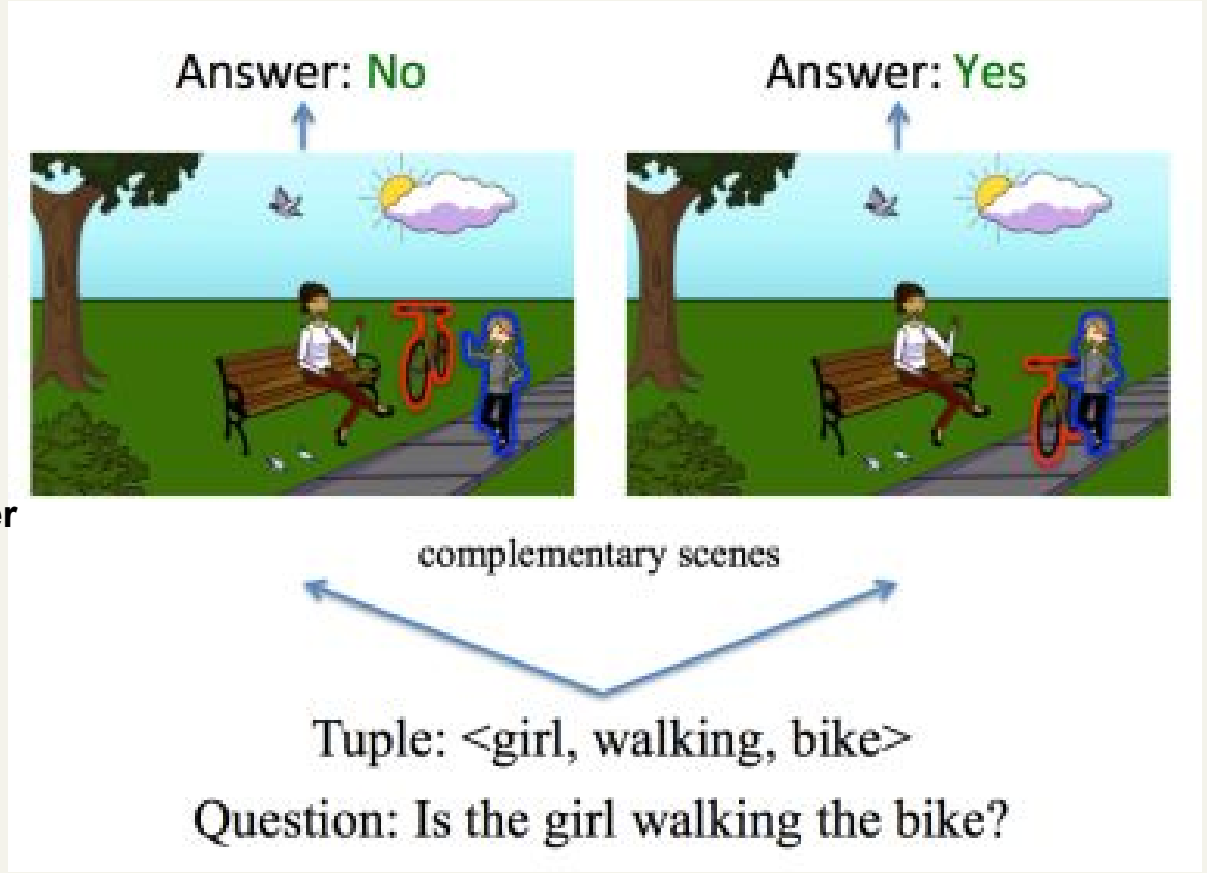
- Strong language priors leads to superficial performance

- Use abstract scenes instead of real scenes
  - focus on the high-level semantics of the VQA task as opposed to the low-level recognition problems
  - Easy to balance the dataset such that language priors are controlled and vision is tested

Background and Motivation (contd...)

- Collect a balanced dataset

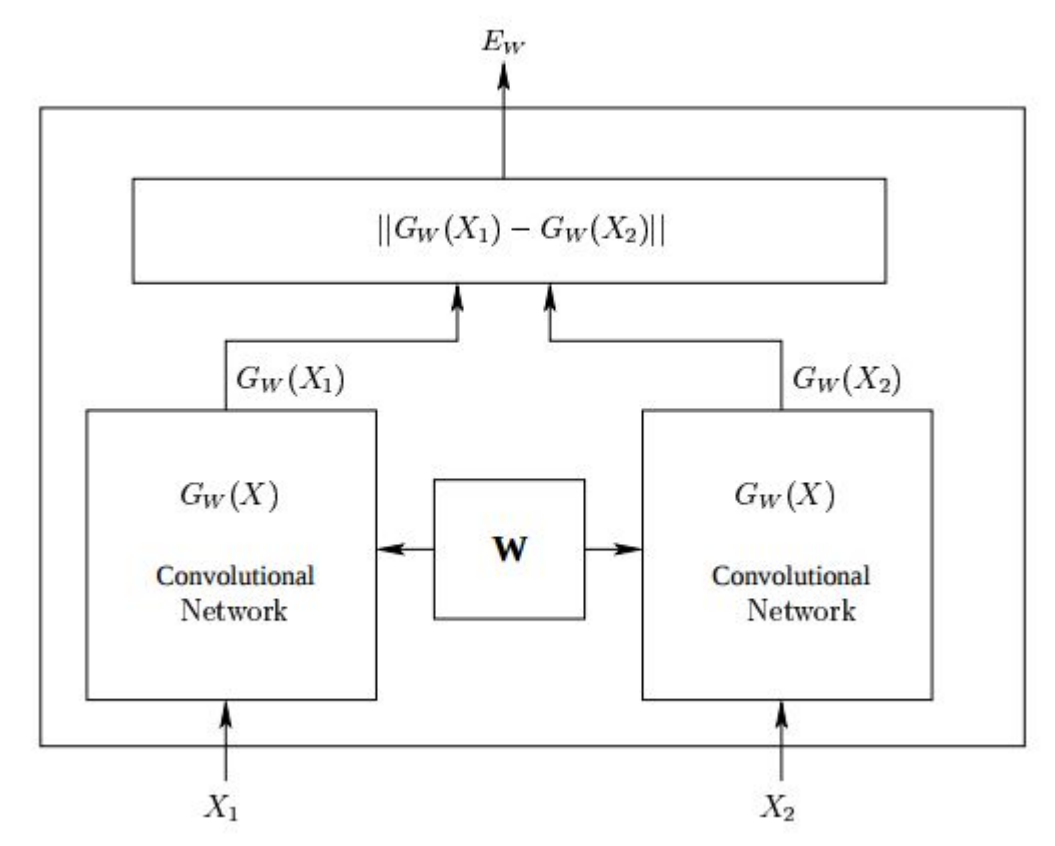




- This model trained on balanced dataset performs best on balanced dataset.

Background and Motivation (contd.)

- Can we exploit the pairwise images in the database to train a model
- Can a model be trained to learn fine grained changes in an image that changes the answer from yes(no) to no(yes)?



Siamese network

Image credit: Learning a Similarity Metric Discriminatively, with Application to Face Verification, CVPR ‘05

Goals (problem statement)

- Create a siamese network model for binary VQA to see if it improves performance
- Compare against the non-siamese baseline

Approach: Model Architecture

VQA architecture

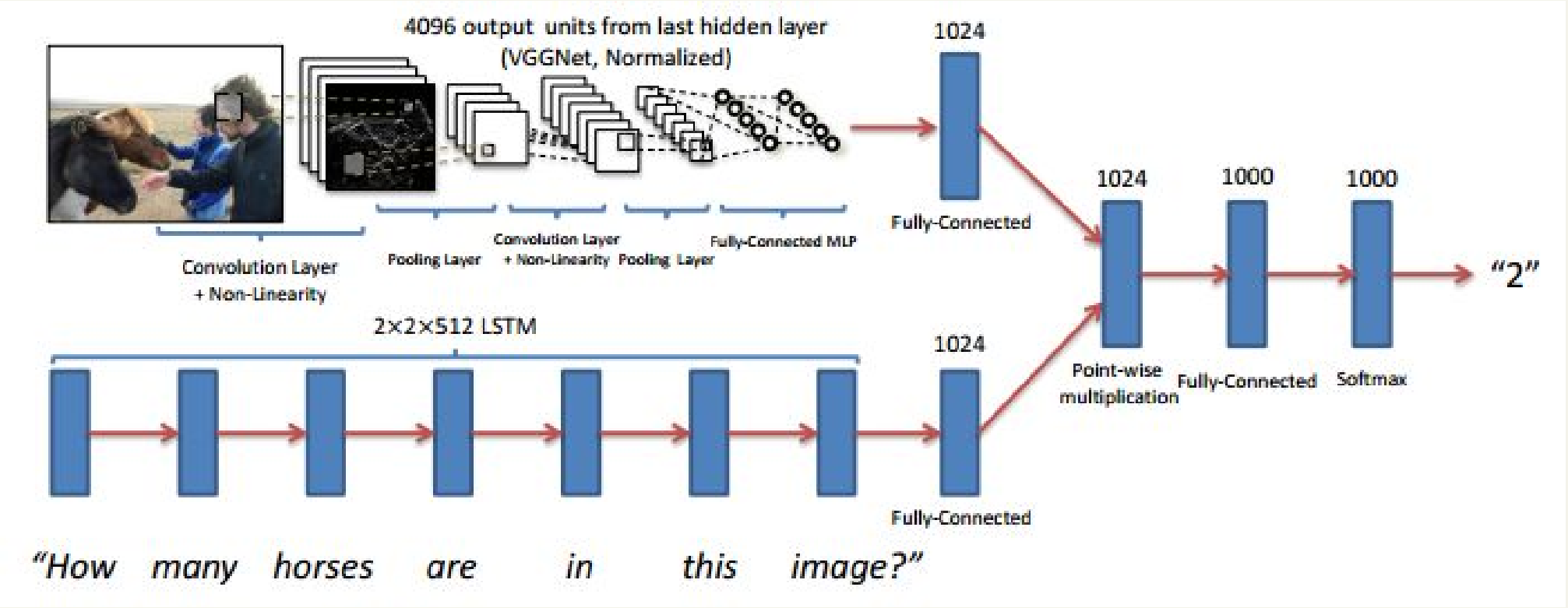
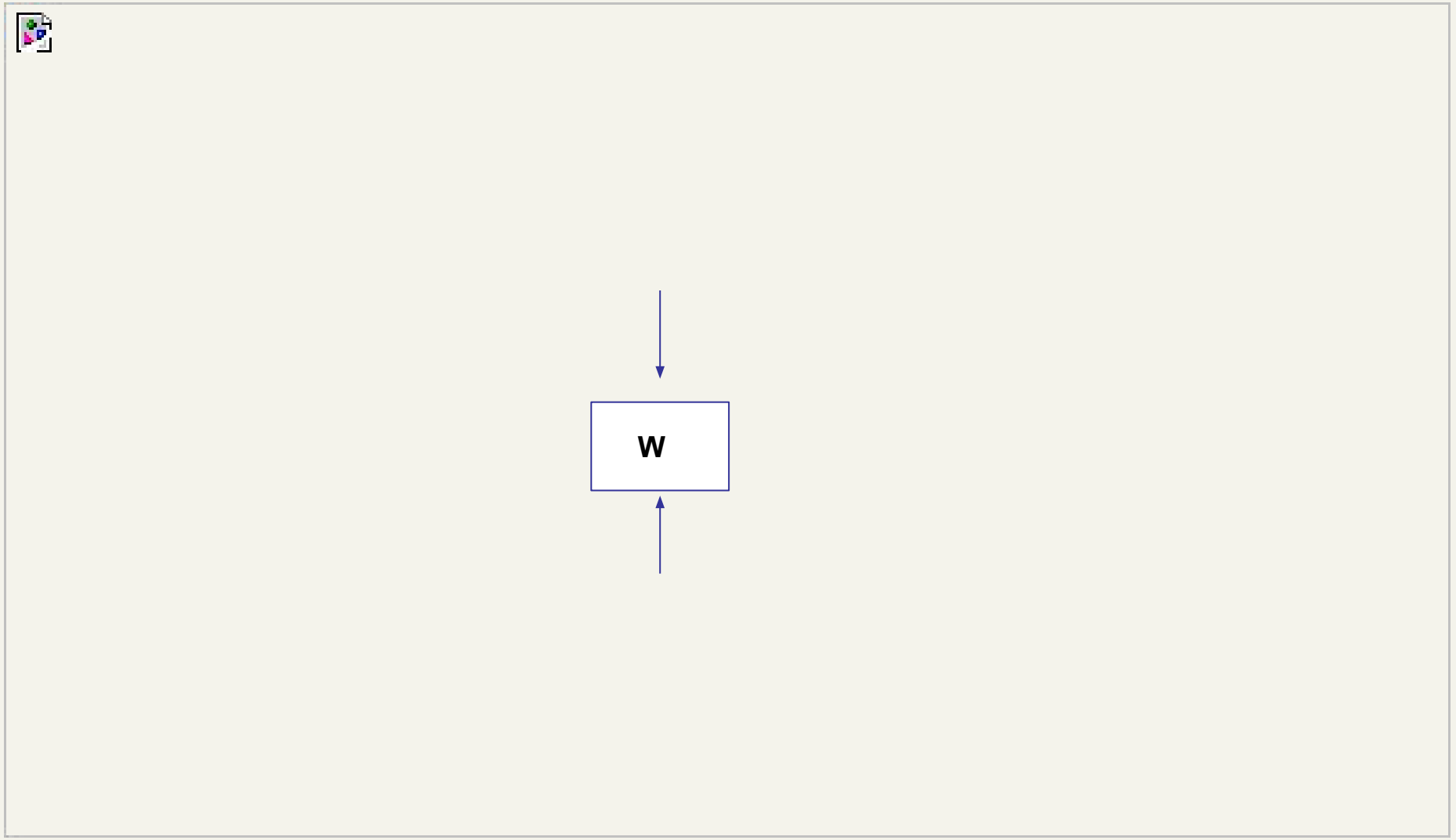


Image credit: VQA: Visual Question Answering, ICCV 2015

Approach: Binary VQA siamese architecture (training)



Loss function

- $$L = \lambda * (\mu * ce_1 + (1 - \mu) * ce_2) + (1 - \lambda) * h$$
  $$\mathrm{ce}_1, ce_2 = cross\ entropy\ losses\ from\ both\ networks$$
  $$H = \max(0, margin\ - (p_{yes_1} - p_{yes_2}))$$
  $$p_{yes_1} := probability\ of\ 'yes'\ class\ from\ network\ 1$$
  $$p_{yes_2} := probability\ of\ 'no'\ class\ from\ network\ 2$$

Technologies used:

- VQA_LSTM_CNN (https://github.com/VT-vision-lab/VQA_LSTM_CNN)
- VQA (https://github.com/VT-vision-lab/VQA)
- Torch

Evaluation and Results

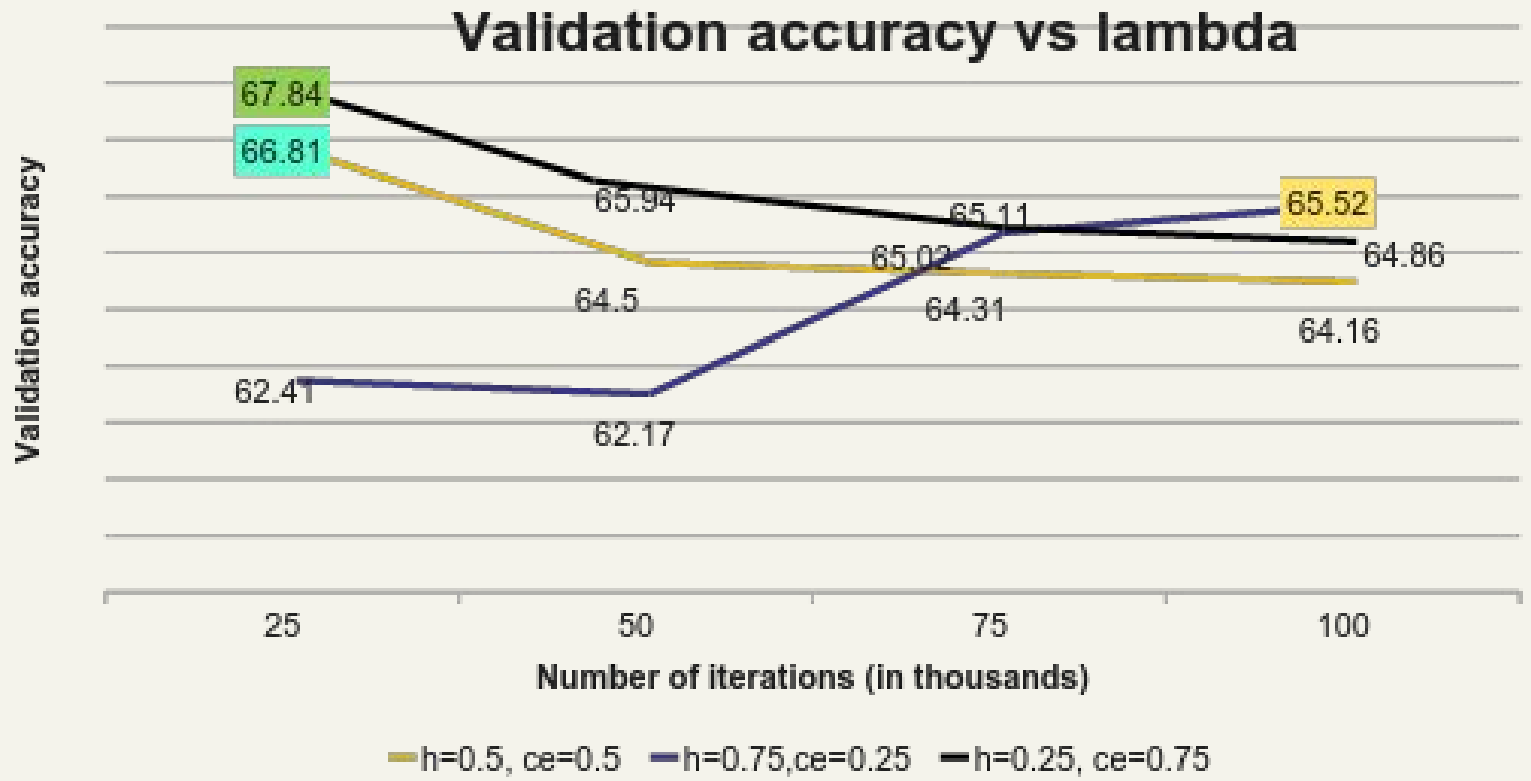
- Dataset: VQA balanced abstract scenes dataset

  [Yin Yang: Balancing and Answering binary visual questions]

- Training
  - 20,578 abstract scenes (10289 scene pairs)
  - 10289 binary questions

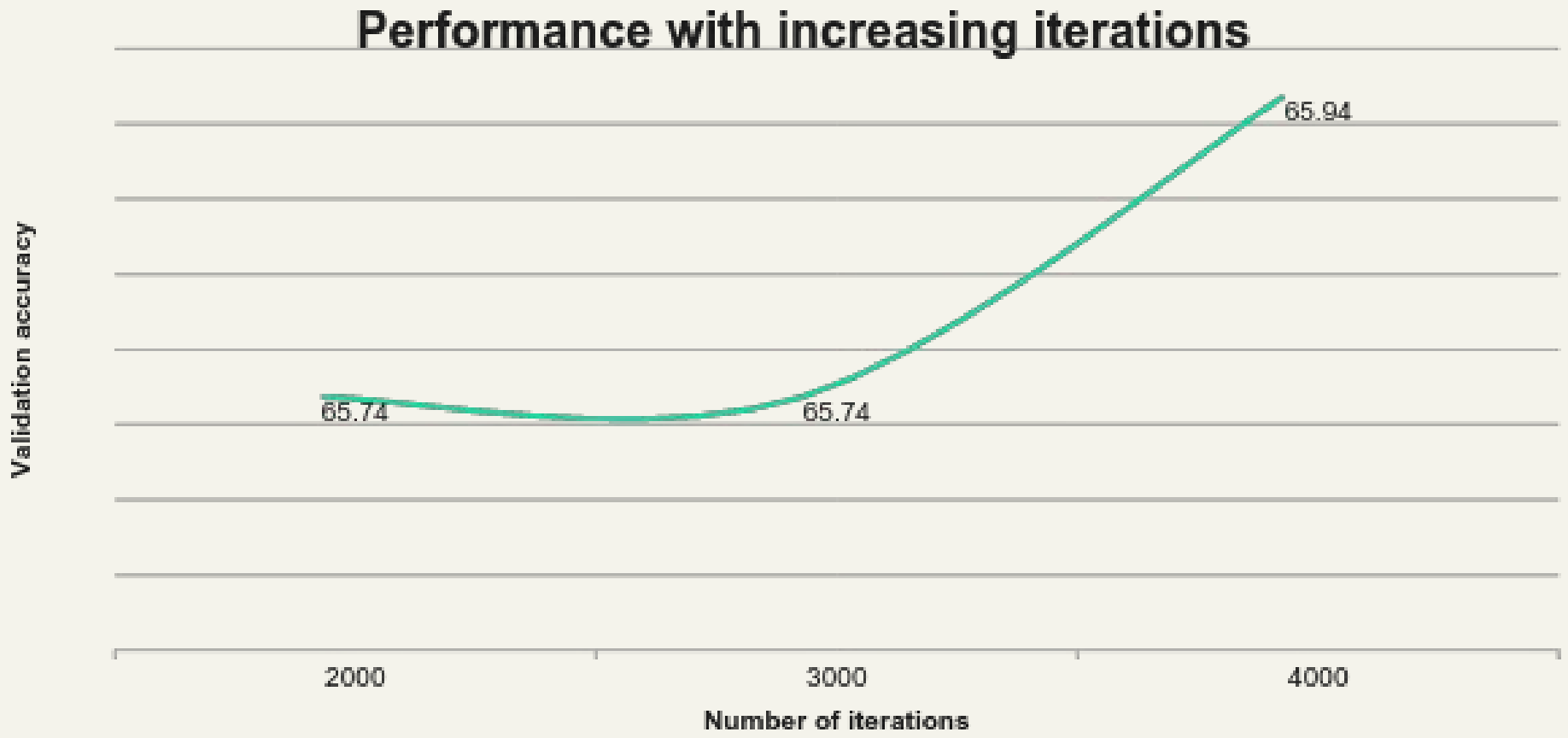
- Validation
  - 13,260 question, scene pairs

Evaluation and Result

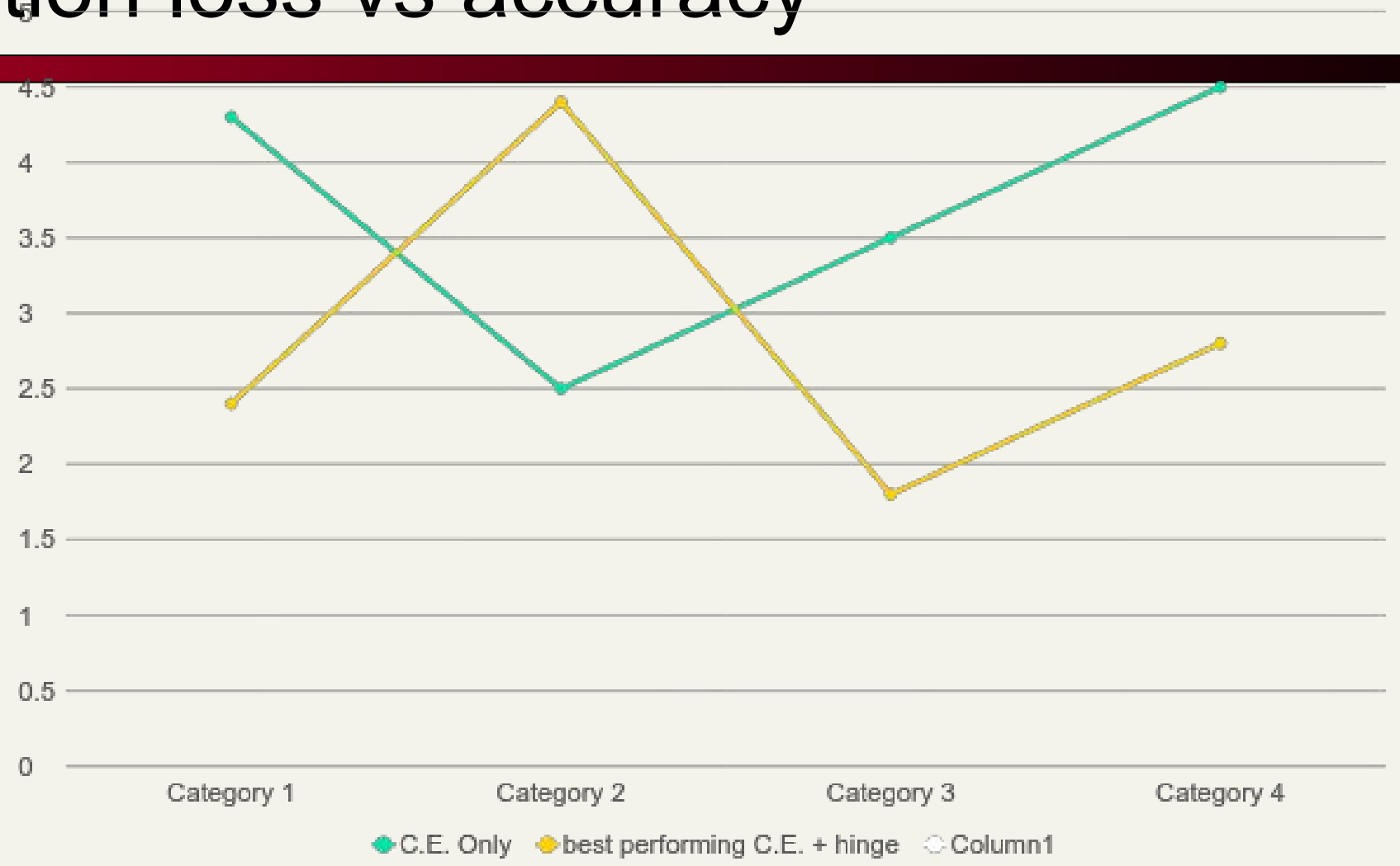


- batch size = 500
- Optimization method = rmsprop
- Learning rate = 0.0003

Evalutation and Results: C.E. only performance



Validation loss vs accuracy



Challenges

- Problems with the data
- Backpropogation through siamese networks
- Tuning the hyperparameters

Conclusion and Future work

- Conclusion
  - Siamese network model indeed performs better than a non-siamese model
- Future work
  - Tune the hyperparameters to beat the state of the art
- Find code at: https://github.com/sumehta/siamese_network_vqa

Thank You!