CS 224N - Final Project

Manan Rai

Generalized Textual and Visual Question Answering using Combined Image-Context Embeddings

YouTube link: https://youtu.be/9bHosmZGbP4

Presented below are the slides used in the video.

CS 224N - Final Project

Manan Rai

Generalized Textual and Visual Question Answering using Combined Image-Context Embeddings

CS 224N - Final Project

Manan Rai

Question Answering

- Given some input and a natural language question, provide a natural language answer to the question.
- First Step: Understand the input as well as the question, i.e. extract features from both.
- Second Step: Relate these features with one another.
- Final Step: Formulate an Answer.

“
The sun shines in the morning. The moon shines at night.
”

CS 224N - Final Project

Manan Rai

When does the moon shine?
At night.

Textual Question Answering

- Input is text, also referred to as context.
- Therefore, this task is also called context-based question answering.
- Dataset: SQuAD 2.0 Dataset.
- Multiple Approaches: Abstractive and Extractive.
- Sample Question: When did the war start? Who scored the game-winning touchdown?
- Loss: Cross-Entropy against the start and end positions.

CS 224N - Final Project

Manan Rai

Is it day or night?
Night.

Visual Question Answering

- Input is an image that the questions pertain to.
- Dataset: VQA 2.0 Dataset.
- Sample questions: How many objects are there in the picture? What are the people doing here? Is it day or night?
- Loss: Cross-Entropy against the ground truth answer.

CS 224N - Final Project

Manan Rai

The Way we Absorb Knowledge

- Consider Wikipedia articles. Note the amount of text, and the amount of pictures.
- This is because images help visualize and bring to life the words we read. Diagrams and graphs help disambiguate difficult-to-comprehend information that words don't do justice to.
- Yet research in question answering treats these two disparate formats of inputs completely differently – therefore adding an unnatural component to how machines approach question answering.

Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools

What links here
Related changes
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export

Create a book
Download as PDF
Printable version

In other projects

Wikimedia Commons
Wikiquote

Languages

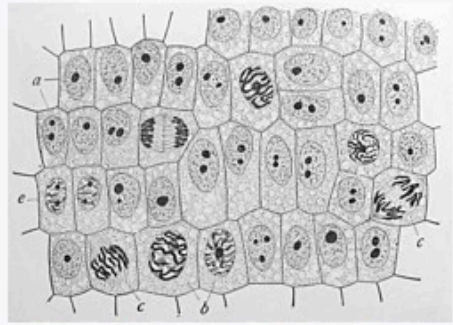
Deutsch
Español
Français
한국어
Italiano
Русский
Tagalog
Tiếng Việt
中文

smallest unit of life. Cells are often called the "building blocks of life". The study of cells is called cell biology or cellular biology.

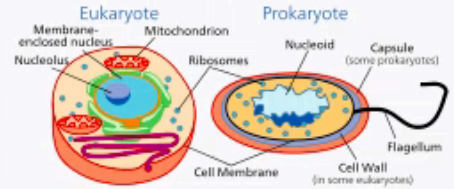
Cells consist of cytoplasm enclosed within a membrane, which contains many biomolecules such as proteins and nucleic acids.[2] Organisms can be classified as unicellular (consisting of a single cell; including bacteria) or multicellular (including plants and animals).[3] While the number of cells in plants and animals varies from species to species, humans contain more than 10 trillion (10^{13}) cells.[4][*clarification needed*] Most plant and animal cells are visible only under a microscope, with dimensions between 1 and 100 micrometres.[5]

Cells were discovered by Robert Hooke in 1665, who named them for their resemblance to cells inhabited by Christian monks in a monastery.[6][7] Cell theory, first developed in 1839 by Matthias Jakob Schleiden and Theodor Schwann, states that all organisms are composed of one or more cells, that cells are the fundamental unit of structure and function in all living organisms, and that all cells come from pre-existing cells.[8] Cells emerged on Earth at least 3.5 billion years ago.[9][10][11]

Contents [hide]



Onion (*Allium cepa*) root cells in different phases of the cell cycle (drawn by E. B. Wilson, 1900)



A eukaryotic cell (left) and prokaryotic cell (right)

Identifiers	
MeSH	D002477
TH	H1.00.01.0.00001
FMA	68646

Anatomical terminology

[edit on Wikidata]

Structure of an animal cell

CS 224N - Final Project

Manan Rai

Textual and Visual Question Answering

- We propose the task of generalized, input-format-independent question answering, where a single, generalized model can answer questions based on *both* textual and visual inputs.
- An important part of this involves being able to capture features from both input formats and expressing them in a combined format that is accessible to neural networks.
- This work presents a technique to build these embeddings, as well as proposes models that can process them and answer questions.

CS 224N - Final Project

Manan Rai

Embeddings

- Machine Comprehension is an important research area, and one that is far from fulfilled.
- This means that our model(s) can't (yet) understand the input as is – and some initial processing is required to get them to be able to understand the input well enough to answer questions about it.
- The embeddings capture the salient features of the image in an easily representable and computable format.

CS 224N - Final Project

Manan Rai

Embeddings: Current State

- Currently, there are several pre-trained as well as non-pre-computed embeddings for both input images and text, but none that are generalizable.

CS 224N - Final Project

Manan Rai

Embeddings: Defining Generalizability

- Currently, there are several pre-trained as well as non-pre-computed embeddings for both input images and text, but none that are generalizable.

- Here, 'generalizability' refers to the ability to capture salient features of both input formats enough that a common model can process the produced embedding with good results on both visual and textual question answering tasks.

CS 224N - Final Project

Manan Rai

Embeddings: Closer Look

- Currently, there are several pre-trained as well as non-pre-computed embeddings for both input images and text, but none that are generalizable.
- Here, ‘generalizability’ refers to the ability to capture salient features of both input formats enough that a common model can process the produced embedding with good results on both visual and textual question answering tasks.
- Let us look at how this is done for each of the input formats.

CS 224N - Final Project

Manan Rai

Context Embeddings

- Textual inputs usually utilize word vectors that place semantically and/or linguistically similar words close to one another in the word vector space.
- Pre-trained embeddings include Global Vectors (GloVe) that are fixed-length word and character embeddings.
- Neural Networks can be used on the fly to create embeddings. A popular choice is the Highway Network.

CS 224N - Final Project

Manan Rai

Image Embeddings

- Image inputs usually require a more complicated neural network for feature extraction.
- These neural networks almost always involve convolutional blocks, since they can effectively capture modular features that may occur at multiple places in the image.
- VGG-19 is a popular network that has performed well on object detection tasks, especially on the ImageNet dataset.

CS 224N - Final Project

Manan Rai

Embeddings: Proposed "General" form

- Throwback to the biggest limitation of current embeddings: they *can* capture features for both input images and text separately, but they *can't* combine the salient features of both input formats.
- We propose a model that can do just that.
- Given an image embedding and a context embedding, we use a simple convolutional neural network to output a single, generalized embedding.

CS 224N - Final Project

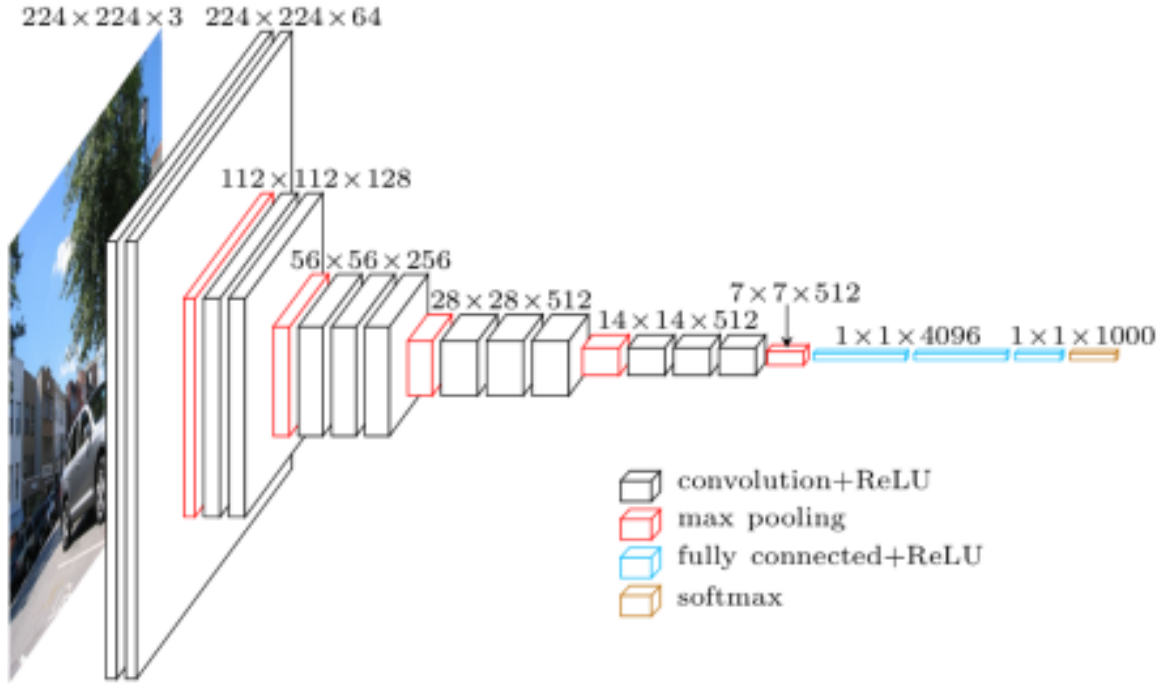
Manan Rai

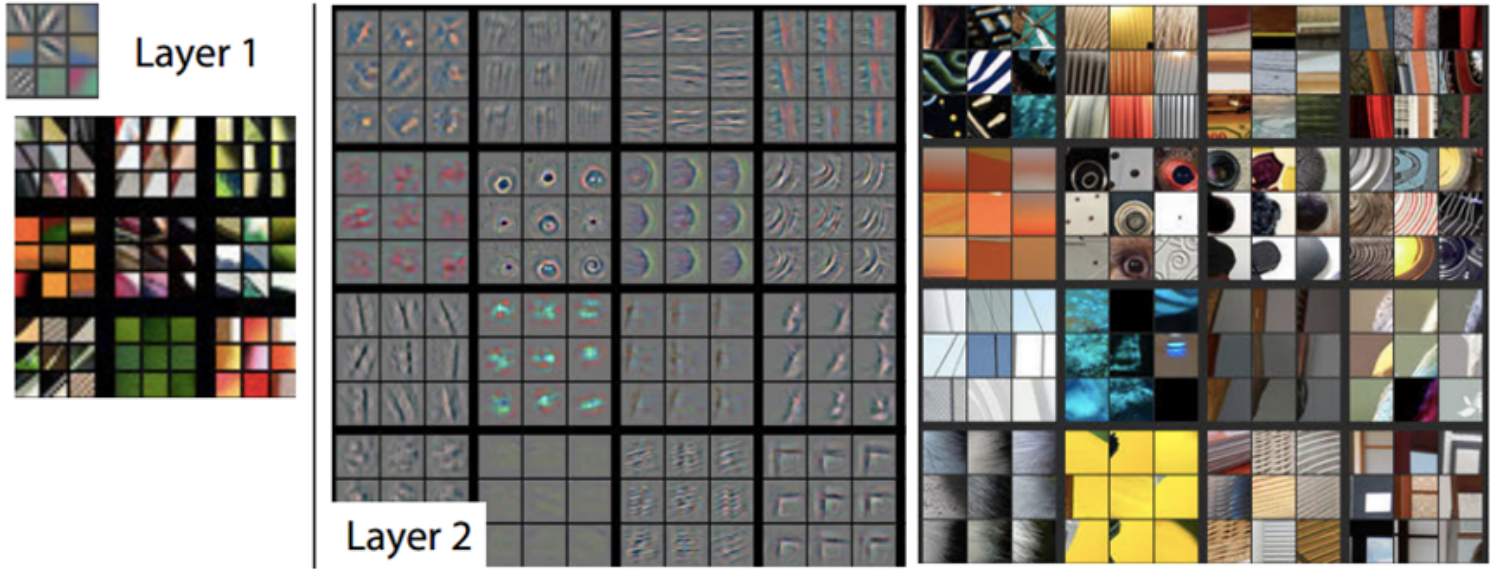
A New Wor(l)d Order

- The first step in building a rock-solid combined embedding, is building a rock-solid embedding for each of the input tasks separately.
- To do so, for image inputs, we experiment with a VGG-19

CS 224N - Final Project

Manan Rai





Visualizations of Layer 1 and 2. Each layer illustrates 2 pictures, one which shows the filters themselves and one that shows what part of the image are most strongly activated by the given filter. For example, in the space labled Layer 2, we have representations of the 16 different filters (on the left)

CS 224N - Final Project

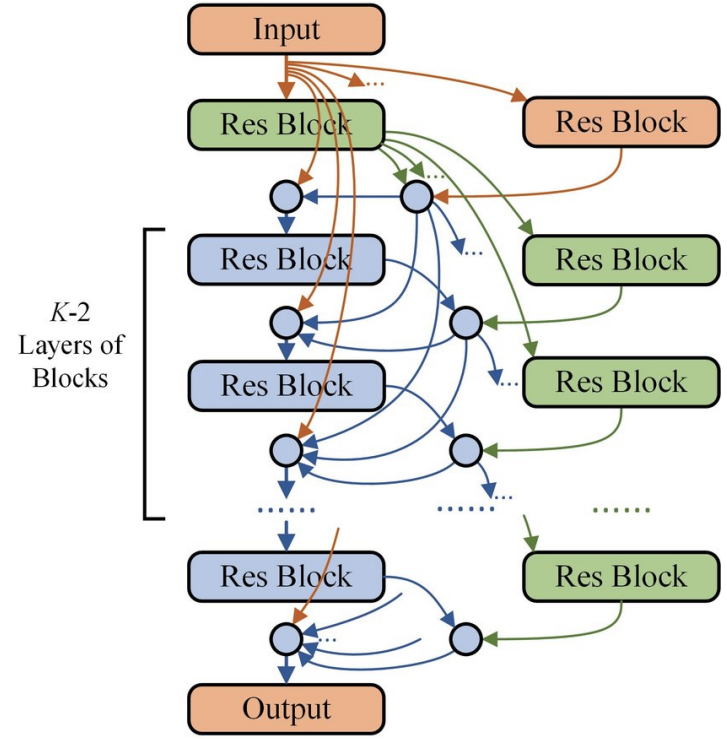
Manan Rai

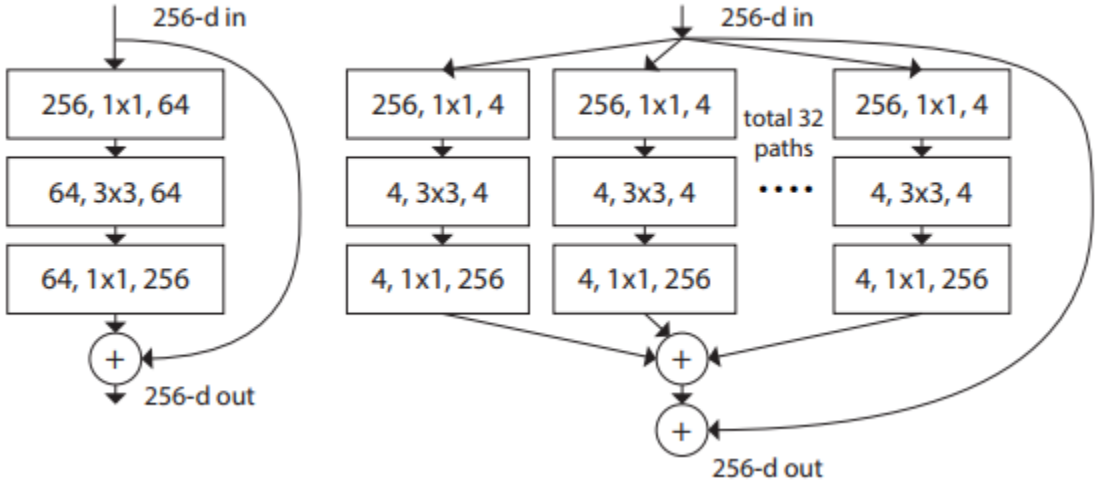
A New Wor(l)d Order

- The first step in building a rock-solid combined embedding, is building a rock-solid embedding for each of the input tasks separately.
- To do so, for image inputs, we experiment with a VGG-19, a ResNeXt

CS 224N - Final Project

Manan Rai





CS 224N - Final Project

Manan Rai

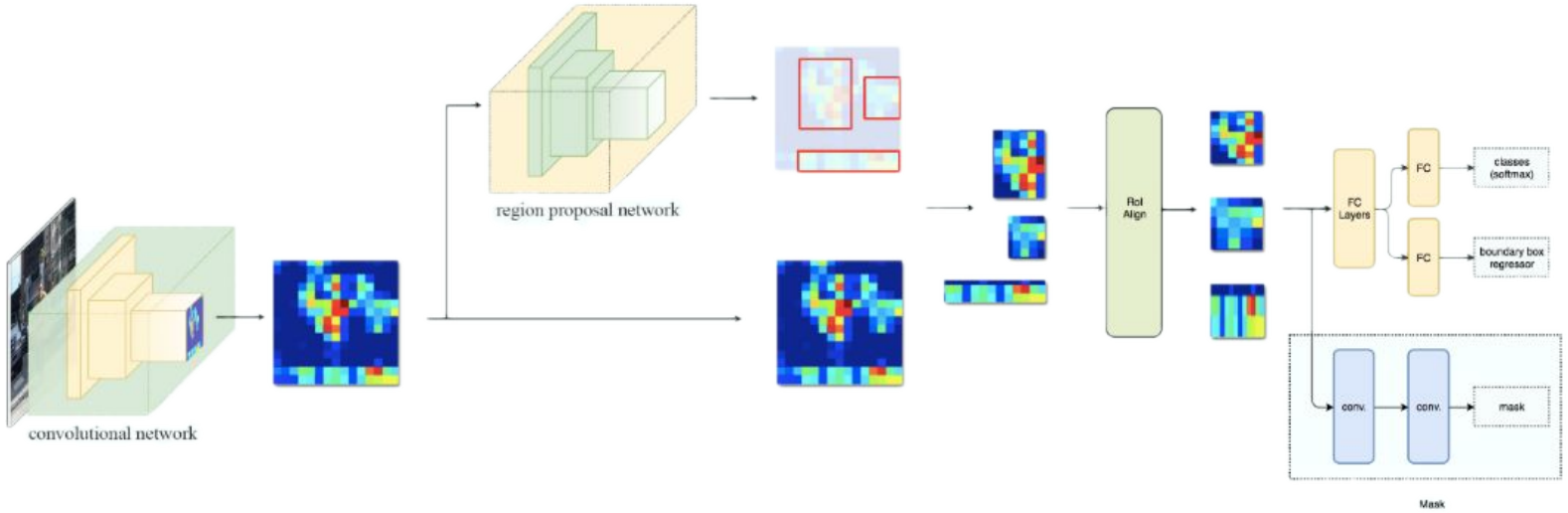
A New Wor(l)d Order

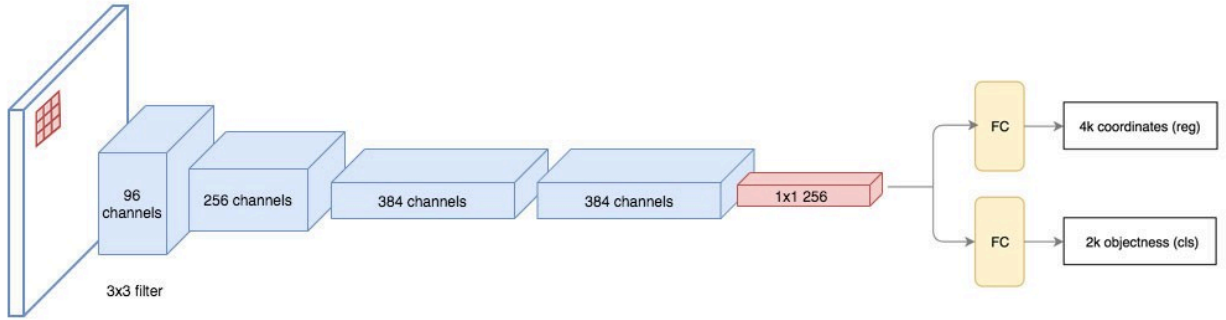
- The first step in building a rock-solid combined embedding, is building a rock-solid embedding for each of the input tasks separately.

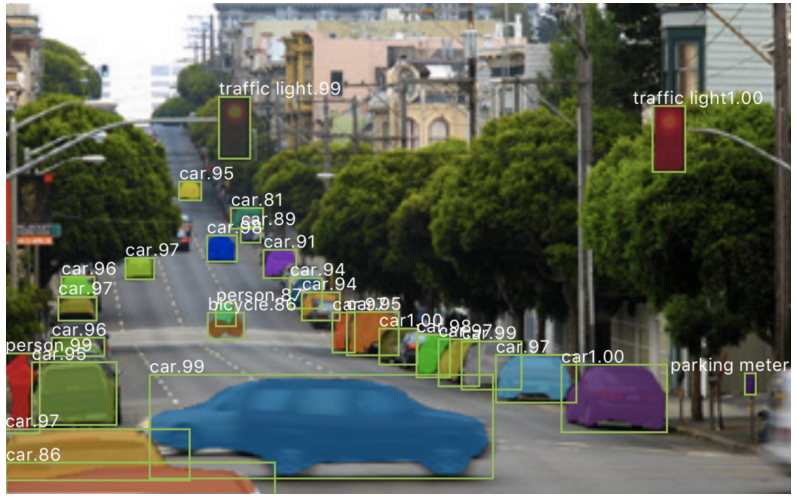
- To do so, for image inputs, we experiment with a VGG-19, a ResNeXt, and a Mask R-CNN with ResNeXt and DenseNet backbones

CS 224N - Final Project

Manan Rai







CS 224N - Final Project

Manan Rai

A New Wor(l)d Order

- The first step in building a rock-solid combined embedding, is building a rock-solid embedding for each of the input tasks separately.

- To do so, for image inputs, we experiment with a VGG-19, a ResNeXt, and a Mask R-CNN with ResNeXt and DenseNet backbones, all with a Multi-Layered Perceptron, to find that a Mask R-CNN with ResNeXt backbone performs the best.

CS 224N - Final Project

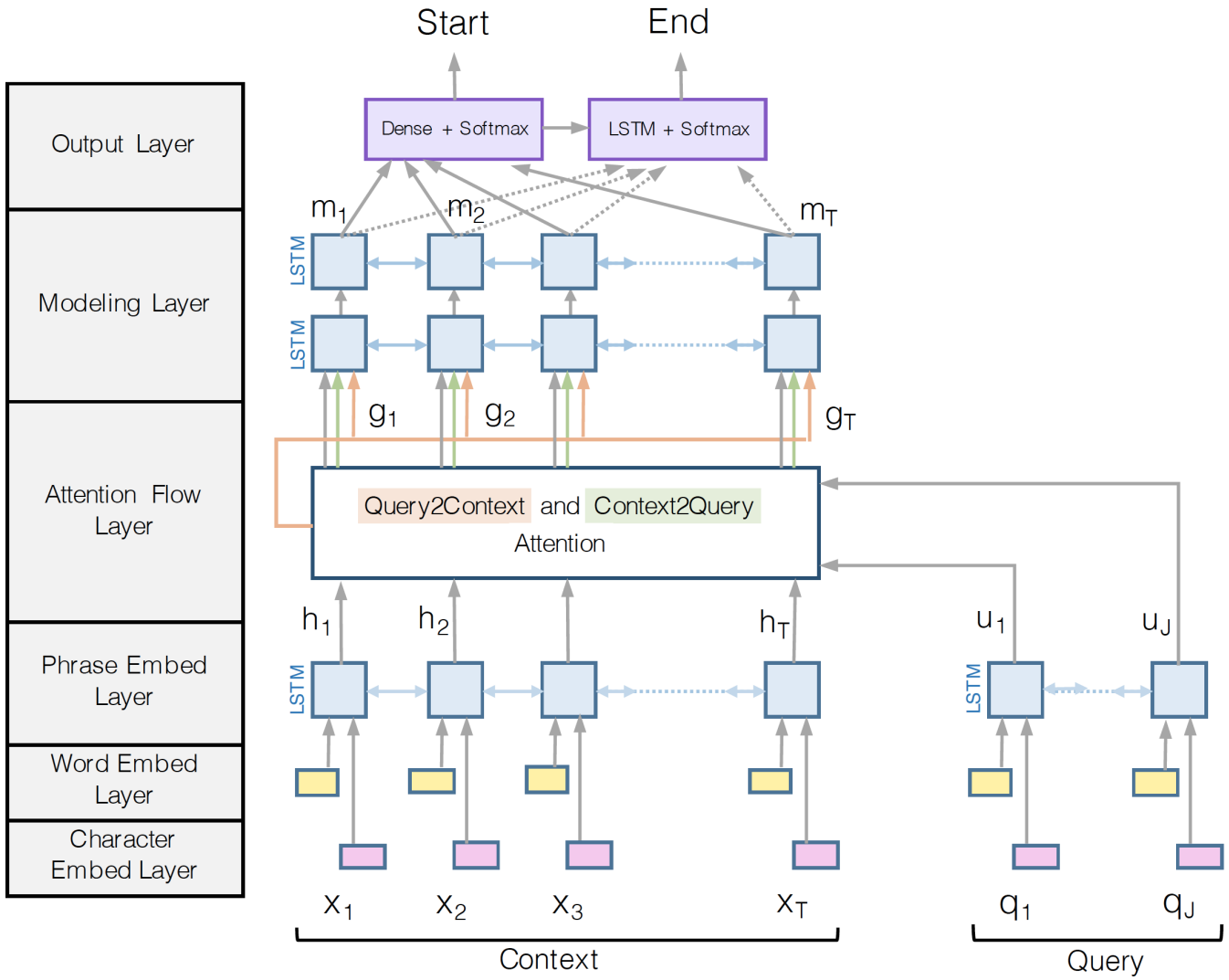
Manan Rai

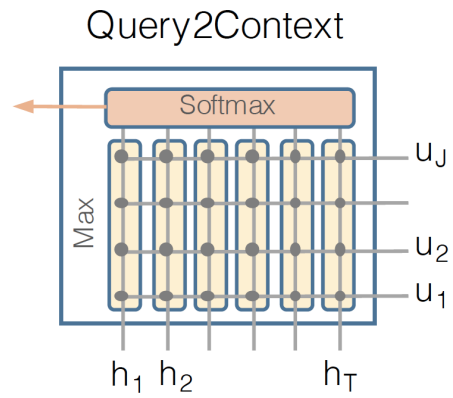
A New Wor(l)d Order

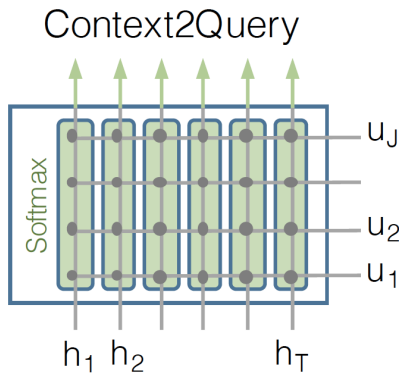
- The first step in building a rock-solid combined embedding, is building a rock-solid embedding for each of the input tasks separately.
- To do so, for image inputs, we experiment with a VGG-19, a ResNeXt, and a Mask R-CNN with ResNeXt and DenseNet backbones, all with a Multi-Layered Perceptron, to find that a Mask R-CNN with ResNeXt backbone performs the best.
- For text inputs, we experiment with a Highway Network with a Bidirectional Attention Flow model

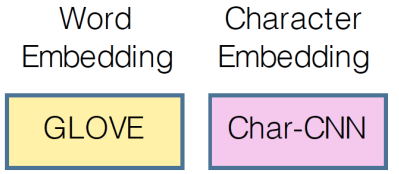
CS 224N - Final Project

Manan Rai









CS 224N - Final Project

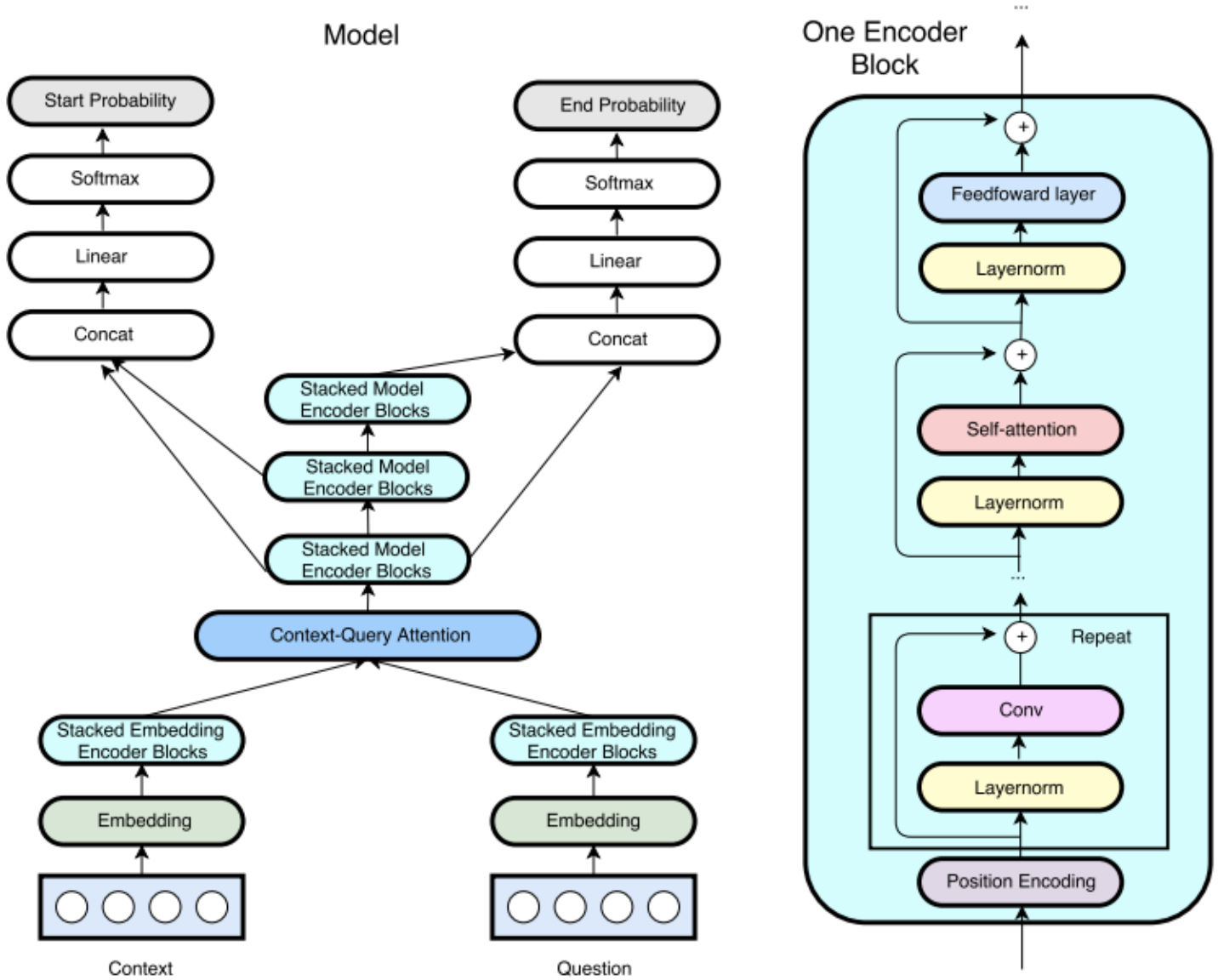
Manan Rai

A New Wor(l)d Order

- The first step in building a rock-solid combined embedding, is building a rock-solid embedding for each of the input tasks separately.
- To do so, for image inputs, we experiment with a VGG-19, a ResNeXt, and a Mask R-CNN with ResNeXt and DenseNet backbones, all with a Multi-Layered Perceptron, to find that a Mask R-CNN with ResNeXt backbone performs the best.
- For text inputs, we experiment with a Highway Network with a Bidirectional Attention Flow model, and GloVe embeddings with a Question Answering Network (QANet)

CS 224N - Final Project

Manan Rai



CS 224N - Final Project

Manan Rai

A New Wor(l)d Order

- The first step in building a rock-solid combined embedding, is building a rock-solid embedding for each of the input tasks separately.

- To do so, for image inputs, we experiment with a VGG-19, a ResNeXt, and a Mask R-CNN with ResNeXt and DenseNet backbones, all with a Multi-Layered Perceptron, to find that a Mask R-CNN with ResNeXt backbone performs the best.

- For text inputs, we experiment with a Highway Network with a Bidirectional Attention Flow model, and GloVe embeddings with a Question Answering Network (QANet), to find that the QANet performs better.

* (Note: Results obtained are presented in long form in the paper.)

CS 224N - Final Project

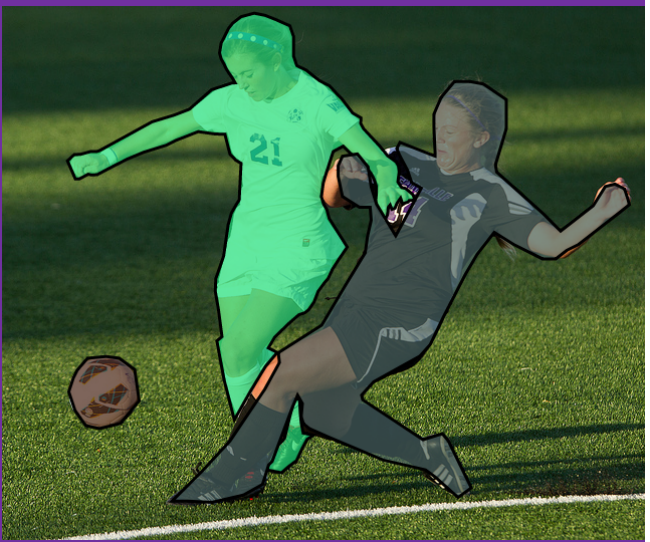
Manan Rai

A New Wor(l)d Order

- The next step is to combine these two embeddings, for which we build a simple convolutional network, and the parameters are learned from testing on both image and context inputs.
- We use the best models for the individual tasks to test the combined embeddings.
- We achieve a score of 49.29 on image inputs, and F1 and EM scores of 48.54 and 45.33 on context inputs.
- Even though these results are not as strong as those achieved by the individual models, they present a solid performance on both tasks together, which was the goal of the project.

CS 224N - Final Project

Manan Rai



What the Model got wrong

- Let us look at some places where the model didn't really hit the right chords.

- Confusing wording throws the model off:
  What are they playing? Soccer
  What are they doing? Playing frisbee

- Social Biases:
  Who is playing? (1) Man (2) Boy (3) Woman
  Are there any women in this picture? No (> 70%)

CS 224N - Final Project

Manan Rai

What the Model got wrong

- Let us look at some places where the model didn't really hit the right chords.
- Given a context passage reporting on a football game, and asked, "How many touchdowns were in this game?" the model is not able to identify that this is a counting problem and adopting an extractive answering strategy, it predicts "29-yard" (from the excerpt "…29-yard touchdown…") instead of the correct answer, "4".
- Other common errors include an incorrect end position (picking an answer longer than necessary), which is known to be solvable by conditioning the end position on the start.

CS 224N - Final Project

Manan Rai

Looking Ahead

- Note that this work presents only the first attempt at generalized input-format-independent question answering.
- We haven't yet utilized Attention techniques to their full extent, and the method of combining the embeddings is currently a simple neural network, that can benefit from techniques similar to those adopted for combining input and question embeddings, such as Hierarchical Co-Attention, which has achieved great success on the VQA task.

CS 224N - Final Project

Manan Rai

Thank You!

I would greatly appreciate any feedback on the project, and would be glad to answer any questions you have!

CS 224N - Final Project

Manan Rai

References:

1. S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh, "Vqa: Visual question answering," in *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2425–2433, Dec 2015.
2. M.J.Seo,A.Kembhavi,A.Farhadi,andH.Hajishirzi,"Bidirectional attention flow for machine comprehension," *CoRR*, vol. abs/1611.01603, 2016.
3. A. W. Yu, D. Dohan, M. Luong, R. Zhao, K. Chen, M. Norouzi, and Q. V. Le, "Qanet: Combining local convolution with global self-attention for reading comprehension," *CoRR*, vol. abs/1804.09541, 2018.

* (Note: Extended references and acknowledgements are presented in the paper.)