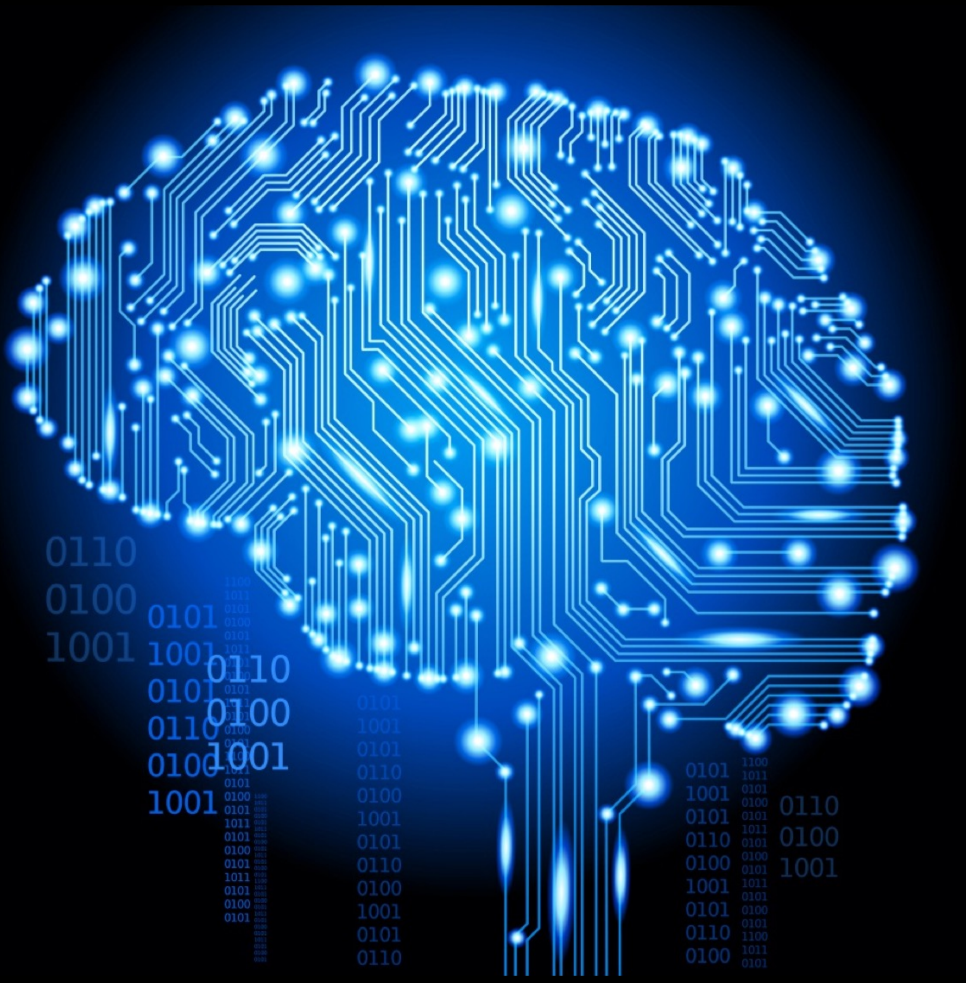
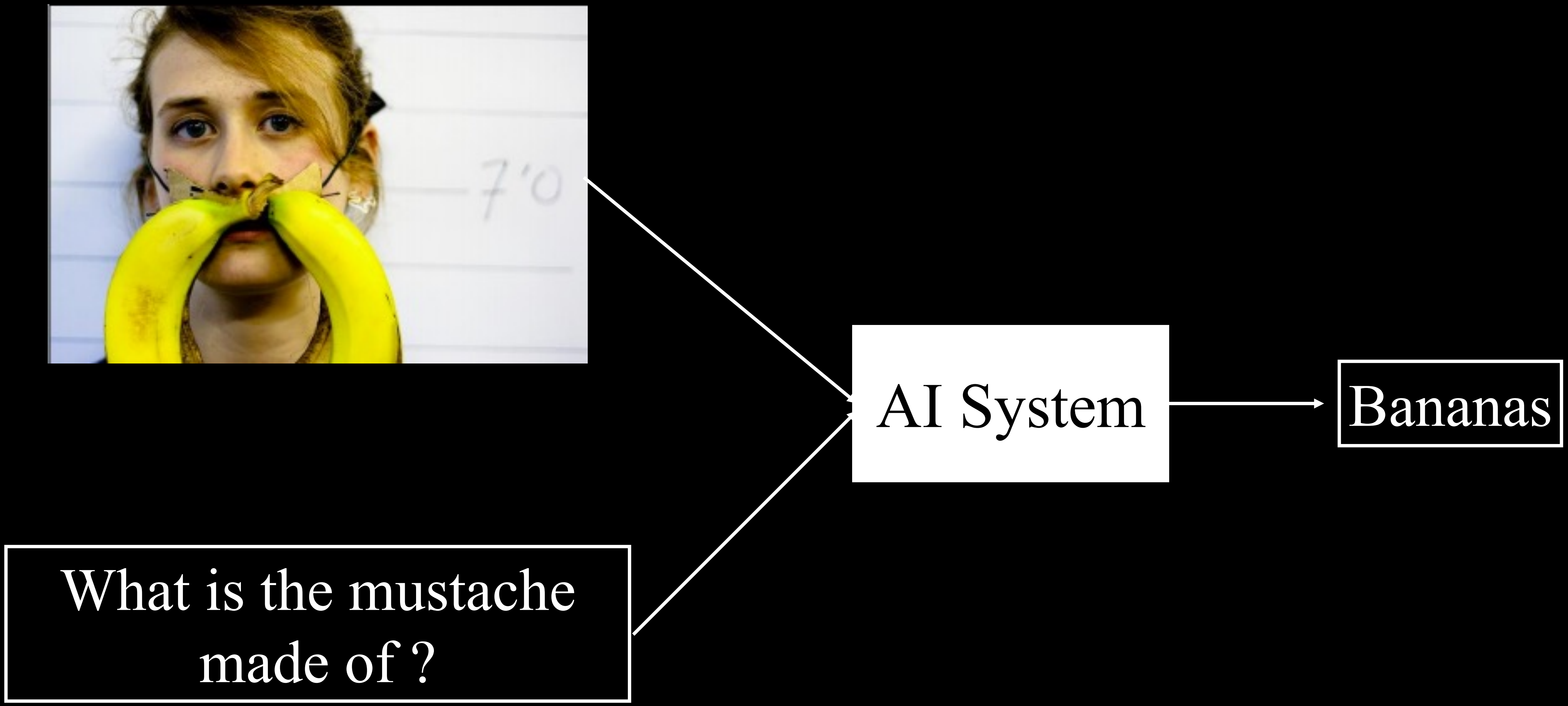
Visual Question Answering
with
Deep Learning



What is VQA?

Visual Question Answering



What is the mustache made of ?

AI System

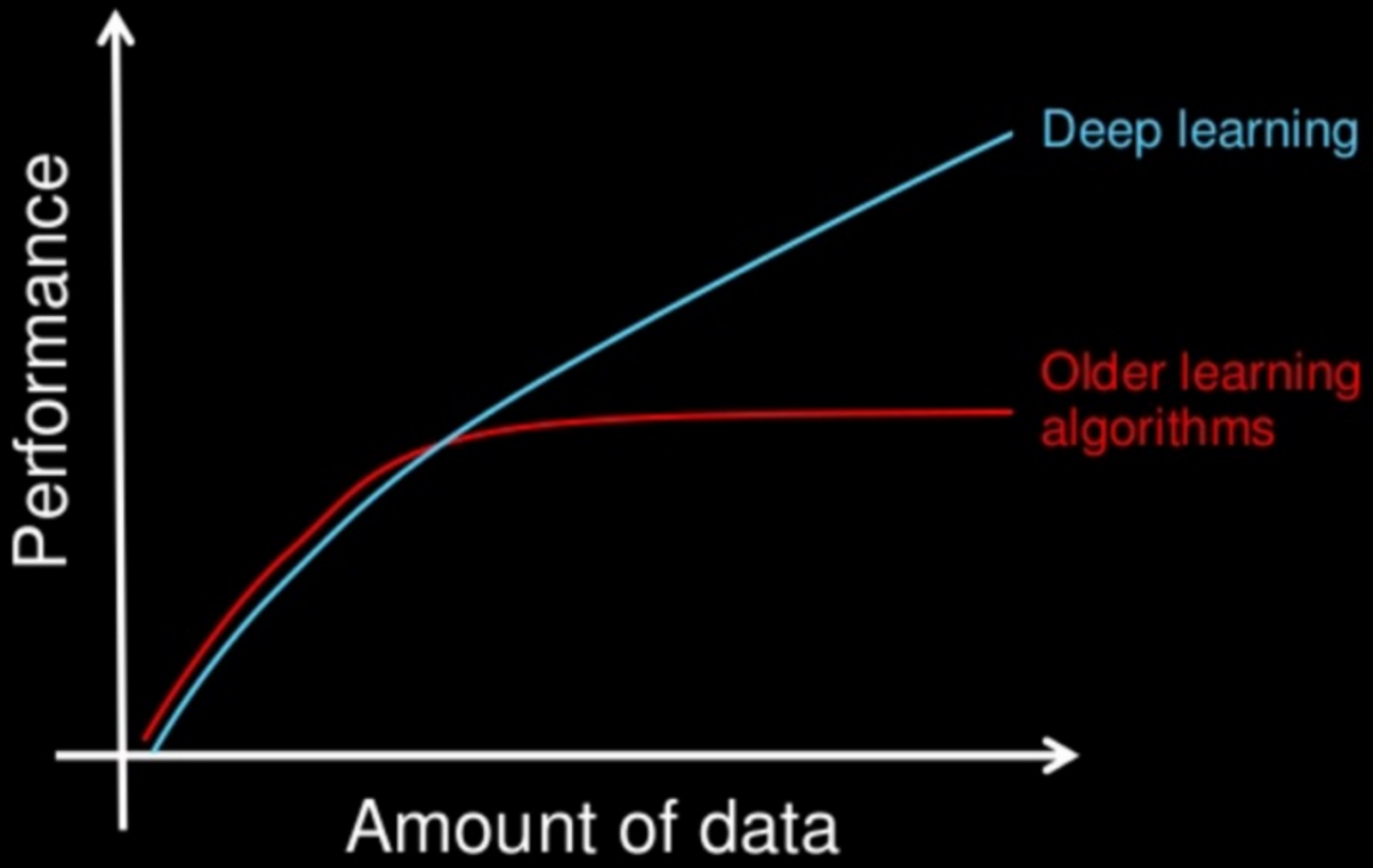
Bananas

Why?

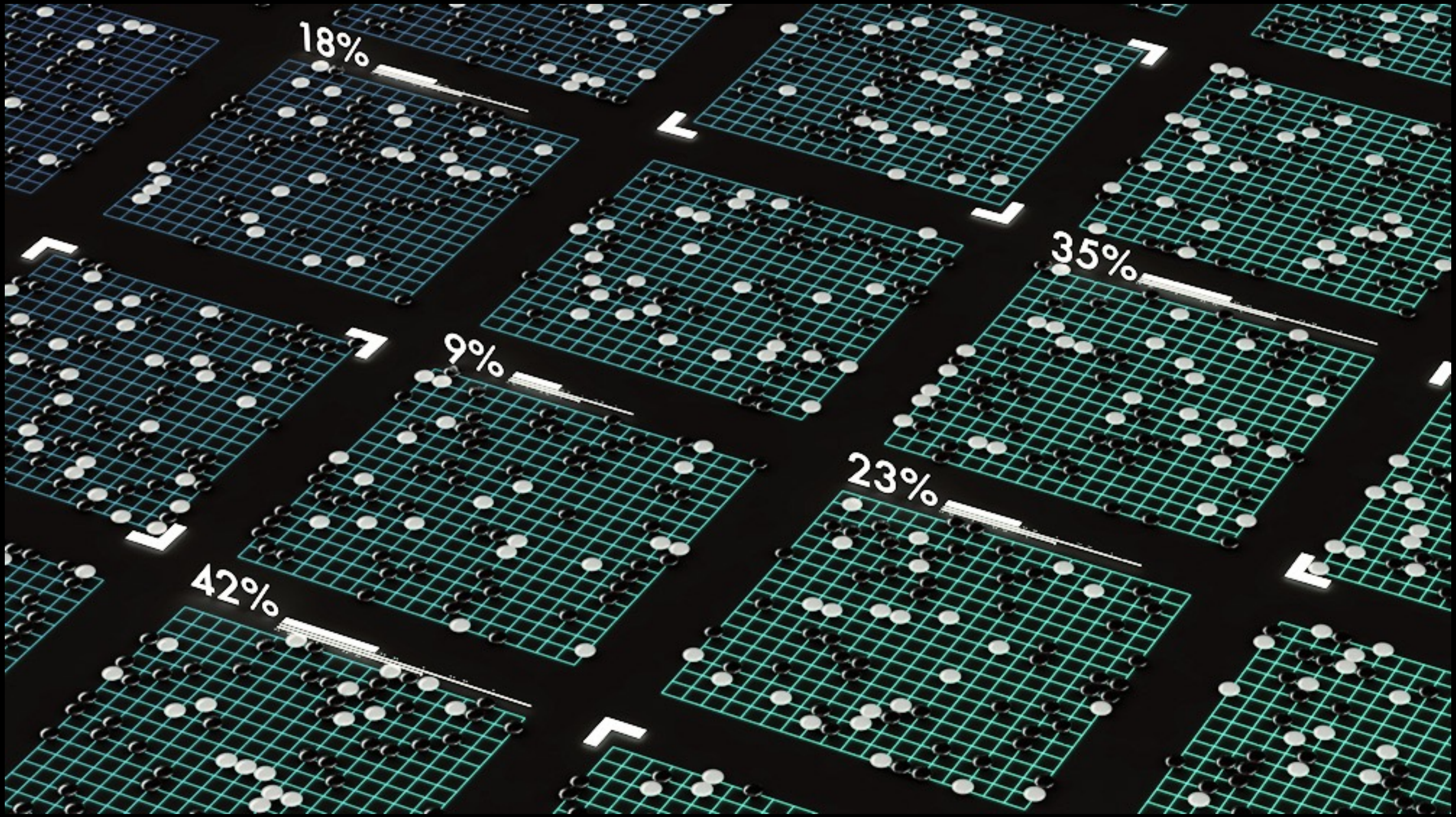
Deep Learning

- Scalability (more data, larger model and using more computation to train)
- Perform automatic feature extraction from raw data
- Be characterized as a rebranding of neural networks
- DNN, CNN, RNN, LSTM

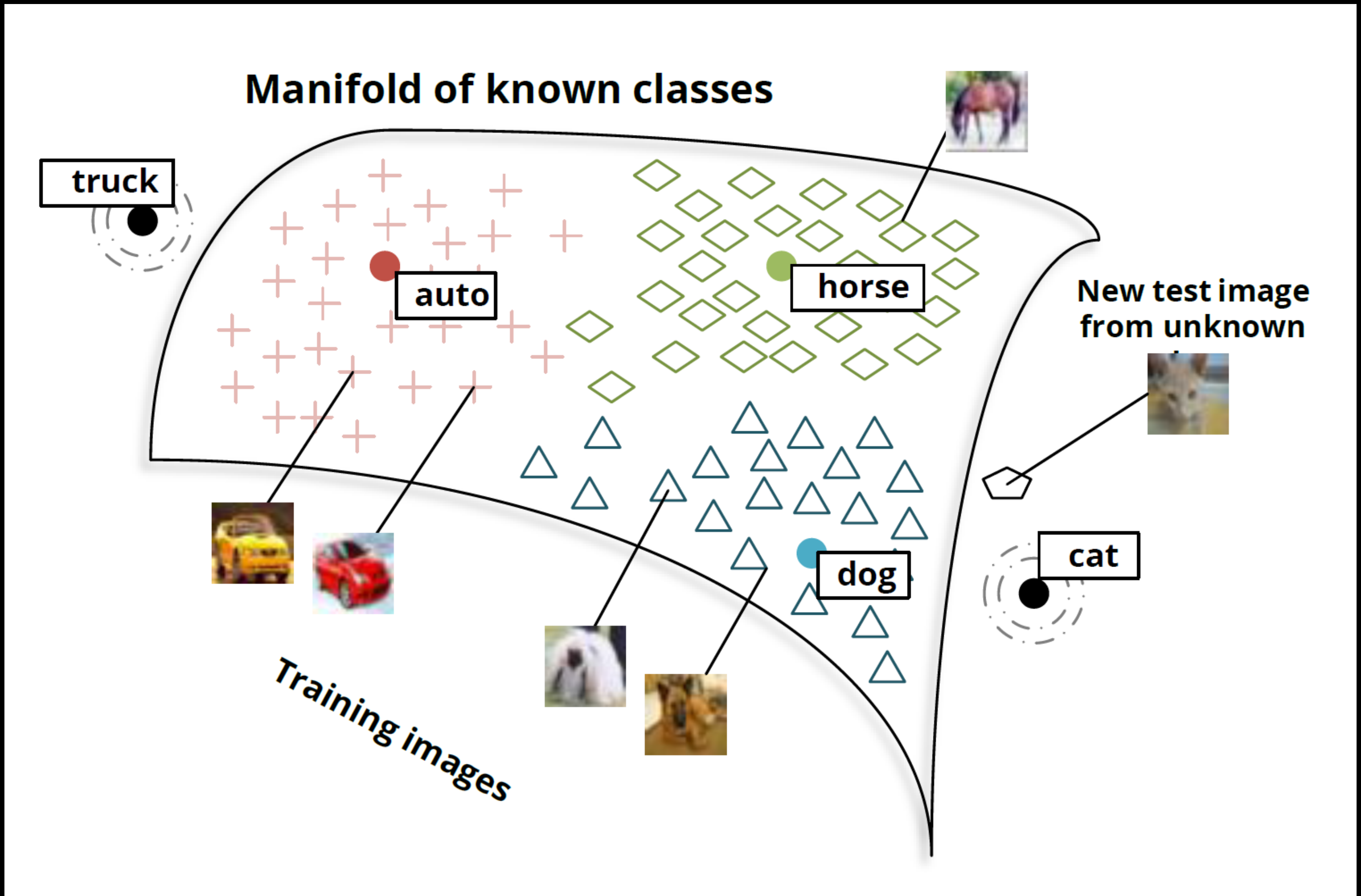
Deep Learning



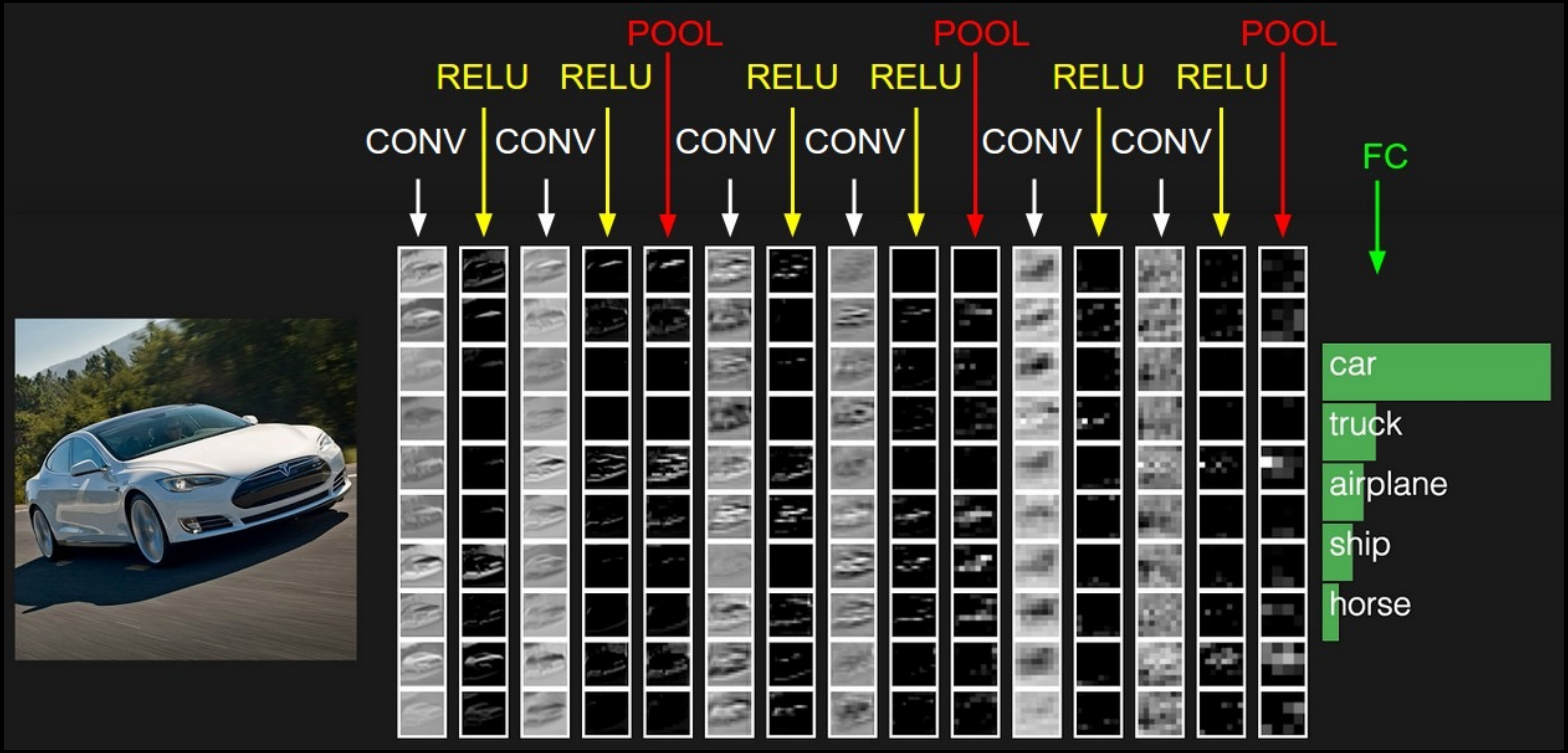
AlphaGo



Natural Language Processing

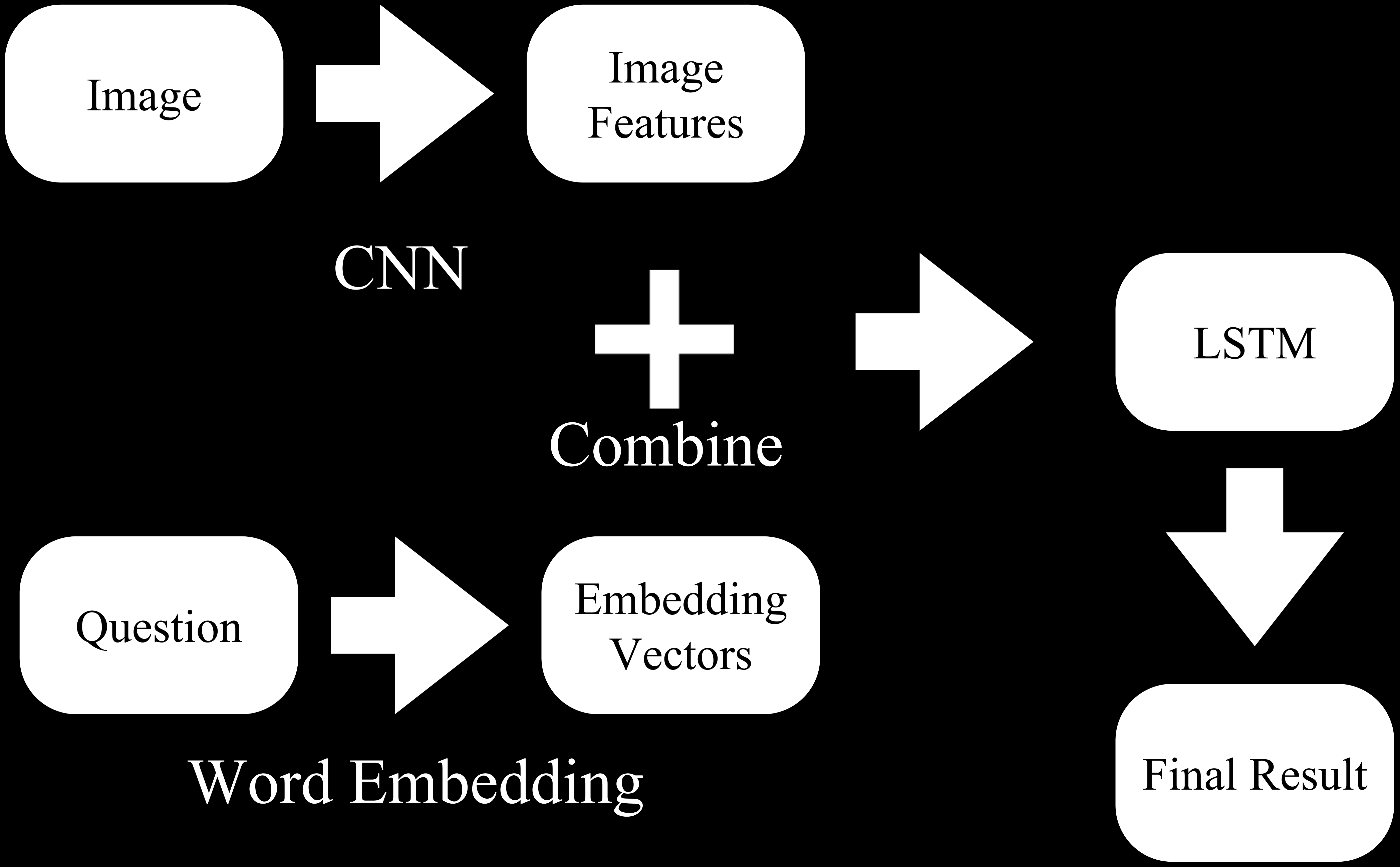


Computer Vision



How to Achieve?

Overall Architecture



Image

CNN

Image Features

Combine

Question

Word Embedding

Embedding Vectors

LSTM

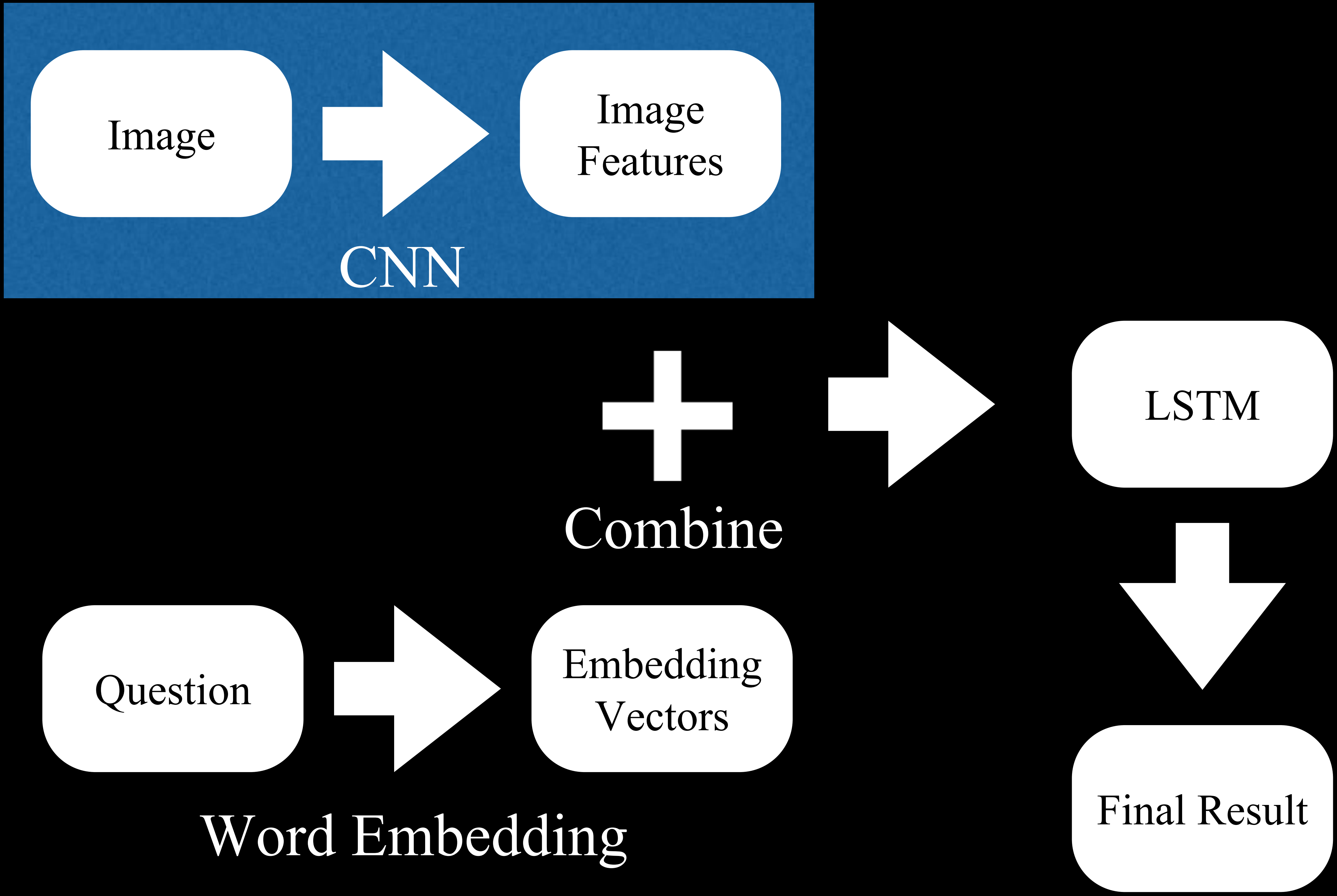
Final Result

- Image Processing
  - CNN (Convolutional Neural Network)

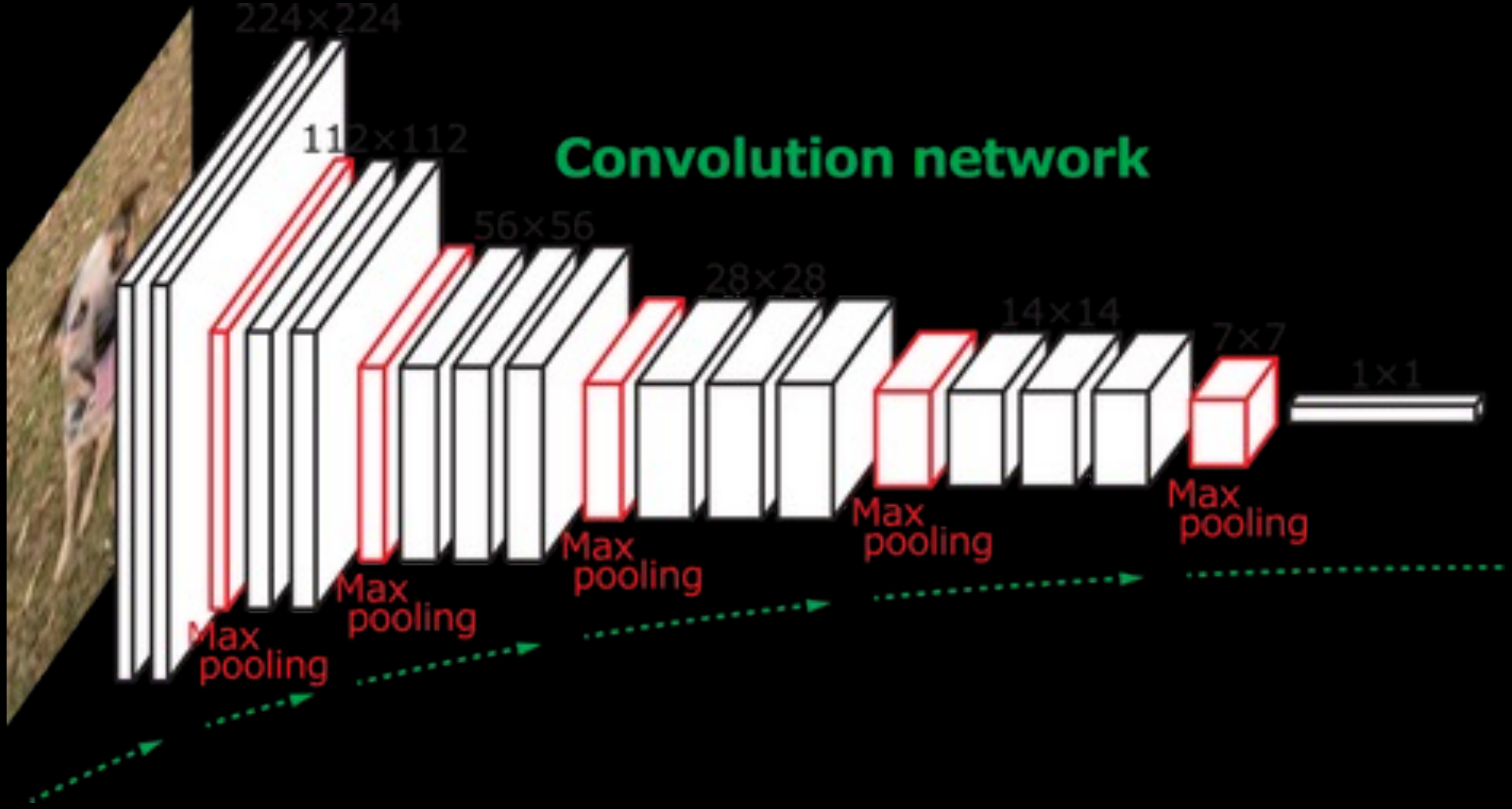
- Text Processing
  - Work Embedding

- LSTM (Long Short-Term Memory)

Overall Architecture

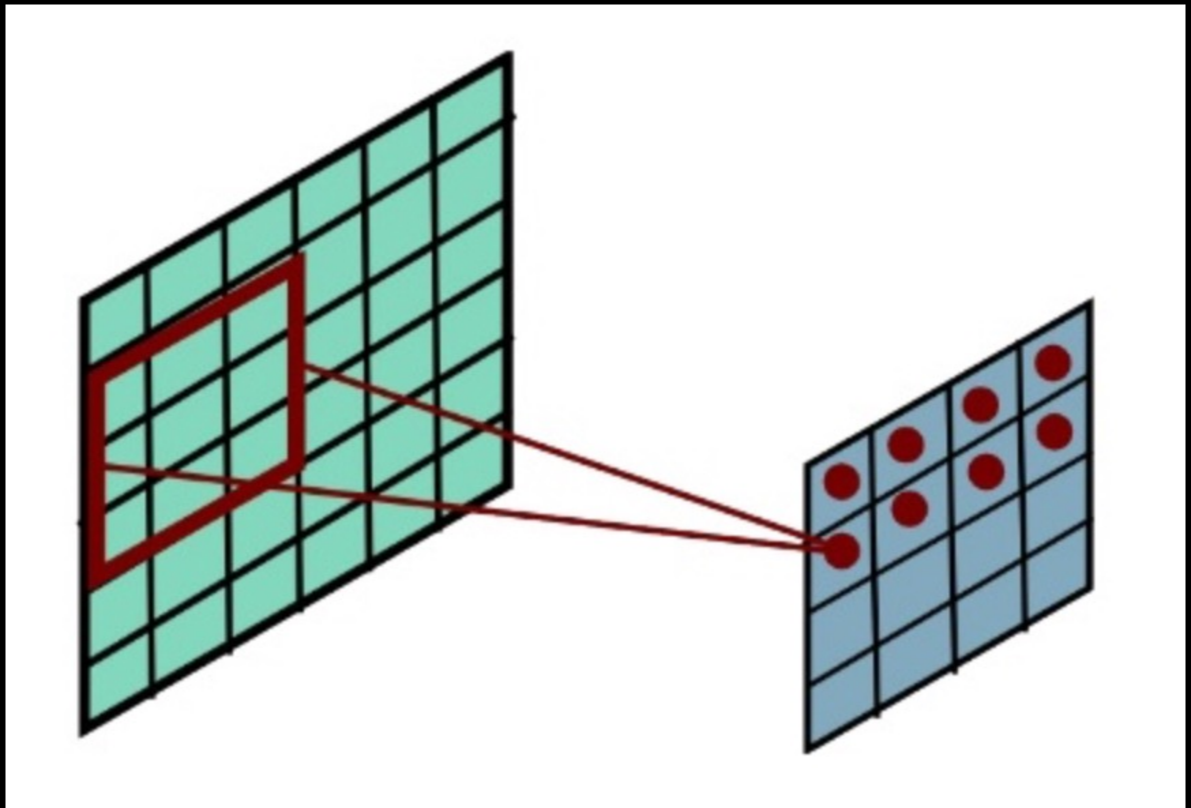


CNN



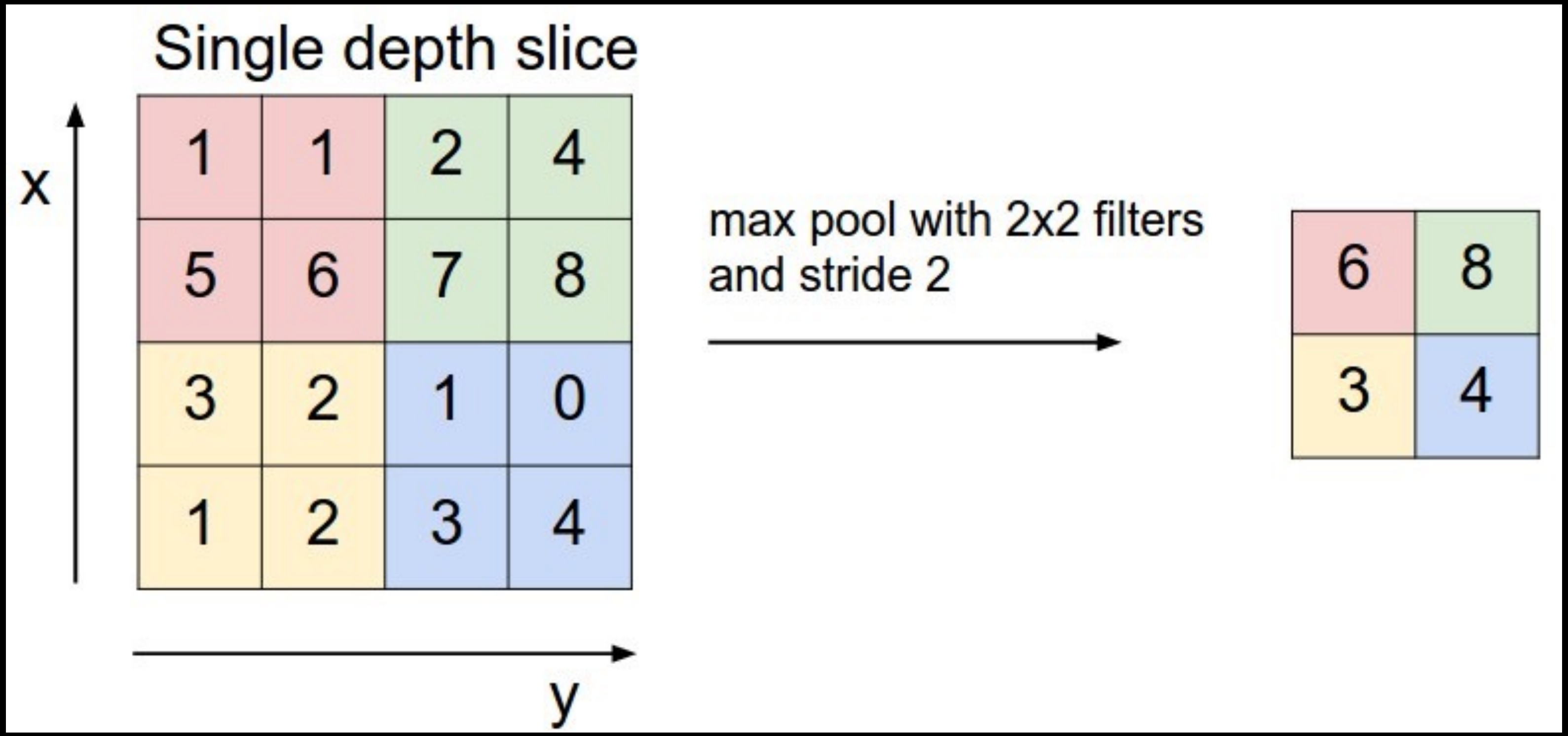
VGG-16 Model

Convolutional Layer



By doing convolution, we can extract features from input.

Pooling Layer

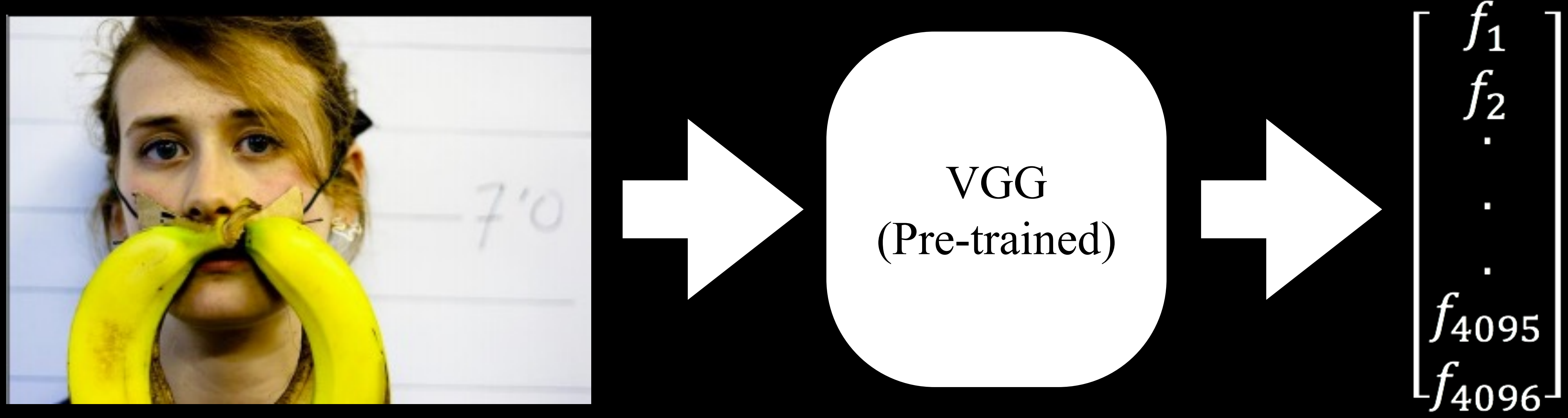


1	1	2	4
5	6	7	8
3	2	1	0
1	2	3	4

6	8
3	4

By doing pooling, Multiple value → One value
Dimension Reduced

Our Implementation



VGG
(Pre-trained)

$$\begin{bmatrix} f_1 \\ f_2 \\ \cdot \\ \cdot \\ \cdot \\ f_{4095} \\ f_{4096} \end{bmatrix}$$

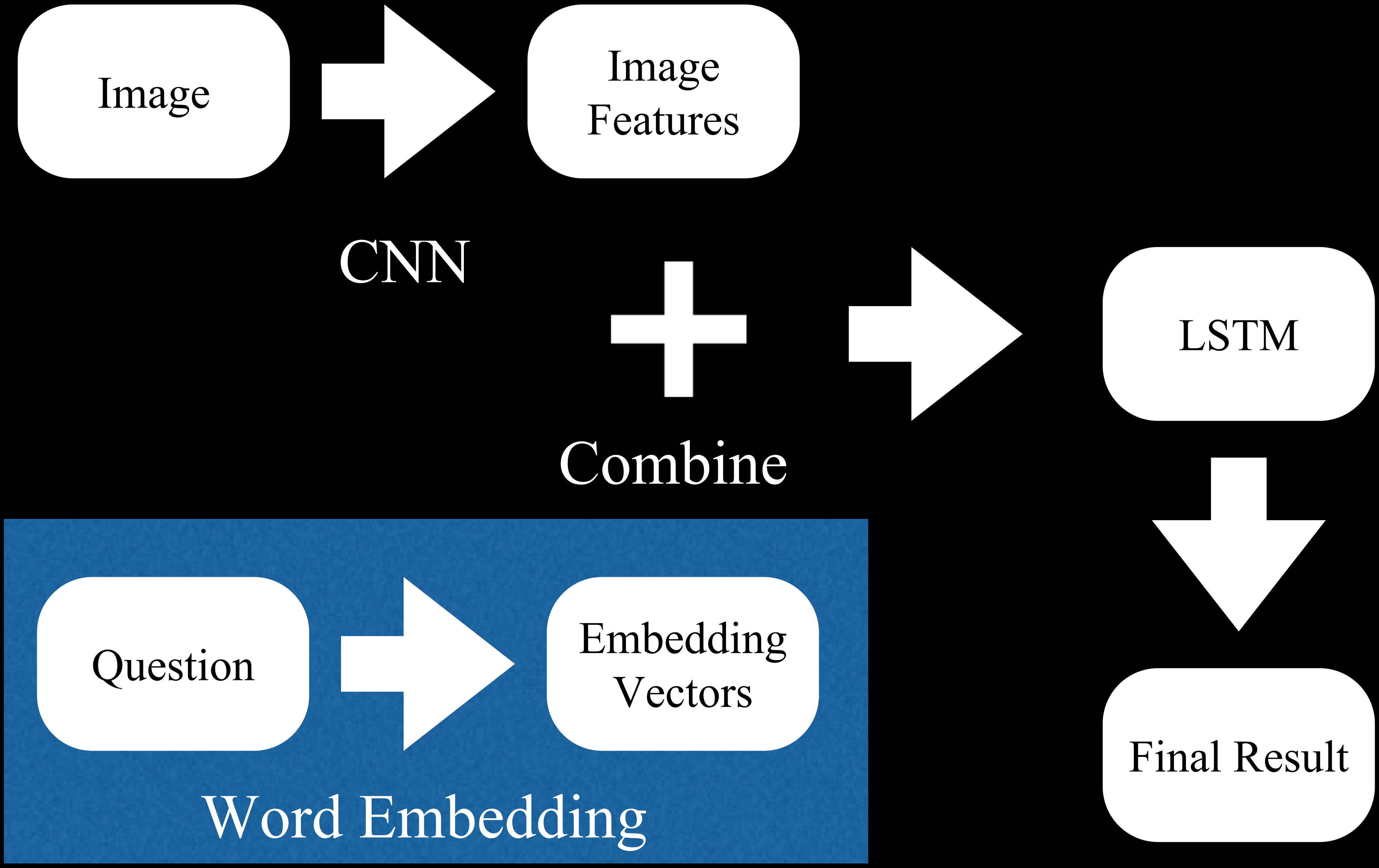
(4096, 1)

Image Features

$$\begin{bmatrix} f_{1,1} & & f_{N,1} \\ & \cdots & \\ f_{1,2} & & f_{N,2} \\ \vdots & \ddots & \vdots \\ f_{1,4095} & & f_{N,4095} \\ & \cdots & \\ f_{1,4096} & & f_{N,4096} \end{bmatrix}$$

(4096, N)
N is the total number
of train images

Overall Architecture



Image

CNN

Image Features

Combine

LSTM

Question

Embedding Vectors

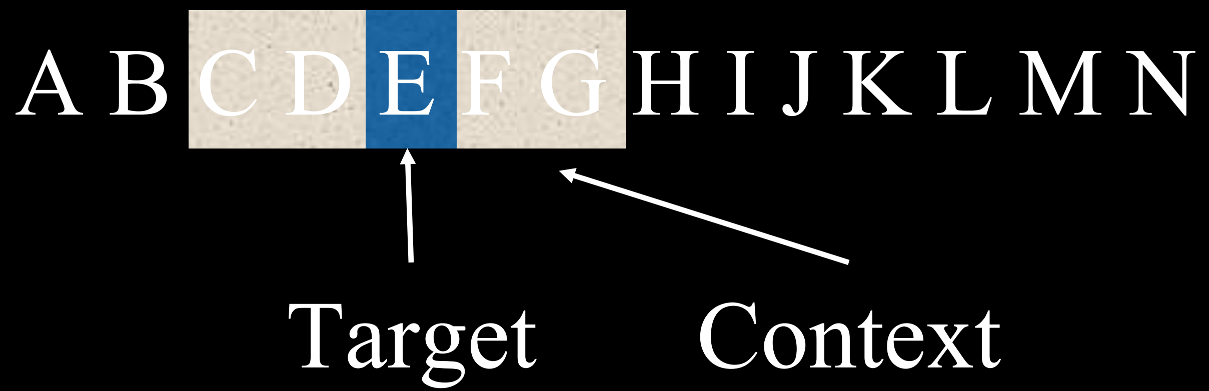
Word Embedding

Final Result

Word Embedding and Word2Vec

Word Embedding: Word → Vector

Word2Vec:

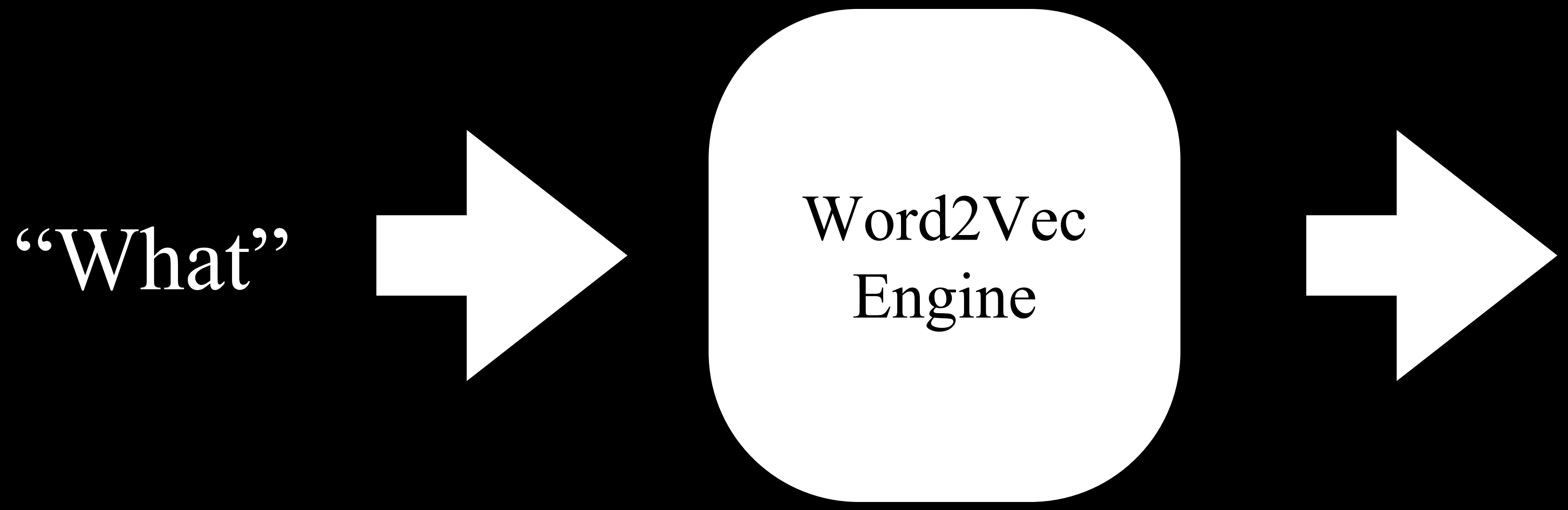


Keep on:

Moving Target and words in context closer and closer.

Moving Target and words outside context further and further.

Our Implementation



“What”

Word2Vec Engine

$$\begin{bmatrix} w_1 \\ w_2 \\ . \\ . \\ . \\ w_{511} \\ w_{512} \end{bmatrix}$$

(512, 1)

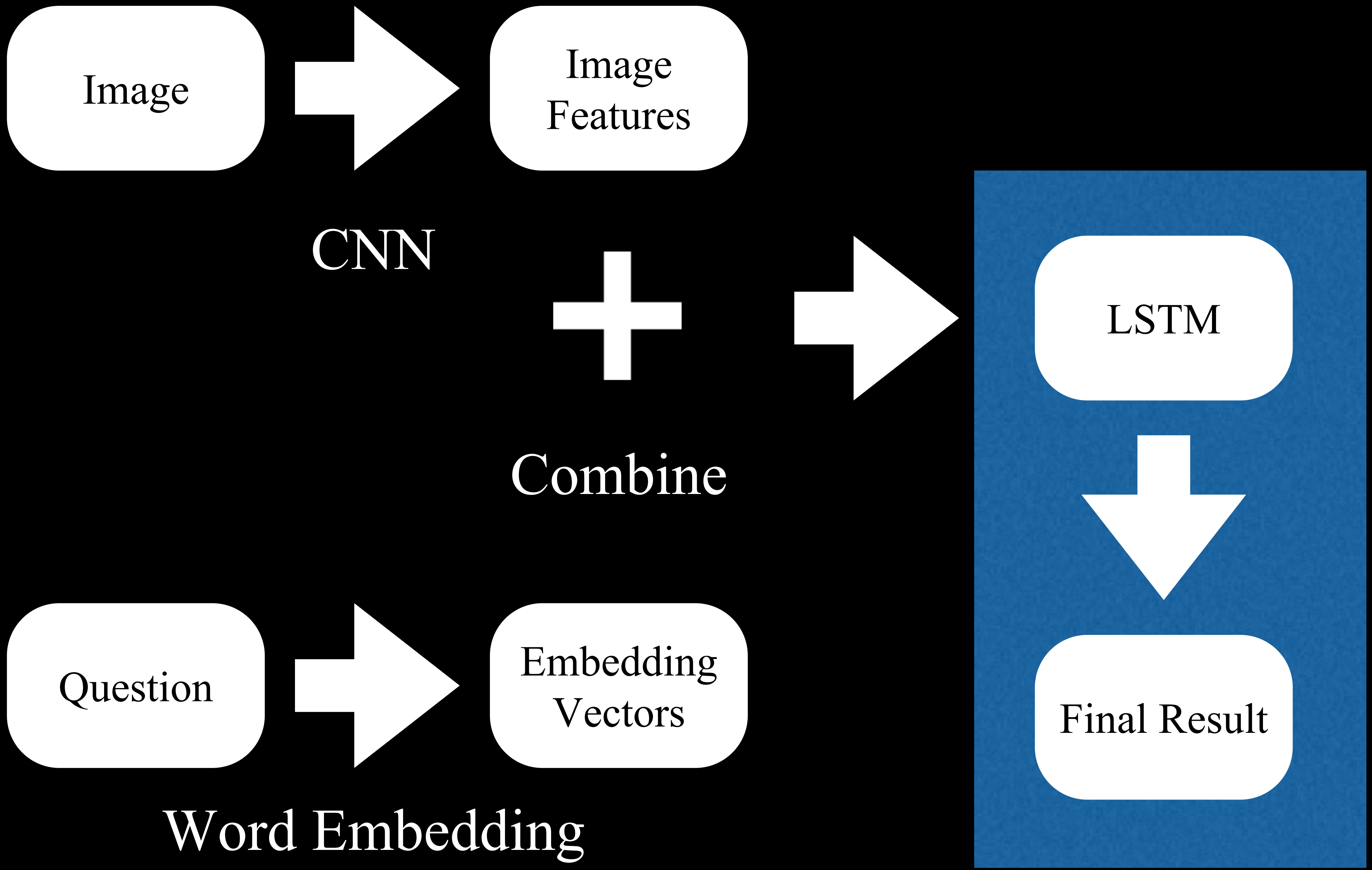
Embedding Vectors

$$\text{What is this?} \Rightarrow \begin{bmatrix} w_{1,1} & \cdots & w_{L,1} \\ w_{1,2} & & w_{L,2} \\ \vdots & \ddots & \vdots \\ w_{1,511} & & w_{L,511} \\ w_{1,512} & \cdots & w_{L,512} \end{bmatrix} \xrightarrow{\text{zero-padding}} \begin{bmatrix} w_{1,1} & \cdots & f_{S,1} \\ w_{1,2} & & f_{S,2} \\ \vdots & \ddots & \vdots \\ w_{1,511} & & f_{S,511} \\ w_{1,512} & \cdots & f_{S,512} \end{bmatrix}$$

(512, L)
L is the length of
the question.

(512, S)
S is the max length of
the question.

Overall Architecture



Image

CNN

Image Features

Question

Word Embedding

Embedding Vectors

Combine

LSTM

Final Result

Combining

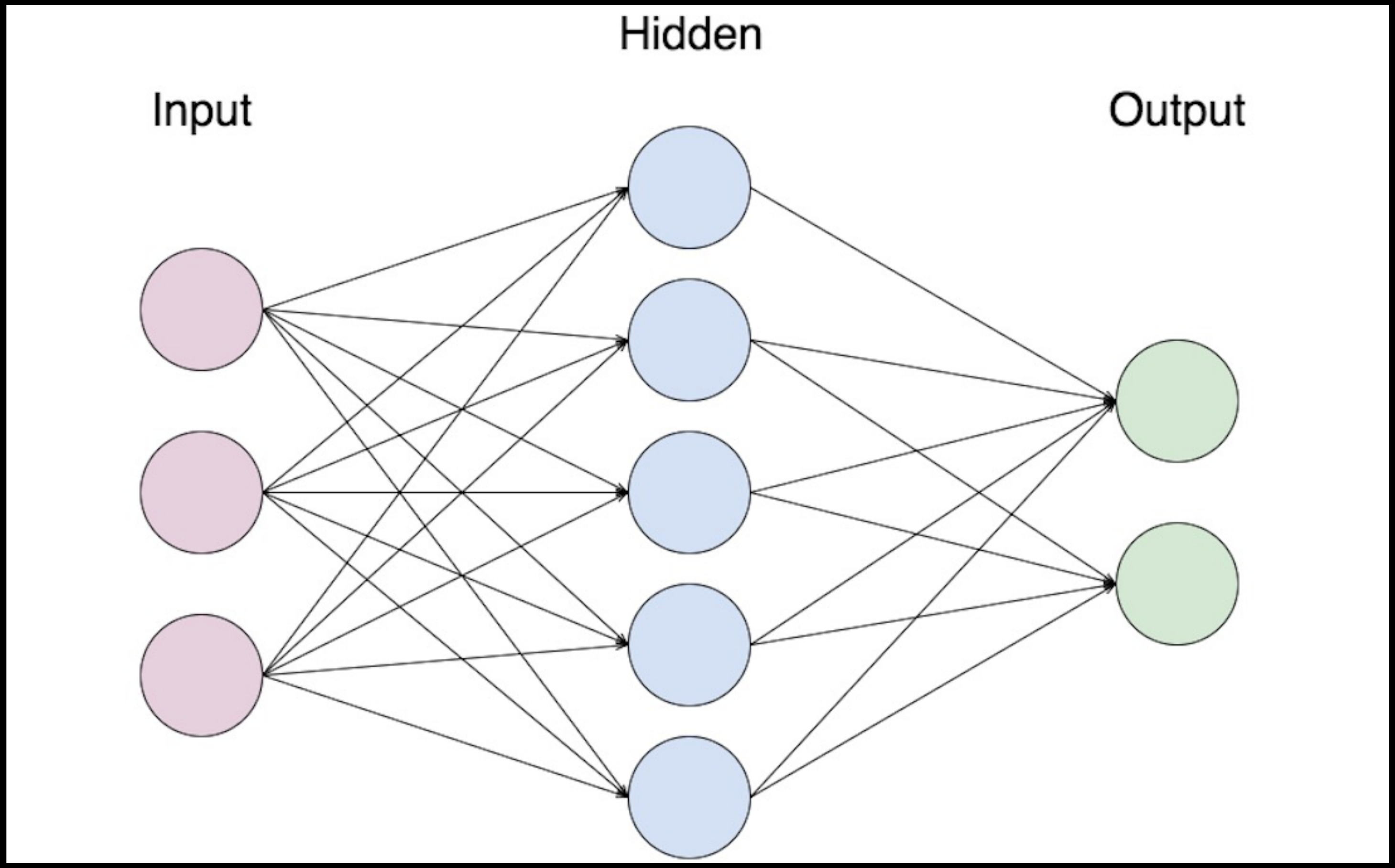
- Image Features
- (4096, 1)

- Words Vectors
- (512, S)

- Generate a matrix whose shape is (4096, 512) from a Truncated normal distribution
- Using this matrix to convert the Image Features into a matrix whose shape is (512, 1)
- Append these two matrix

LSTM

Traditional Structure



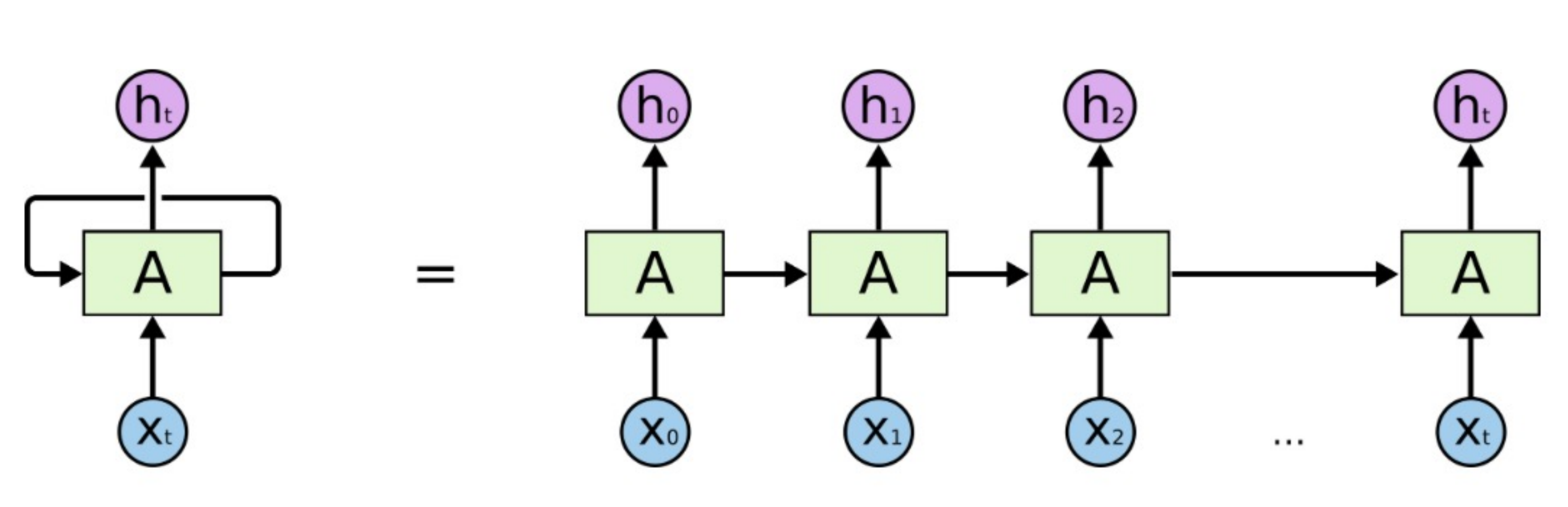
Full-Connected

&

No Connection Between Nodes in Some Layer

LSTM

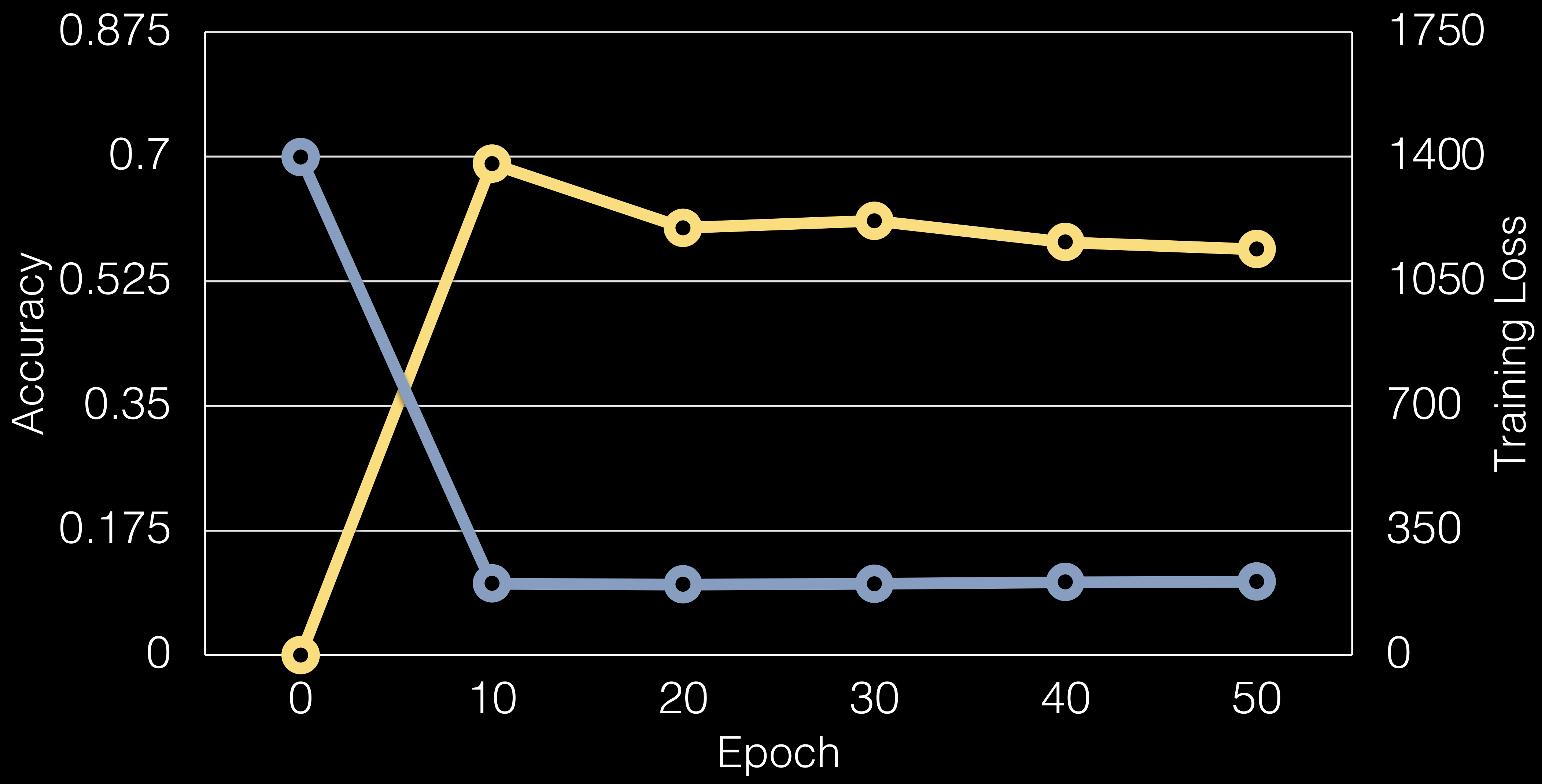
- Designed to solve "Long-Term dependencies" problem



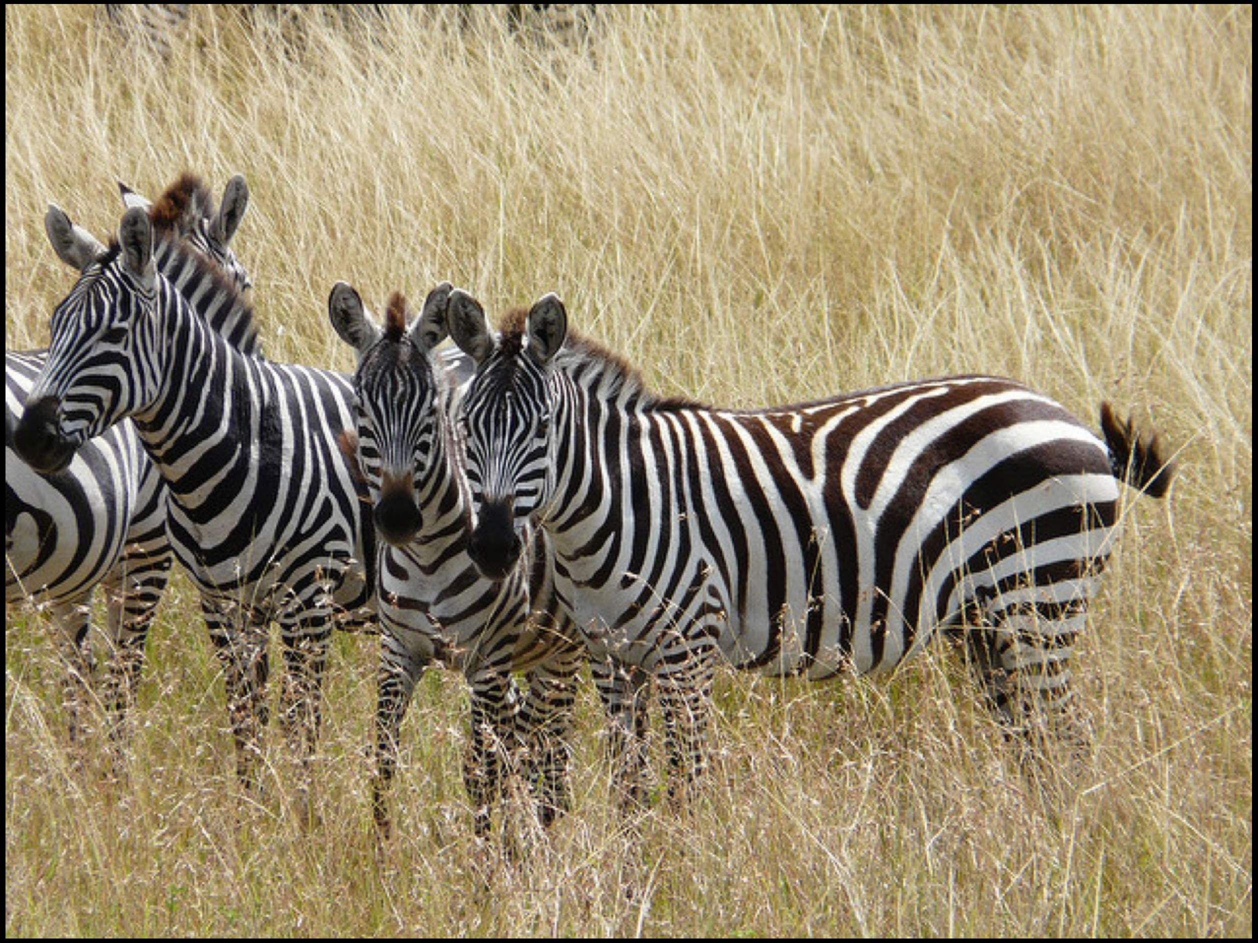
Structure Detail

- 512 * (S + 1) Nodes in Input Layer
- 2 Hidden Layer (first one has 512 * 4 node, second one has 512 node)
- Output Layer (Softmax)

Training Process



Example 1



What is this animal?

zebra, giraffe, horse, cow, zebras

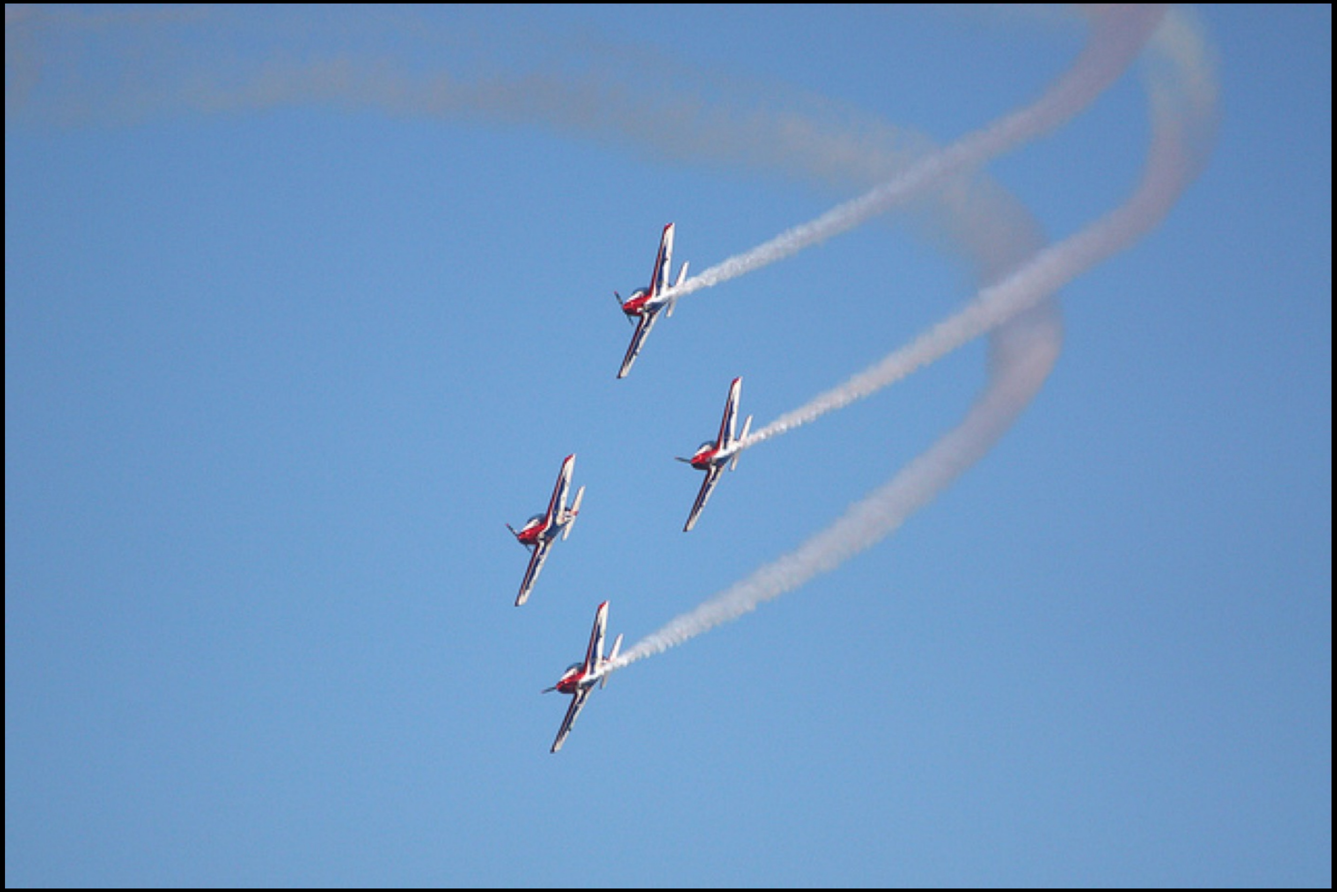
How many animals are there?

2, 3, 4, 1, 5

What is the color of this animal?

black and white, white, brown, black, gray

Example 2



What are flying through the sky?

kites, plane, kite, clouds, airplane

What is the color of background?

blue, red, green, orange, yellow

How many objects in the sky?

13, 10, 4, 5, 1

Accuracy

Yes/No	Number	Other	Overall
74.62	31.76	31.32	49.12

1. The overall accuracy is not very high, only 49.12%.
2. The accuracy on number-related question is very low, this model is not good at counting.
3. The accuracy on Yes/No question is relatively high, this model is good at classification.

Key: Convolutional Neural Network

Future Work

- Improve the accuracy of our model, especially the accuracy of number-related question. (Using R-CNN)

- Extend the question to not only in English but also in Chinese.

Thank You