Introduction to Vision-Language Understanding

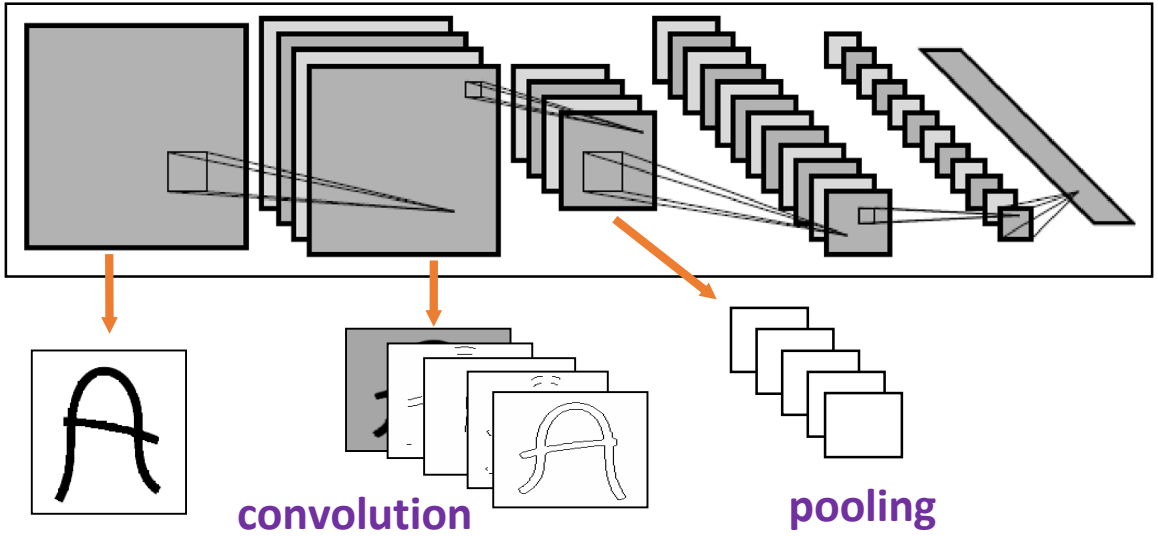
CS 585 Image and Video Computing - Spring 2020
Guest Lecture by Bryan A. Plummer

Today's Outline

- **Visual features:** Review/quick overview of convolutional neural networks
- **Language features:** one-hot vector, word2vec
- **Language models:** averaging, recurrent neural networks
- **Task learning:** bidirectional retrieval, image captioning, visual question answering

Previous lectures? Convolutional neural networks (CNN)

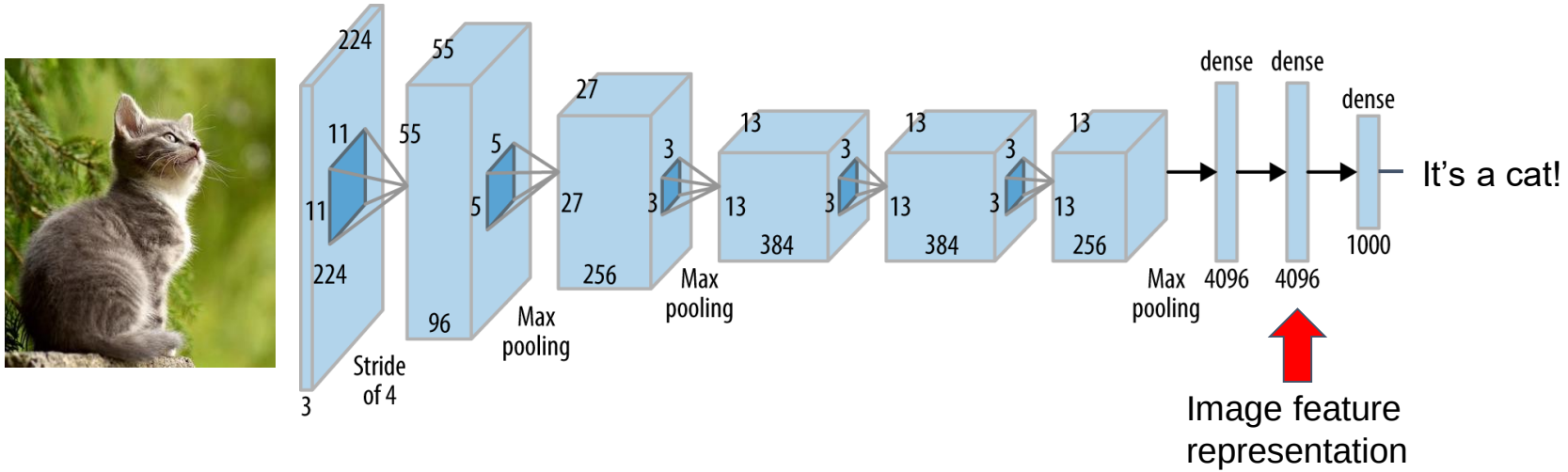
- Each convolutional layer outputs a feature map
- New layers are stacked on top of previous ones
- Subsample feature map with pooling layers



Slide credit: Brian Kulis & Kate Saenko

Previous lectures? Convolutional neural networks (CNN)

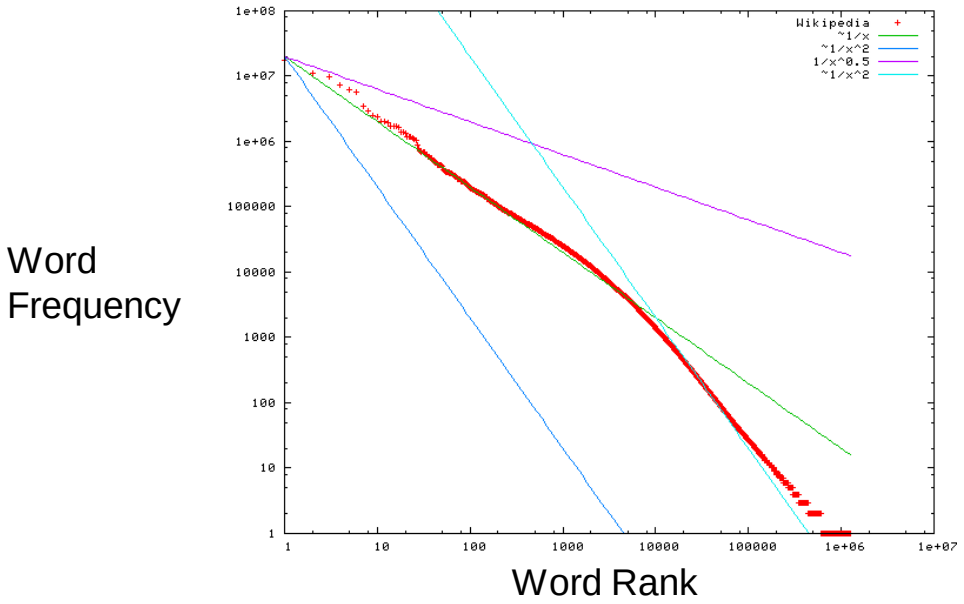
Krizhevsky et al. “ImageNet classification with deep convolutional neural networks.” NeurIPS, 2012.



How do vision-language tasks differ from tasks like image classification?

Zipf's Law - the frequency of a word is inversely proportional to its rank in a frequency table

Word Frequency on Wikipedia (November 27, 2006)



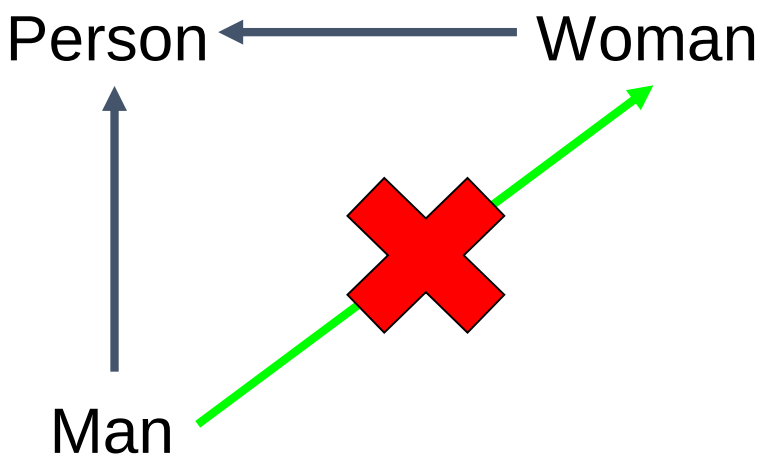
How would you describe this image?

An orchestra performing in a concert hall.
The Dublin Symphony Orchestra playing in an ornate theatre.

How would you describe this image region?

a man
a bearded man
a person with sunglasses
…

Non-transitive semantic relationships



Requires abstract/commonsense reasoning

Two siblings are walking on rocks across a river

Some of the challenges for vision-language tasks

- Words can relate to objects, their attributes, relationships between entities, or require commonsense reasoning
- Non-transitive semantic relationships
- Very sparsely labeled
- Many words seen at test time may not be present in the training data
- Word ordering may change what kinds of images you would expect to see

Learning a similarity function between features

Neural Net

Feature Set A

“An orchestra performing in a concert hall.”

Neural Net

Feature Set B

Shared Semantic Space

Today's Outline

- **Visual features:** Review/quick overview of convolutional neural networks
- **Language features:** one-hot vector, word2vec
- **Language models:** averaging, recurrent neural networks
- **Task learning:** bidirectional retrieval, image captioning, visual question answering

One-hot vector representation

One-hot vector representation

car

One-hot vector representation

zebra

Neural Net

Feature Set A

“An orchestra performing in a concert hall.”

Neural Net

Feature Set B

Shared Semantic Space

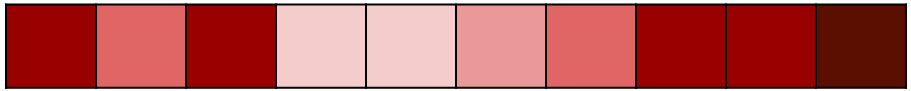
Semantic relationships between words must be learned by the neural network

kid

child

Word vector representations

kid d( ,

child ) = 0.86

A closer look at word context

1. A man with dark hair is wearing a white jacket with printed flames and is holding a pair of tongs.

1. Two people, one in a white and blue jacket and one in a blue polo bending at the knees in front of some bushes.

1. A small white and gray dog in a jacket standing in the snow.

A closer look at word context

1. A man with dark hair is wearing a white jacket with printed flames and is holding a pair of tongs.

1. Two people, one in a white and blue jacket and one in a blue polo bending at the knees in front of some bushes.

1. A small white and gray dog in a jacket standing in the snow.

A closer look at word context

1. A man with dark hair is wearing a white jacket with printed flames and is holding a pair of tongs.

1. Two people, one in a white and blue jacket and one in a blue polo bending at the knees in front of some bushes.

1. A small white and gray dog in a jacket standing in the snow.

A closer look at word context...

1. A man with dark hair is wearing a white jacket with printed flames and is holding a pair of tongs.

1. Two people, one in a white and blue jacket and one in a blue polo bending at the knees in front of some bushes.

1. A small white and gray dog in a jacket standing in the snow.

A closer look at word context...

Training Set

a white jacket

a blue and white jacket

a jacket

Fill in the blank

a _____ jacket

Idea from Bengio et al. 2003

Word2Vec

- Train a classifier to predict words based on context
- For any sampled word, the other words in a lexicon as negative samples
- Train a classifier using logistic regression (learned classifier becomes word embeddings)
- Can be trained without requiring additional labels

Demo: http://bionlp-www.utu.fi/wv_demo/

Today's Outline

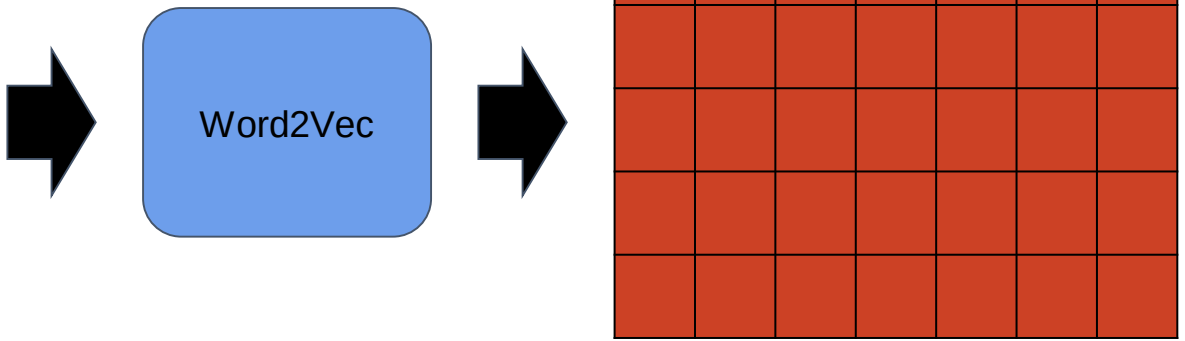
- **Visual features:** Review/quick overview of convolutional neural networks
- **Language features:** one-hot vector, word2vec
- **Language models:** averaging, recurrent neural networks
- **Task learning:** bidirectional retrieval, image captioning, visual question answering

Representing multiple words

N words

N x d features

An orchestra performing in a concert hall.



Neural Net

Feature Set A

“An orchestra performing in a concert hall.”

Word2Vec

Avg Pool

Neural Net

Feature Set B

Shared Semantic Space

Averaging word vectors

- Simple and efficient, often works well
- Loses information like word ordering
- Can’t be used in generative tasks like image captioning

Recurrent Neural Networks

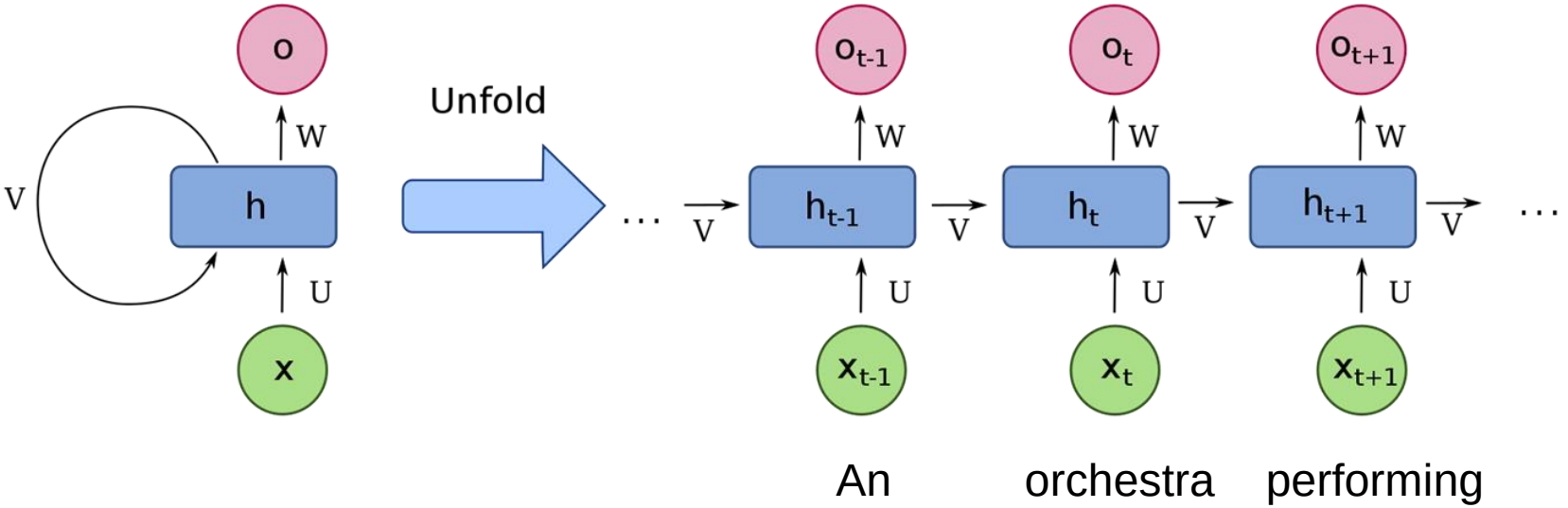


image source: https://en.wikipedia.org/wiki/Recurrent_neural_network

Recurrent neural networks

- Can represent sequences of data
- Adds additional parameters to the network
- Can be relatively slow for large layers

Today's Outline

- **Visual features:** Review/quick overview of convolutional neural networks
- **Language features:** one-hot vector, word2vec
- **Language models:** averaging, recurrent neural networks
- **Task learning:** bidirectional retrieval, image captioning, visual question answering

Image-sentence matching

Neural Net

Feature Set A

“An orchestra performing in a concert hall.”

Neural Net

Feature Set B

Shared Semantic Space

Phrase localization

Localize: “a domino mask”

A man in “a domino mask” carrying a speaker.

Neural Net

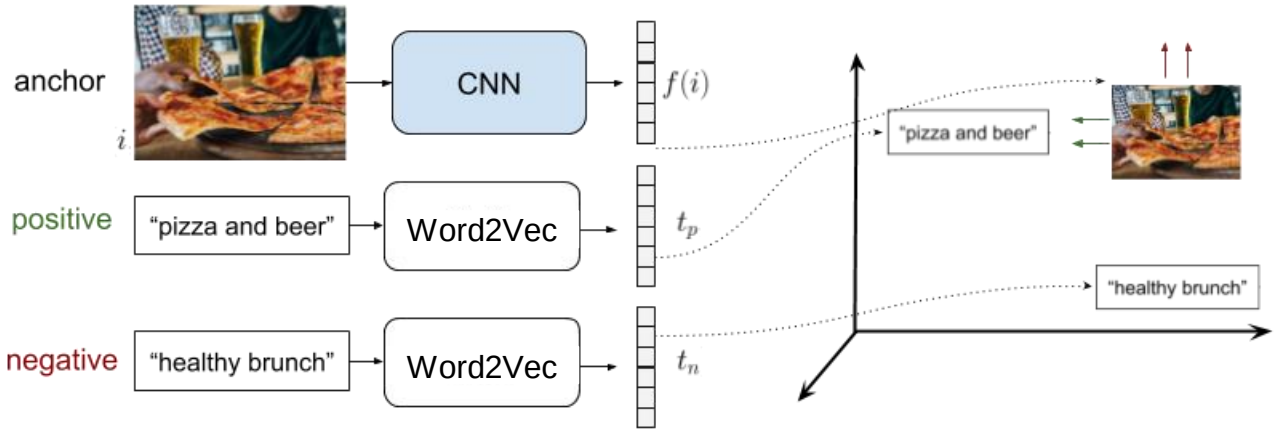
Neural Net

Feature Set A

Feature Set B

Shared Semantic Space

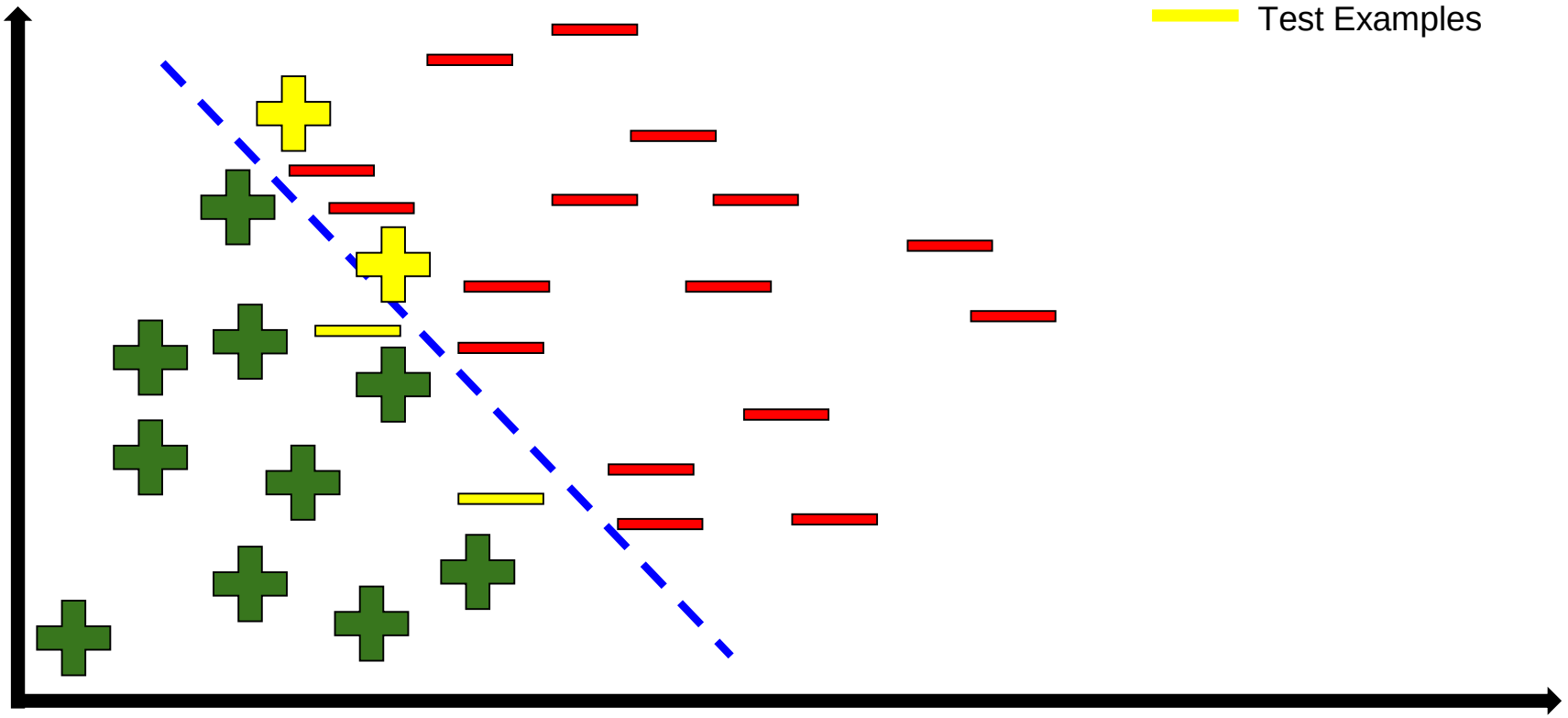
Pairwise ranking loss



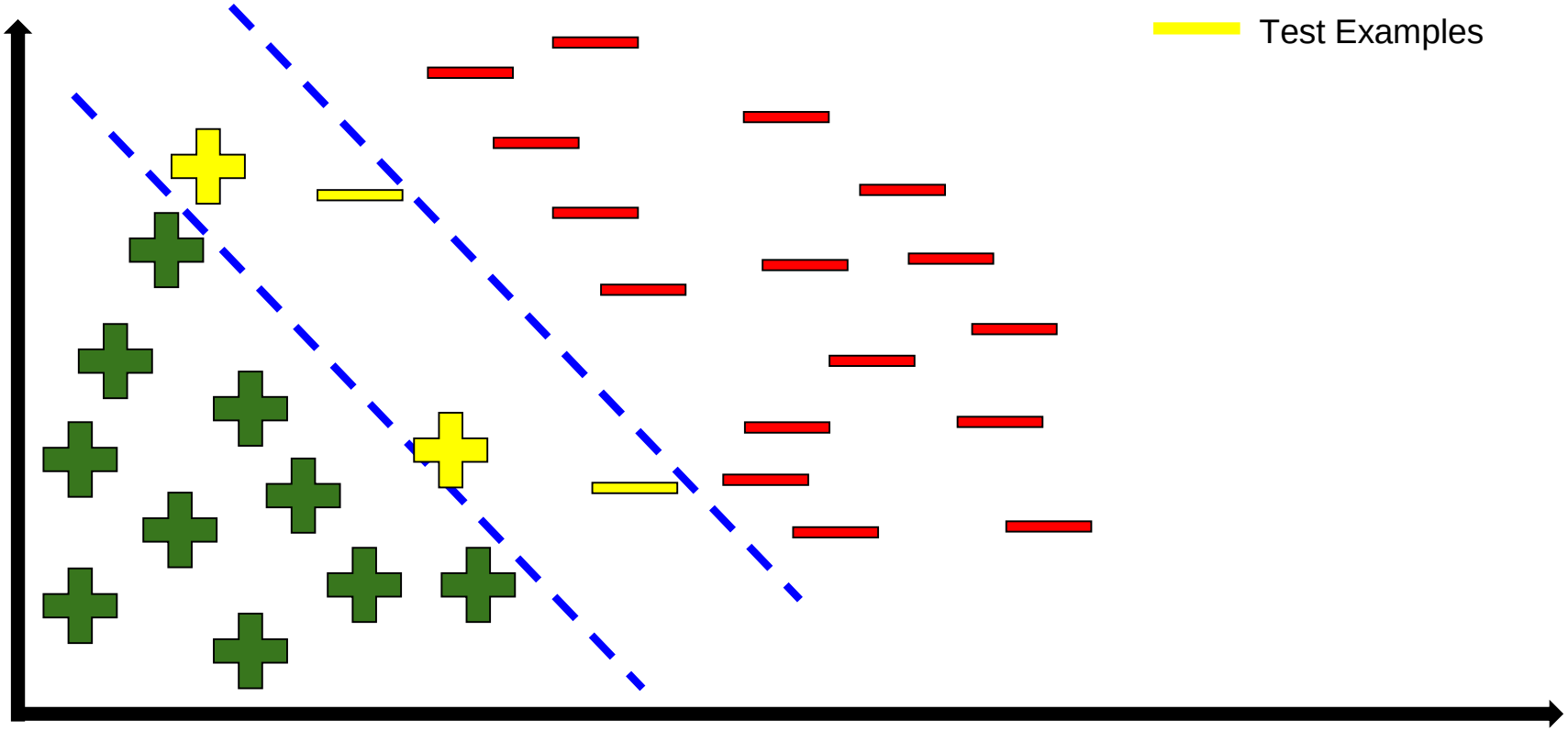
$$PairwiseRanking = \begin{cases} \|f(i) - t_p\|, & if\ Positive \\ \max(0, m - \|f(i) - t_n\|), & if\ Negative \end{cases}$$

image source: https://gombru.github.io/2019/04/03/ranking_loss/

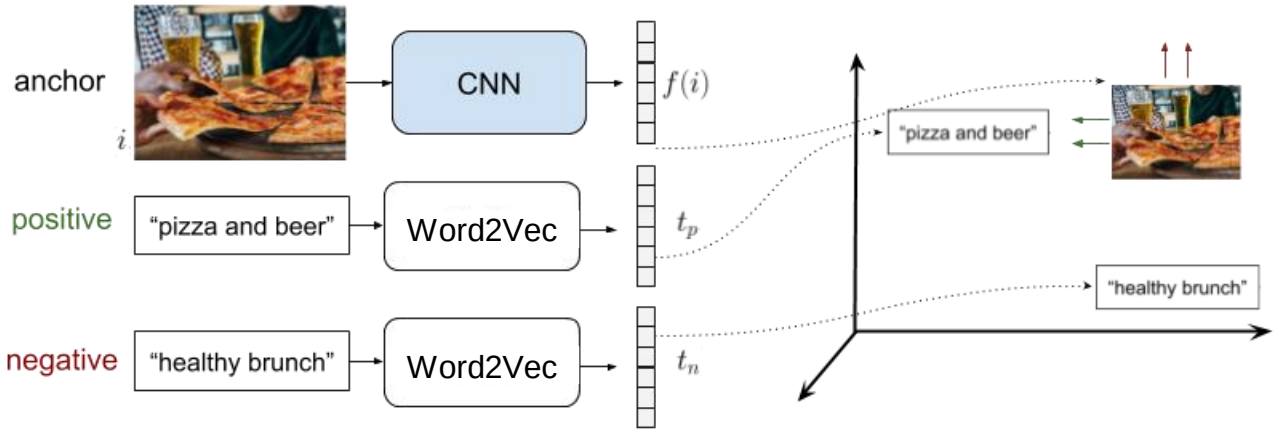
Issues in practice



Better to enforce a margin between samples



Triplet ranking loss

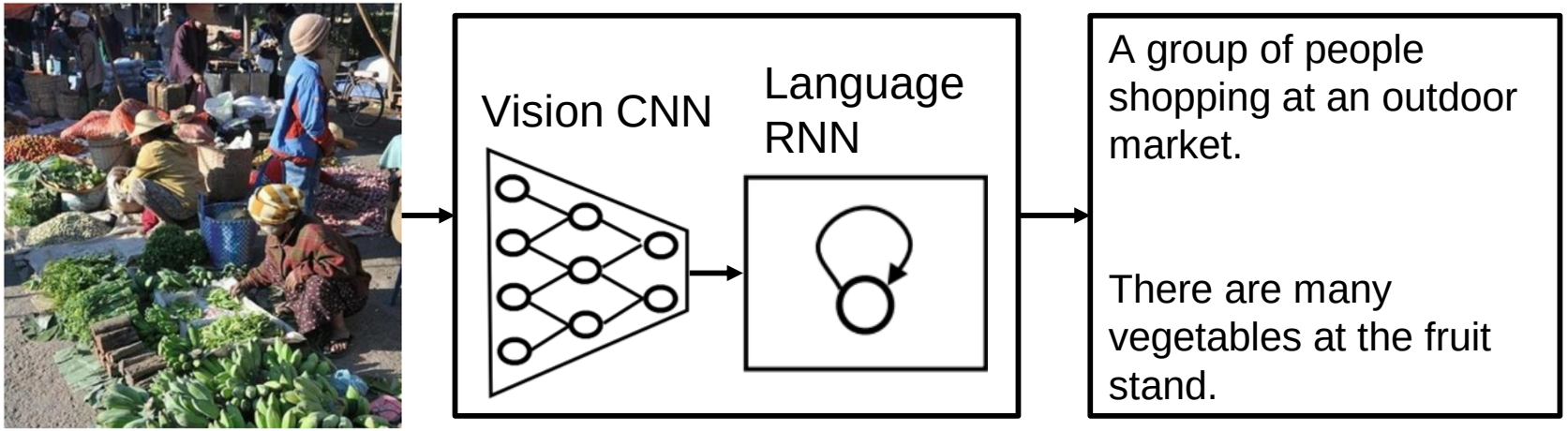


$$TripletRanking = \max(0, m + \|f(i) - t_p\| - \|f(i) - t_n\|)$$

image source: https://gombru.github.io/2019/04/03/ranking_loss/

Image captioning

Vinyals et al., Show and Tell: A Neural Image Caption Generator, *CVPR*, 2015.



LSTM inputs

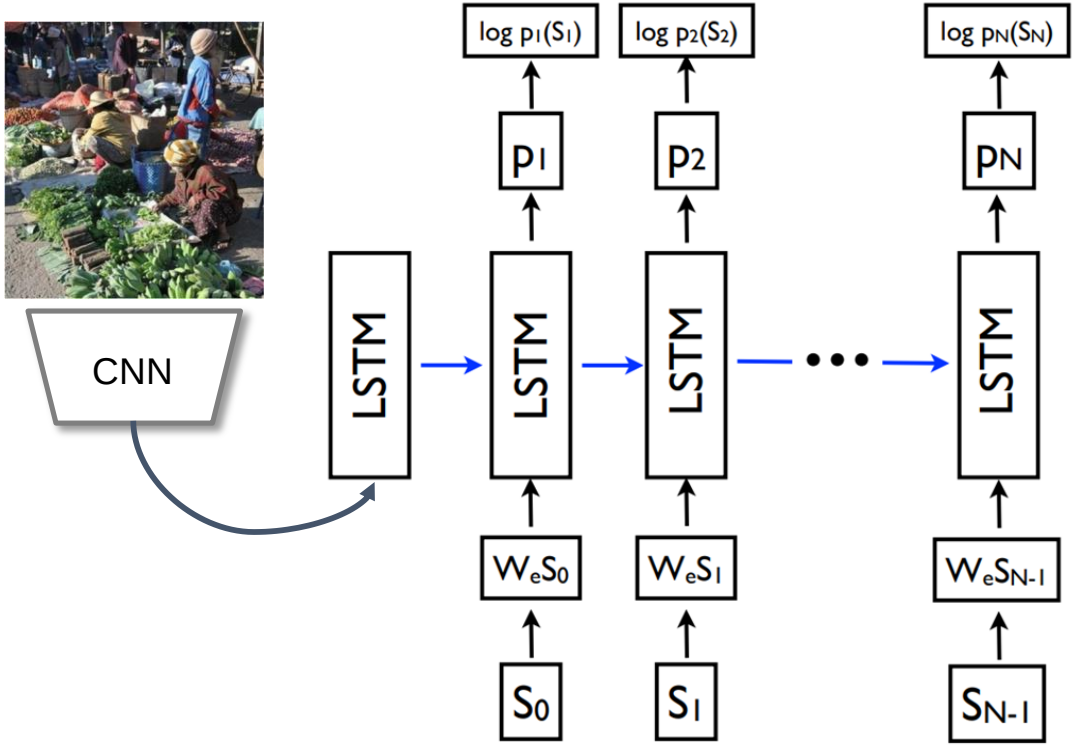


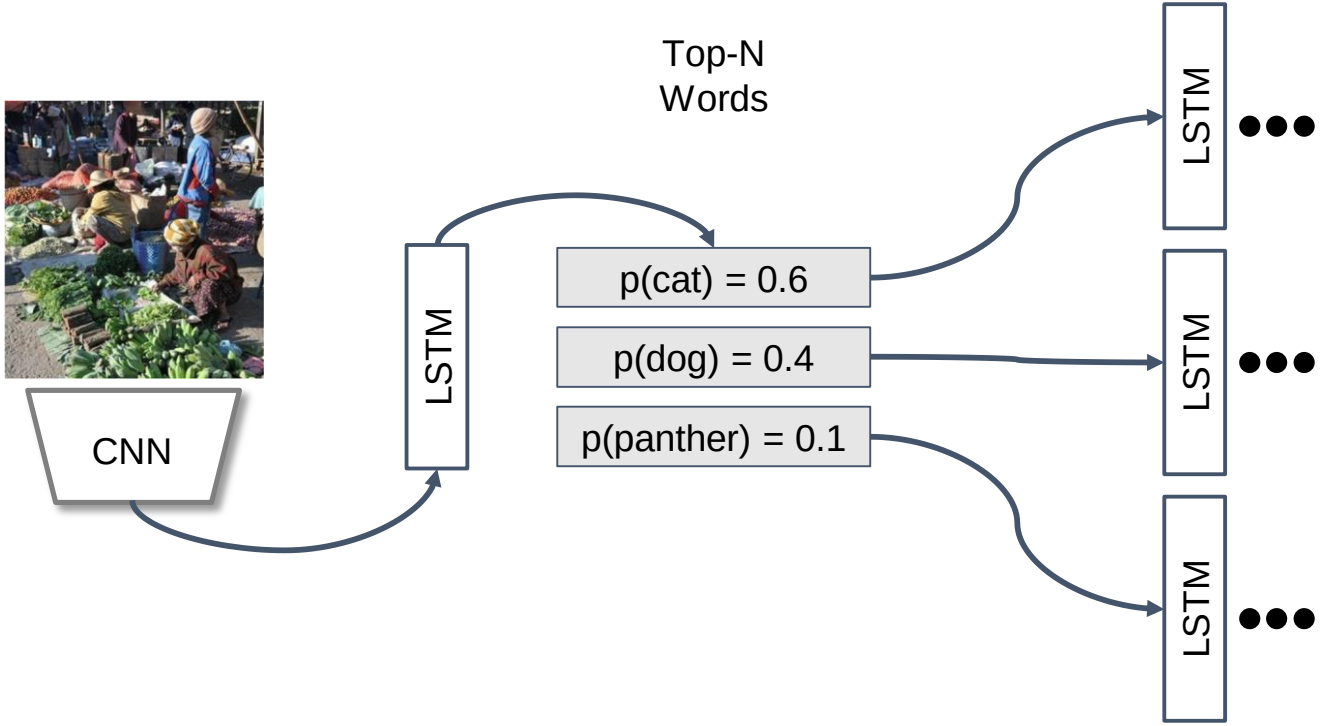
image source: Show and Tell: A Neural Image Caption Generator, CVPR, 2015

Negative log likelihood loss

Trains a model to maximize the likelihood of a caption

$$NegLogLikelihood = -\sum_{t=1}^{N} \log p_t(S_t)$$

Generating a caption - beam search



Demo: http://dbs.cloudcv.org/captioning

VQA Example

Inputs

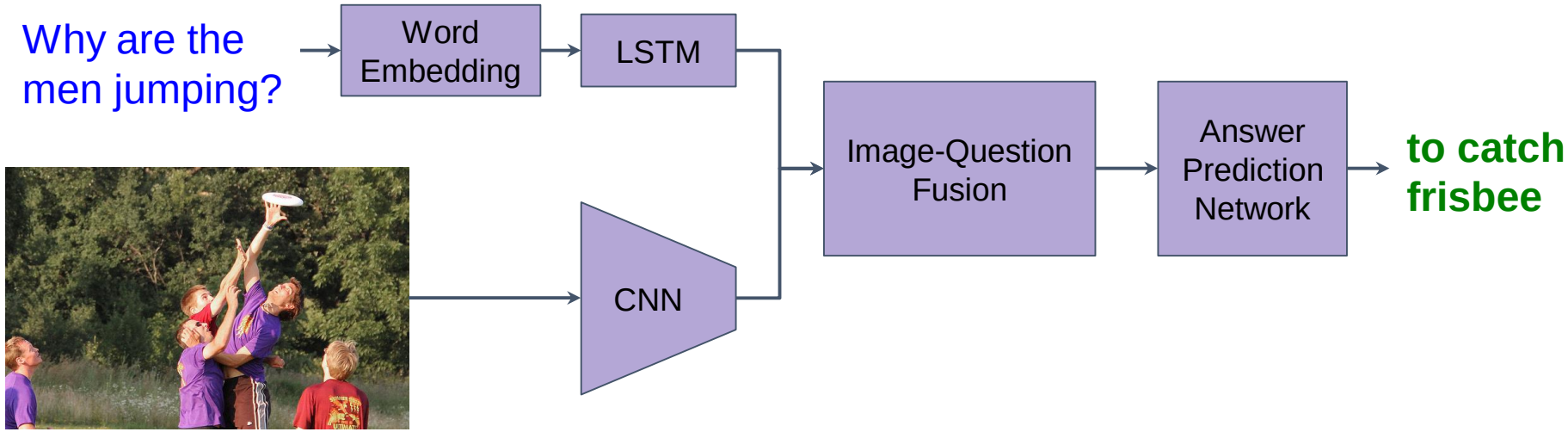
Question: Why are the men jumping?

Outputs

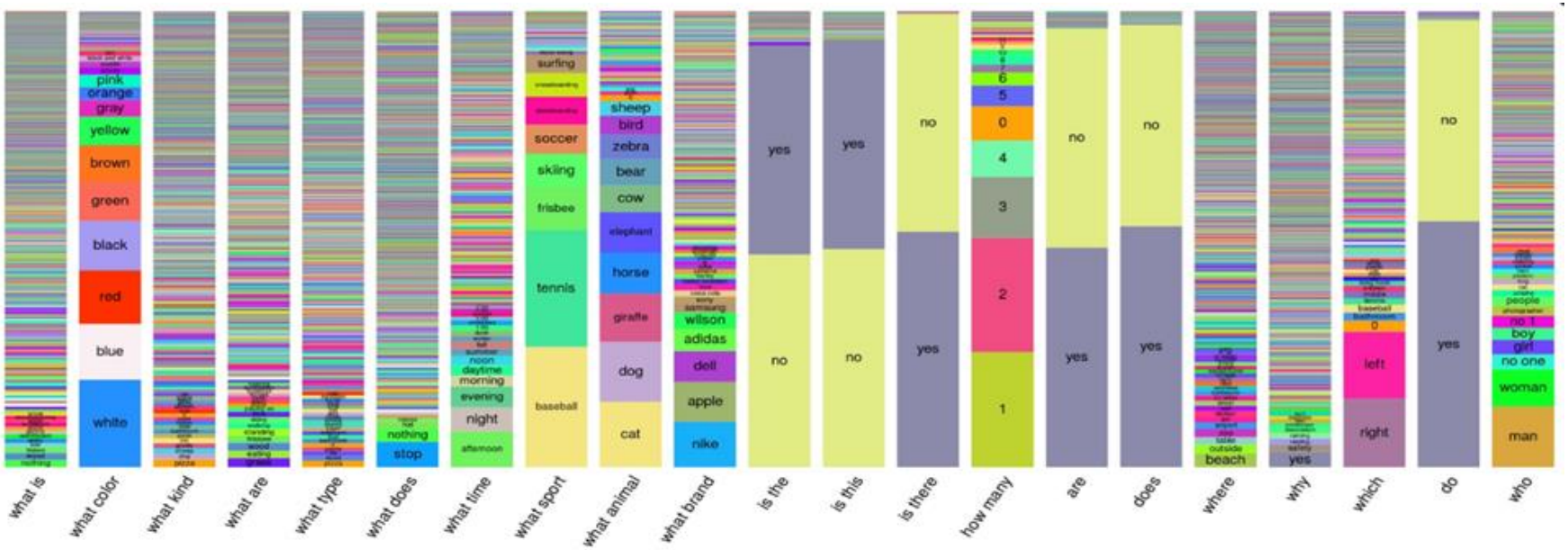
Answer: **to catch frisbee**

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR, 2017.

VQA Generic Approach Components

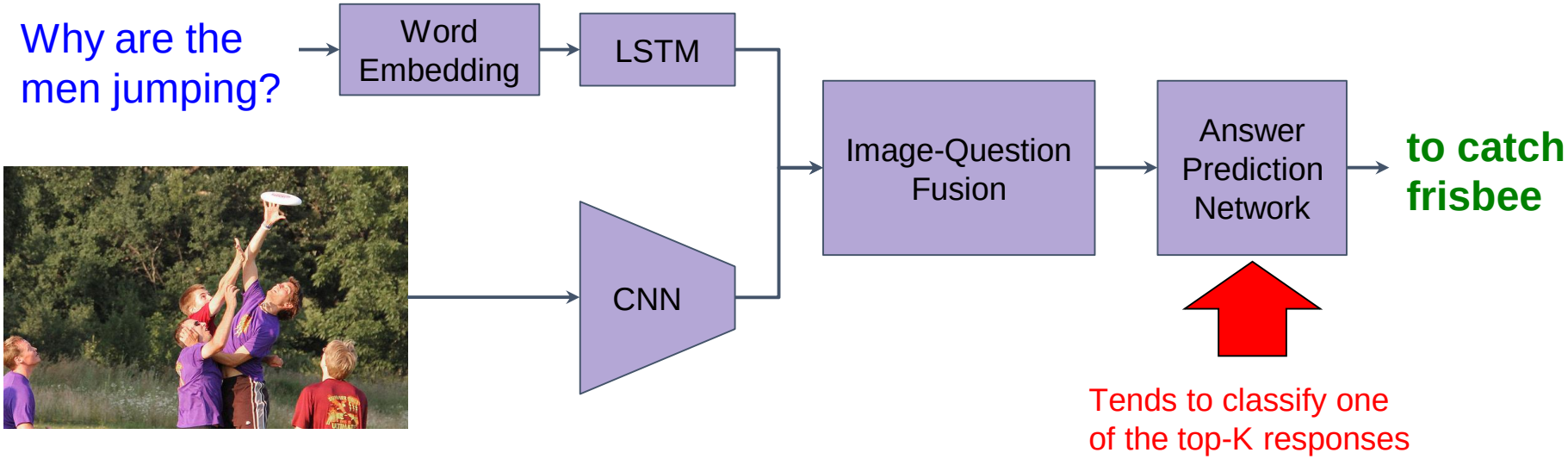


Answer distribution

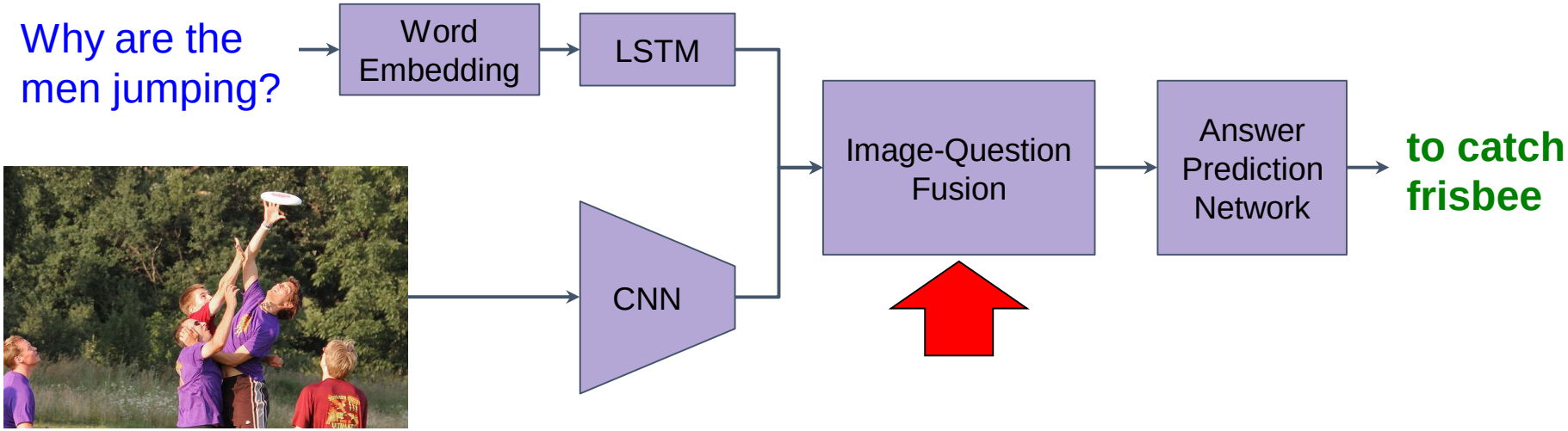


Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR, 2017.

VQA Generic Approach Components



VQA Generic Approach Components



Methods of fusing Image-Question Information

- Concatenate
- Elementwise Sum
- Elementwise Product (inner product)
- Bilinear Pooling (outer product)

A feature vector of size 2048 and desired output dimension of 3000 would require 12.5 billion parameters!

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, Marcus Rohrbach. Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding. EMNLP, 2016.

Performance of different fusion methods on VQAv1

Integration method	Accuracy
Concatenation	57.49
Elementwise Sum	56.50
Elementwise Product	58.57
Bilinear Pooling	59.83

Summary of some takeaways

- Combing vision and language is challenging
- Word vector representations tends to outperform alternatives like one-hot vector representations
- Margin based triplet losses are more robust than simple pairwise losses, but triplet sampling is important
- How vision-language features are combined can significantly impact model accuracy

Additional reading/useful references

- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. ICML, 2015.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, Lei Zhang. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. CVPR, 2018.
- Peter Anderson, Basura Fernando, Mark Johnson, Stephen Gould. SPICE: Semantic Propositional Image Caption Evaluation. ECCV, 2016.
- Chao-Yuan Wu, R. Manmatha, Alexander J. Smola, Philipp Krähenbühl. Sampling Matters in Deep Embedding Learning. ICCV, 2017.
- Andrea Burns, Reuben Tan, Kate Saenko, Stan Sclaroff, Bryan A. Plummer. Language Features Matter: Effective Language Representations for Vision-Language Tasks. ICCV, 2019.