Visual Question Answering with Deep Learning

Review

What is the mustache made of

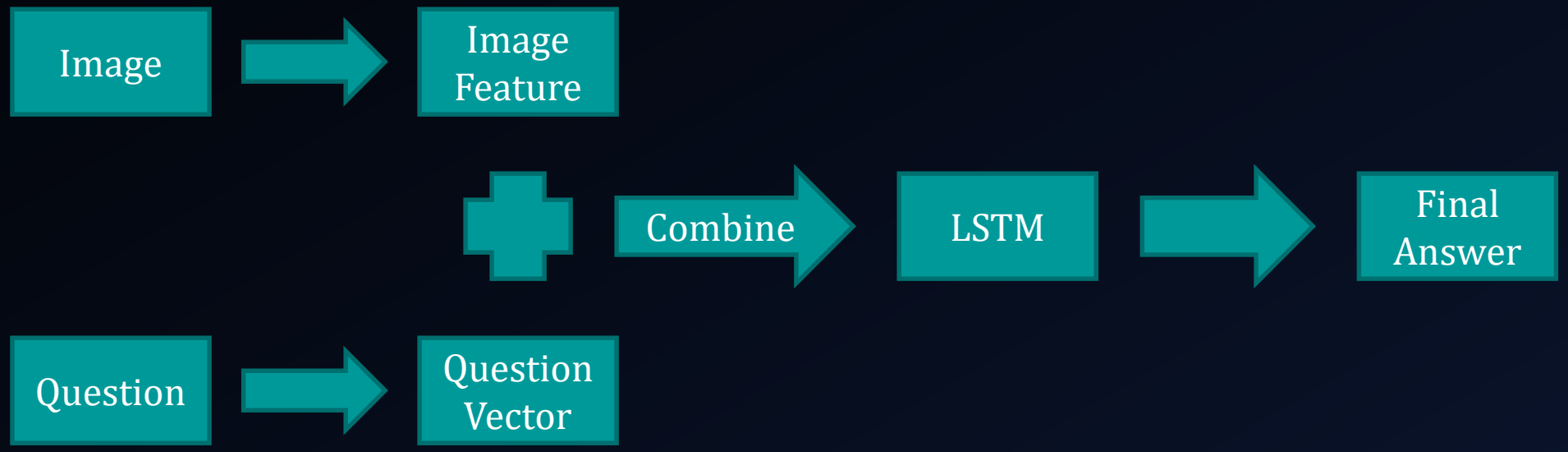
AI System

Bananas

New Features

- Model
  - Naïve Method => Neural Reasoner Based Model
- Language
  - English => French
- Content
  - Image => Video

Previous Model



Neural Reasoner Based Model

NEURAL REASONER

- Neural Reasoner is a framework for neural network-based reasoning over natural language sentences
- Reasoning is widely used in natural language processing tasks
- Architecture
  - Encoder Layer
  - Reasoning layer
  - Answering Layer

Neural Reasoner Based Model

NATURAL LANGUAGE PROCESSING

- $Q \xrightarrow{encode} q^{(0)}, F_k \xrightarrow{encode} f_k^{(0)}, k = 1,2, \dots, K.$
- $\left\{q^{(l)}, f_1^{(l)} \dots, f_K^{(l)}\right\} \xrightarrow{reasoning} \left\{q^{(l+1)}, f_1^{(l+1)} \dots, f_K^{(l+1)}\right\}$

Neural Reasoner Based Model

REASONING

- Question-Fact Iteration
  - $\left[q_k^{(l)}, f_k^{(l)}\right] \stackrel{\text{def}}{=} gDNN_l\left(\left[\left(q^{(l-1)}\right)^T, f_k^{(l-1)^T}\right]^T; \Theta_l\right)$
- Pooling
  - Max/Average Pooling
  - Gating
  - Model-Based (CNN / RNN)

Neural Reasoner Based Model

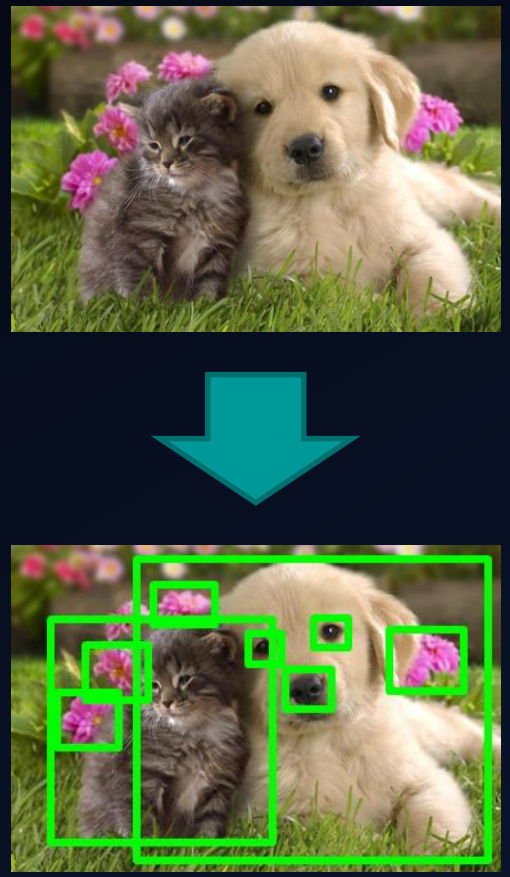
IMAGE PROCESSING

- $Input\ Image \xrightarrow{Object\ Localization} F_k, k = 1,2, \dots, K.$
- $Q \xrightarrow{encode} q^{(0)}, F_k \xrightarrow{encode} f_k^{(0)}, k = 1,2, \dots, K.$
- $\left\{q^{(l)}, f_1^{(l)} \dots, f_K^{(l)}\right\} \xrightarrow{reasoning} \left\{q^{(l+1)}, f_1^{(l+1)} \dots, f_K^{(l+1)}\right\}$

Neural Reasoner Based Model

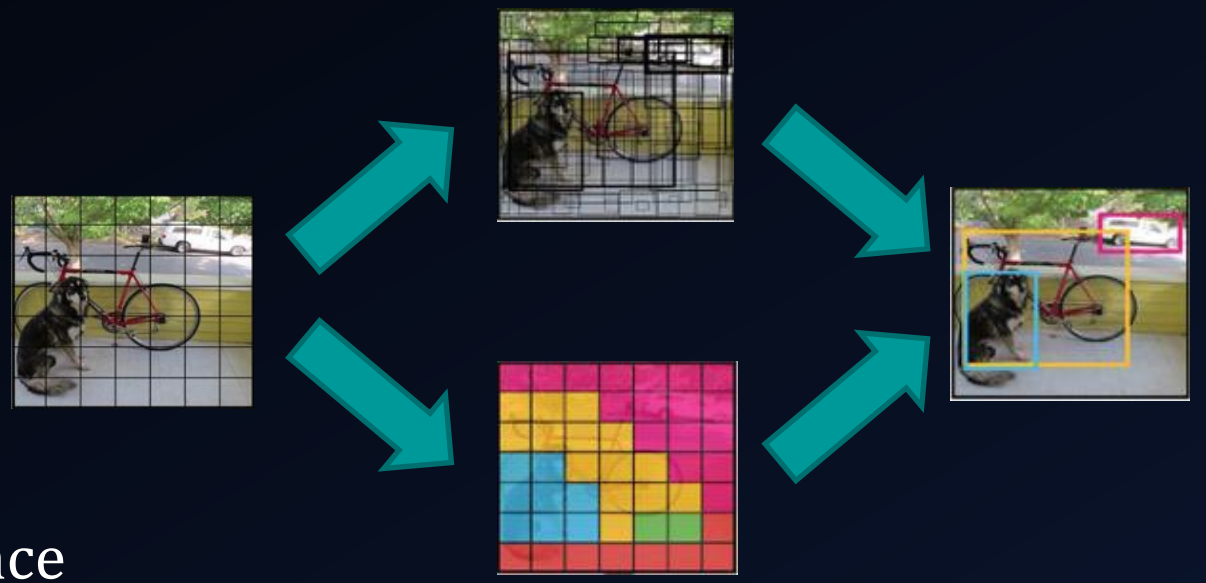
OBJECT LOCALIZATION

- BING ++
- Edge Boxes
- Objectness
- YOLO Darknet ✓



Neural Reasoner Based Model

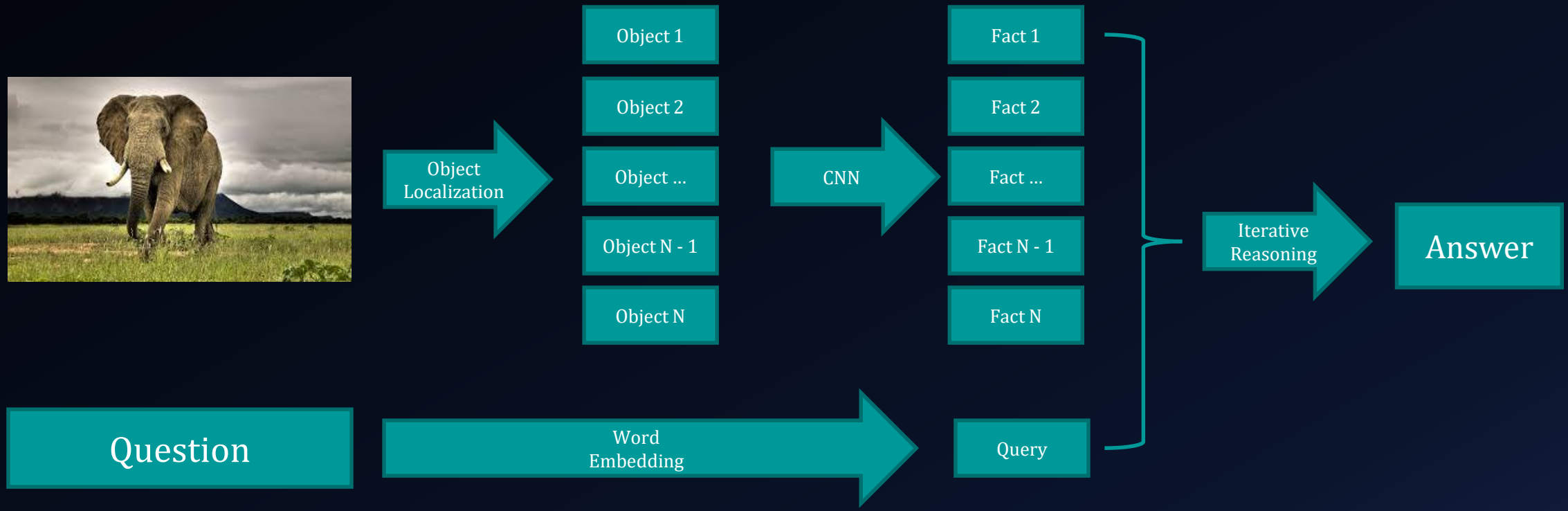
YOLO



- Divide image into S * S grid
- Within each grid cell predict:
  - B Boxes: 4 coordinates + confidence
  - Class scores: C numbers
- Regression from image to 7 * 7 * (5 * B + C) tensor
- Direct prediction using a CNN

Neural Reasoner Based Model

OVERALL STRUCTURE



French Support

FEATURES:

- Support French Q&A (French Question and French Answer)
- Not need French VQA dataset for training.

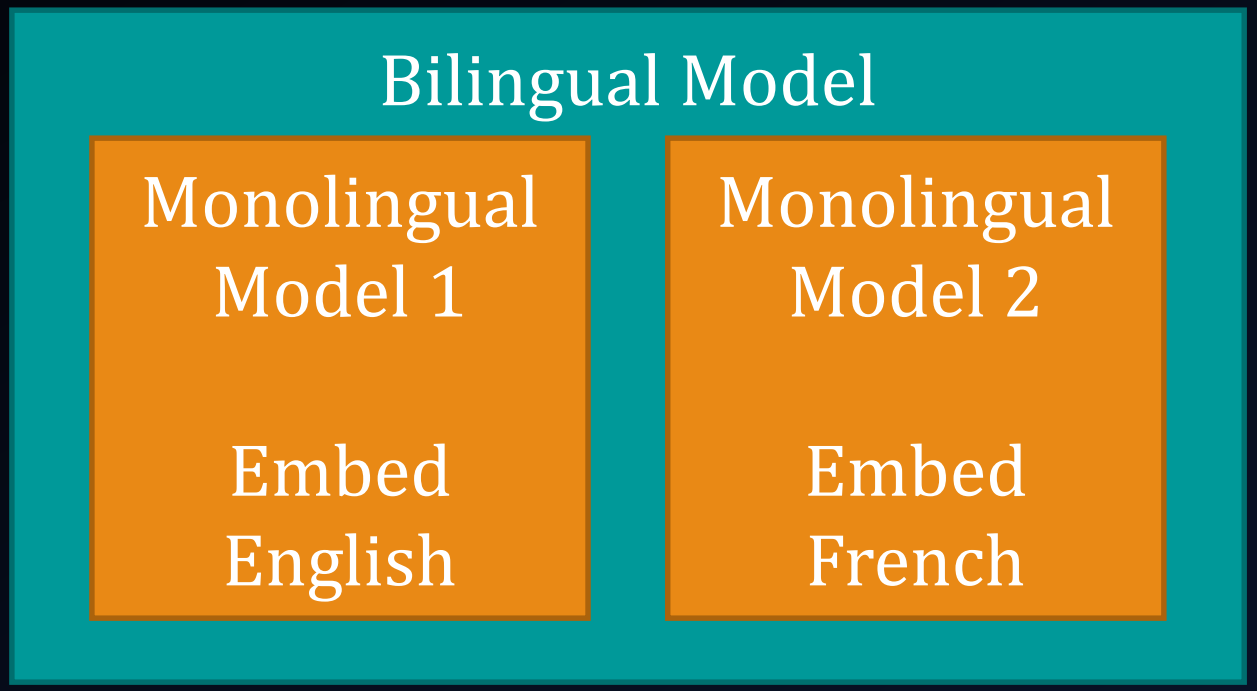
French Support

BASIC IDEA:

- Key component: Bilingual Model
- Described in paper: *Bilingual word representations with monolingual quality in mind*.
- By using Bilingual Model, we can achieve the equivalence between English and French.

French Support

BILINGUAL MODEL STRUCTURE



These two monolingual models embed words in two languages with same meaning into 2 vectors that are close.

For example:
“Cat” (in English)
“Chat” (in French)

French Support

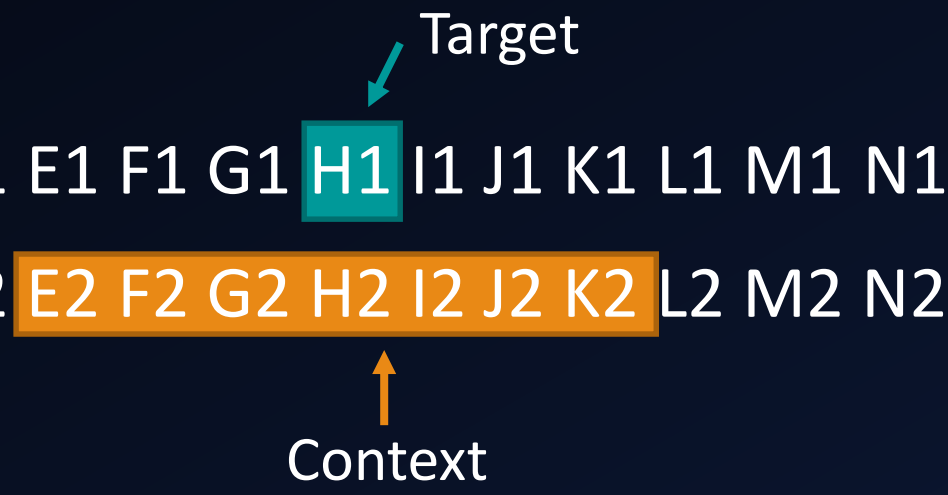
BILINGUAL MODEL TRAINING

- Step 1: Training two monolingual models separately using Skip-Gram.
- Keep on
  - Moving words in context closer and closer.
  - Moving words outside context further and further.

French Support

BILINGUAL MODEL TRAINING

- Step 2: Training two monolingual model together using Biskip-Gram.
- Biskip-Gram is based on the idea of Skip-Gram.



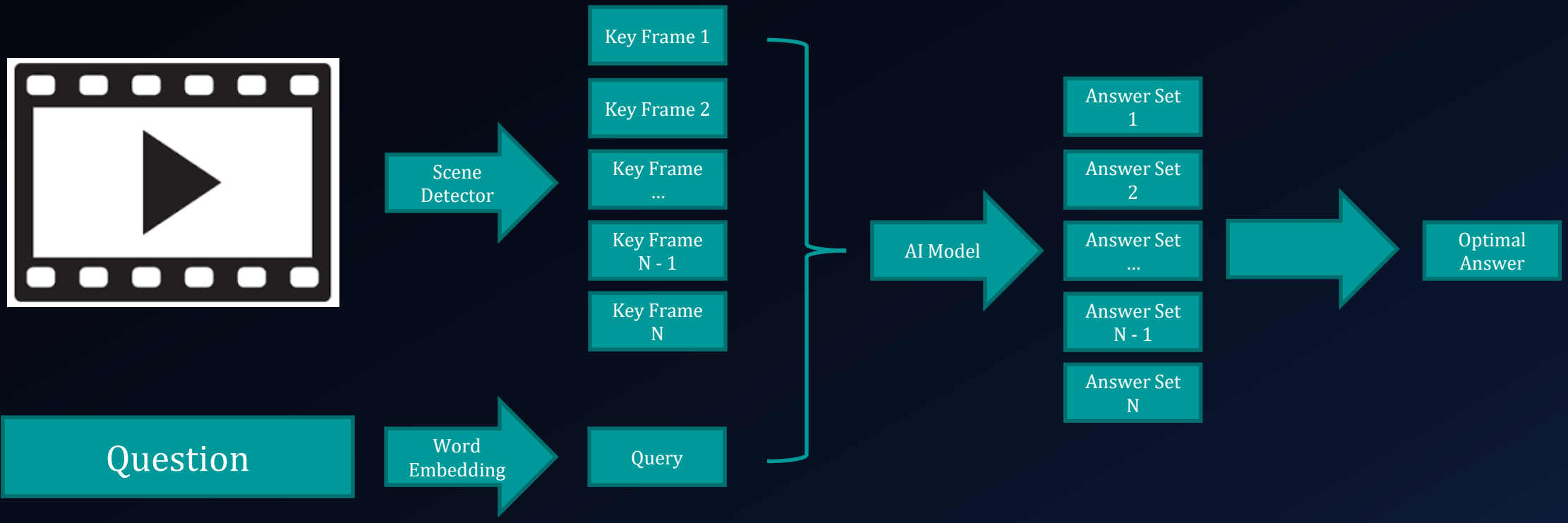
French Support

TRAINING & PREDICTING

- Training: Based on the properties of bilingual model, we can consider French words and English words as the similar vectors. Therefore, we only need English-version dataset for training.

Video Question Answering

METHODOLOGY



Video Question Answering

VIDEO SCENE DETECTOR

- Content-Aware Detector
- Threshold Detector

Video Question Answering

CONTENT-AWARE DETECTOR

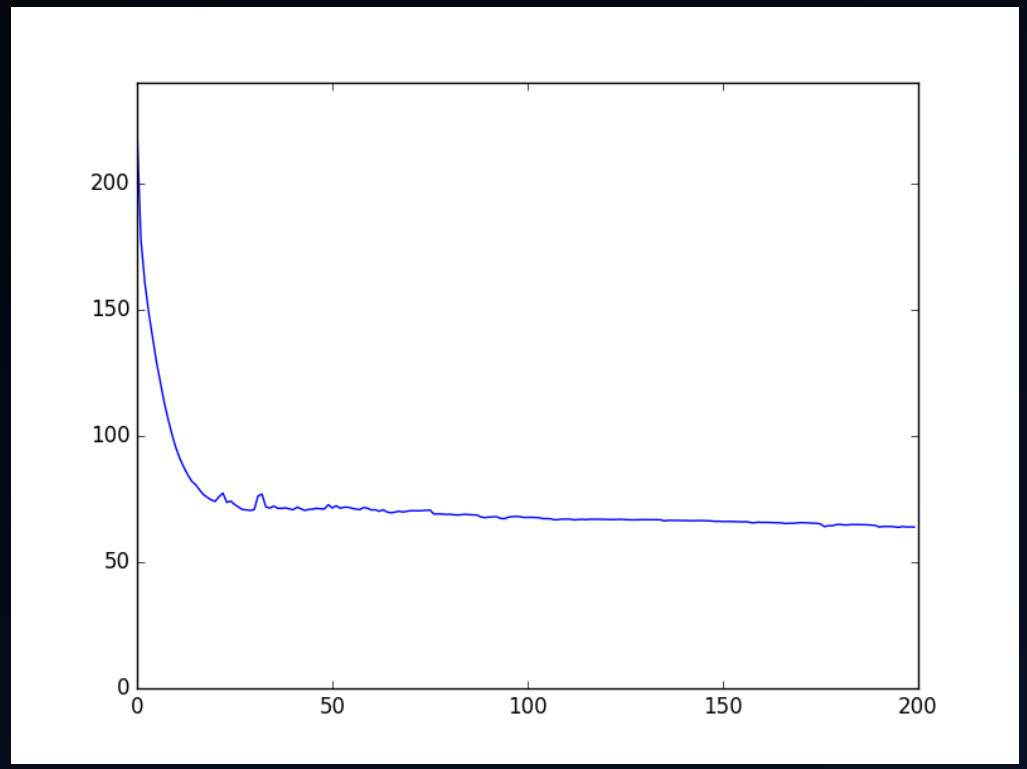
- The content-aware scene detector finds areas where the difference between two subsequent frames exceeds the threshold value that is set

THRESHOLD DETECTOR

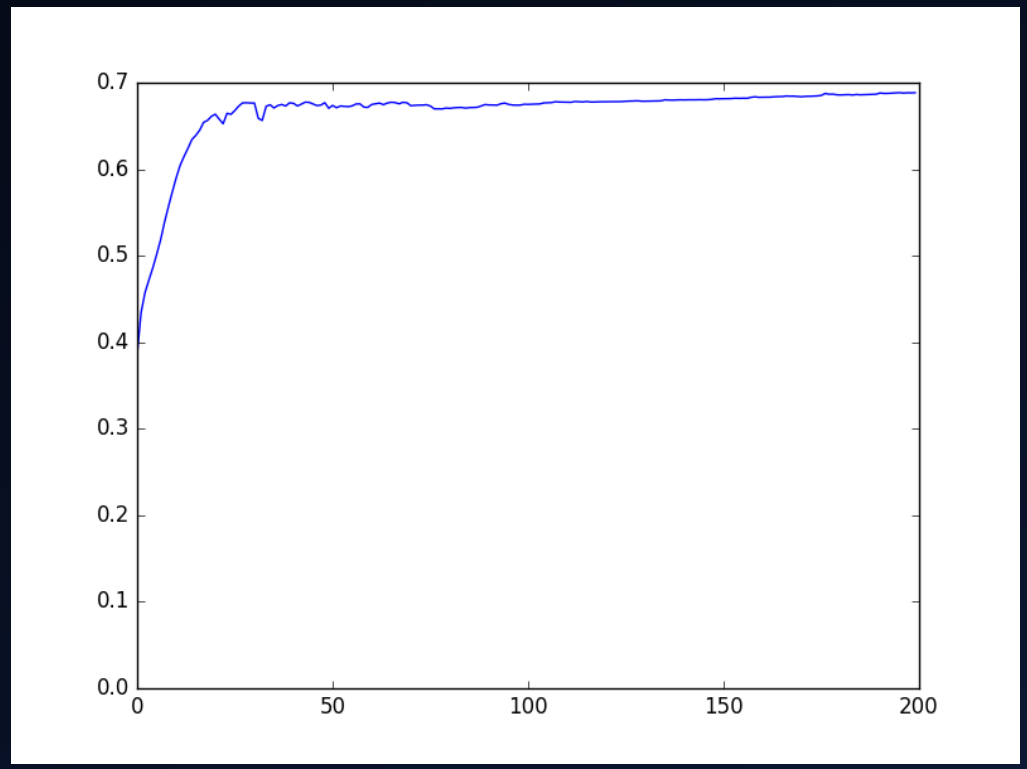
- The threshold-based scene detector compares the intensity/brightness of the current frame with a set threshold, and triggering a scene cut/break when this value crosses the threshold.

Training

NEURAL REASONER BASED MODEL



Loss



Accuracy

Result

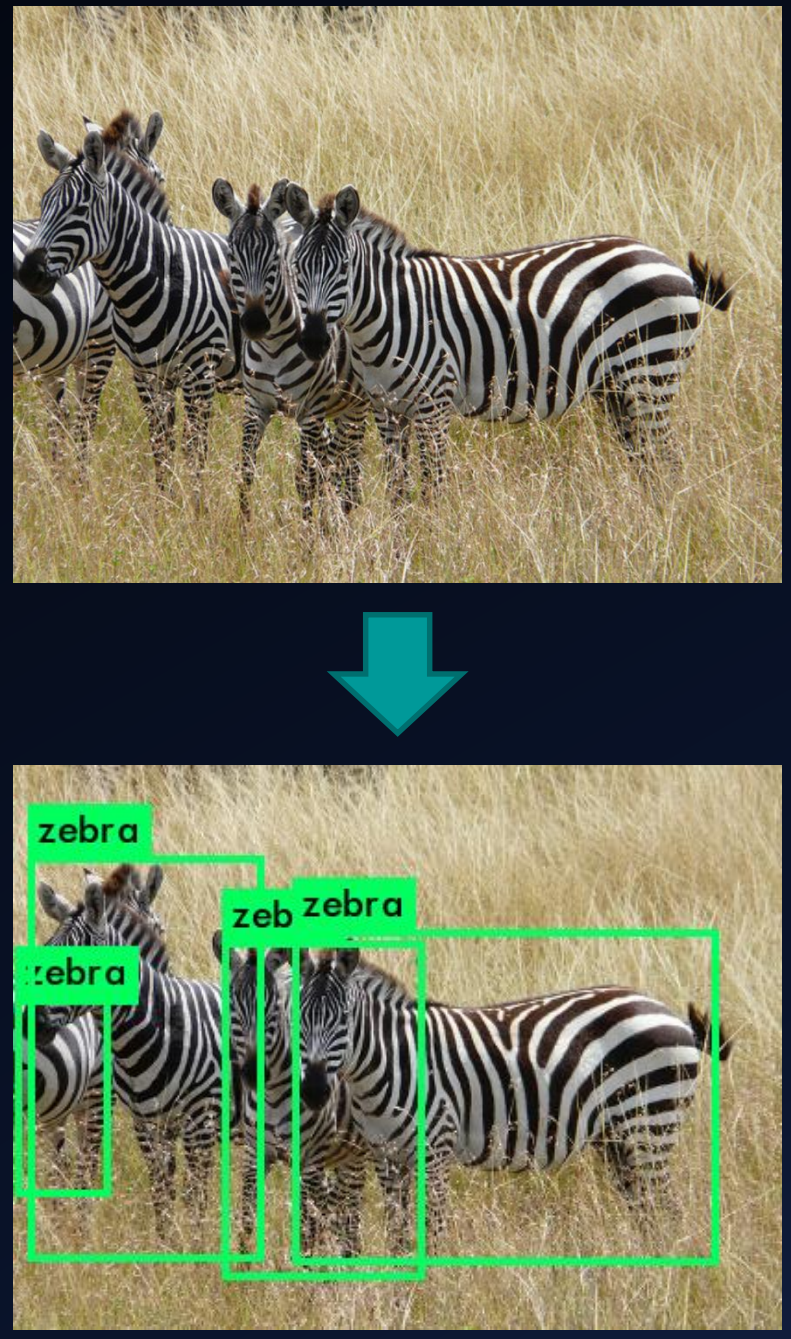
NEURAL REASONER BASED MODEL

- Accuracy

	First Semester	Second Semester
Yes/No	74.62	80.62
Number	31.76	31.78
Other	31.32	40.02
Overall	45.87	48.53

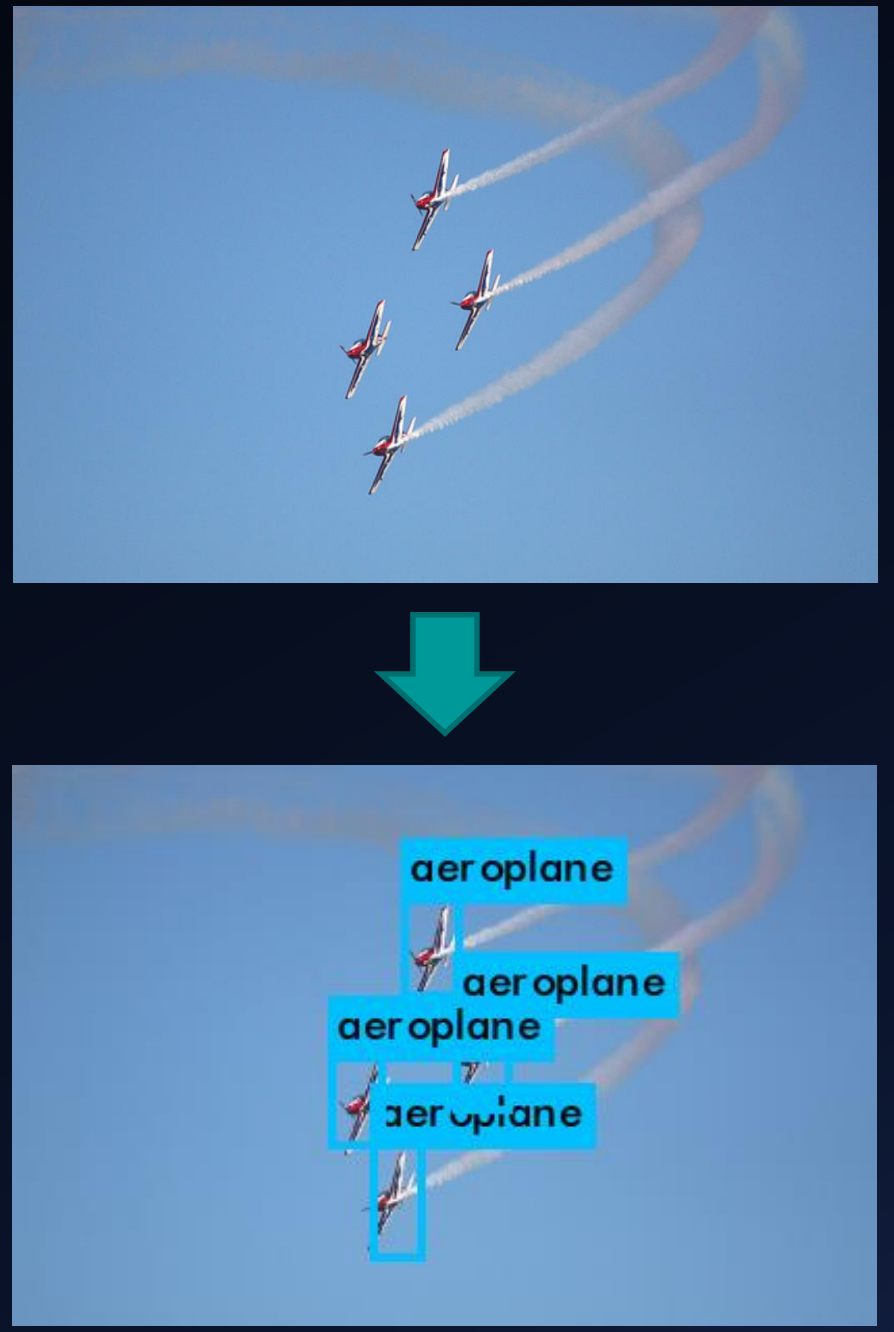
Result

- First Semester
  - What is this animal?
    - zebra, giraffe, horse, cow, zebras
  - How many animals are there?
    - 2, 3, 4, 1, 5
- Second Semester
  - What is this animal?
    - zebra, zebras, giraffe, cow, horse
  - How many animals are there?
    - 4, 3, 2, 1, 6



Result

- First Semester
  - What are flying through the sky?
    - kites, plane, kite, clouds, airplane
  - How many objects in the sky?
    - 13, 10, 4, 5, 1
- Second Semester
  - What are flying through the sky?
    - plane, airplane, kites, kite, clouds
  - How many objects in the sky?
    - 4, 5, 2, 3, 1



Result

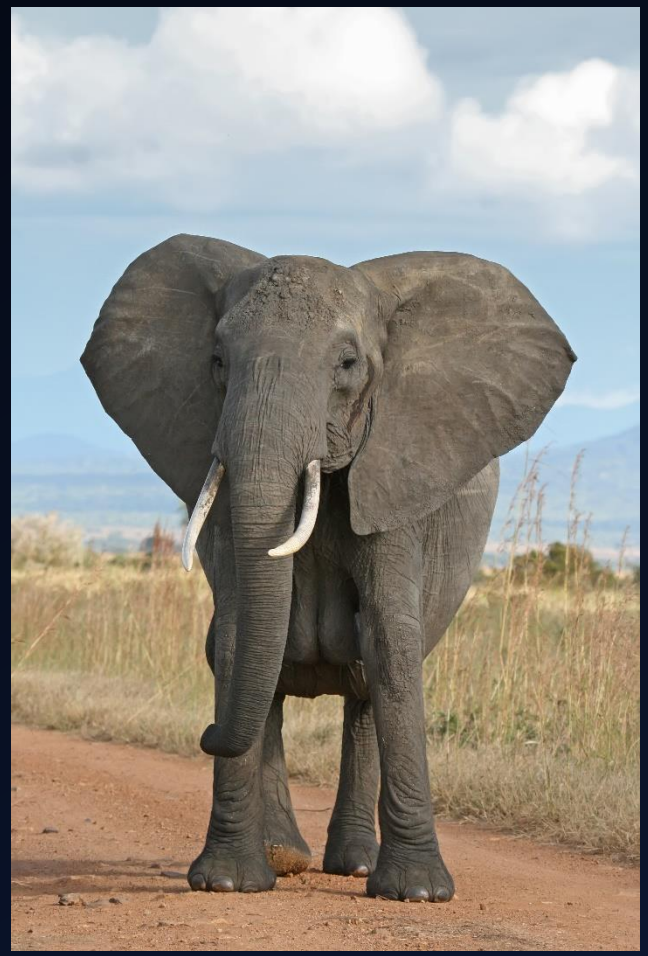
FRENCH Q&A

- *English:*
  - What is this boy doing?
  - Eating.
- *French:*
  - Qu’est-ce qu’il fait?
  - Manger

Result

FRENCH Q&A

- *English:*
  - What animal is this?
  - Elephant
- *French:*
  - Qu’est-ce que c’est animal?
  - l'éléphant



Result

VIDEO QUESTION ANSWERING

Result

VIDEO QUESTION ANSWERING

- Which animal is this?
  - Dog
- What are they doing?
  - Posing

Discussion

NEURAL REASONER BASED MODEL

- Positive
  - Neural Reasoner
- Negative
  - Object localization algorithms
  - Training set

Discussion

VIDEO QUESTION ANSWERING

- Regards each frame as individual (Ignore the relation between frames and frames)
  - Motions
  - Actions
- Key frames to represent the whole video (Lose information)

Discussion

FRENCH SUPPORT

- Although we can find some correct answers in our French Q&A, the most of answers are not correct. After analyzing the model, we think there are two reasons causing this problem.

Discussion

FRENCH SUPPORT

- Reason 1: Semantic structures of French and English are not the same.
- What animal is this
- Qu’est-ce que c’est animal

Discussion

FRENCH SUPPORT

- Reason 2: The bilingual model is not accurate enough.
- v1 = English_word_to_vec(“cat”)
- v2 = French_word_to_vec(“chat”)
- |v1 – v2| / |v1| = 1.113
- |v1 – v2| / |v2| = 0.818

Thank You