# Compréhension multimodale du langage à travers le corpus *GuessWhat?*

Ibrahim Souleiman

Parcours IAAA
Master Informatique
Université d'Aix Marseille

5 Septembre 2019

# Contexte : Compréhension multimodale du langage

- Combinaison texte + image
  - document textuel avec illustrations
  - **dialogue (écrit/parlé) à propos d'une image**
- Compréhension multimodale
  - Récupère des informations complémentaires
  - **Amélioration de la prédiction dans le cas ou une même information est présente dans des différents type de donnees**
- Problématique
  - Comment faire une analyse jointe texte/image ?
  - Est-ce que l'ajout des données supplémentaires permettra d'améliorer le résultat ?

# Sujet du Stage

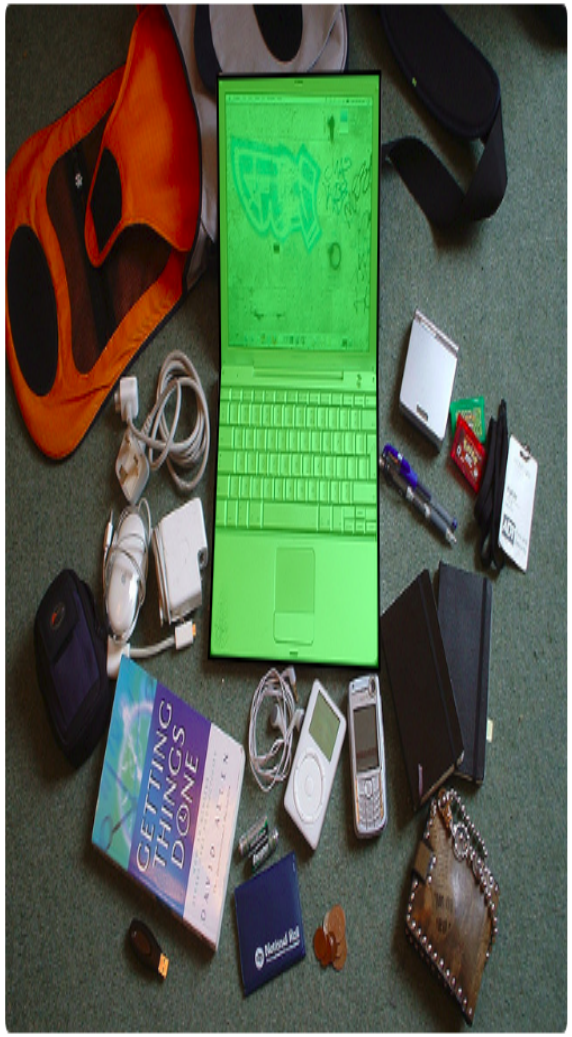
- Compréhension multimodale
  - Etudier cette problématique sur le corpus *GuessWhat?*
- Corpus *GuessWhat?*
  - Corpus de dialogue entre un *Oracle* et un *Guesser* à propos d'une image
  - *GuessWhat?! Visual object discovery through multi-modal dialogue, (H de Vries et al.)*
- L'intérêt du Corpus *GuessWhat?*

# Lieu de Stage

- L'equipe *TALEP*
  - Traitement Automatique du Langage Ecrit et Parlé
- Travaux de l'equipe *TALEP*
  - Analyse syntaxique
  - Expressions polylexicales
  - Annotations linguistiques
  - **Traitement de données multimodales**
  - ...

# Traitement Automatique des Langues / Analyse d'image

- *Guesser*
  - Génération de questions
  - Gestion du dialogue




#29043

is it electronic? Yes

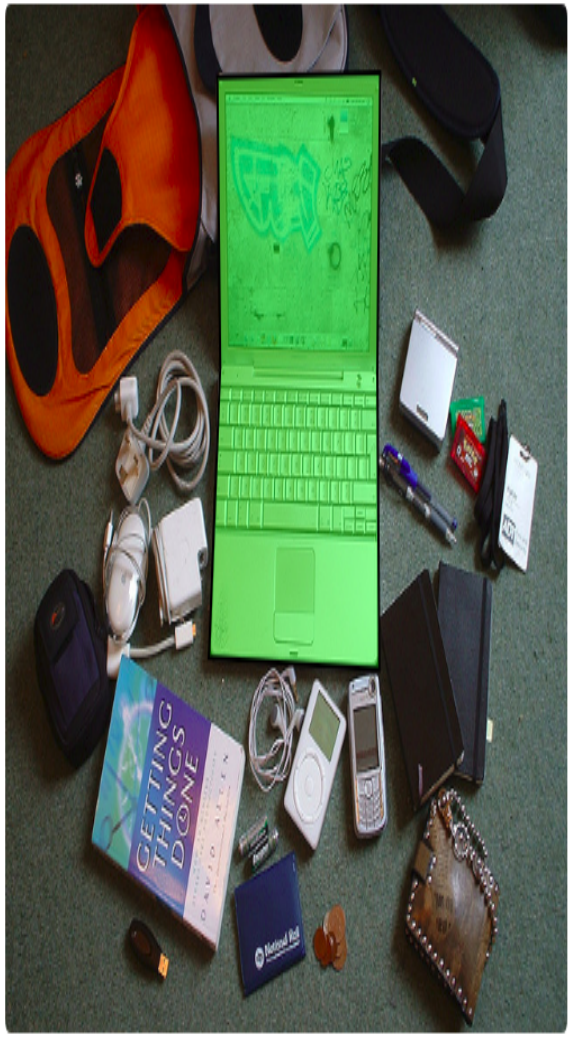
is it a charger? No

is it the laptop? Yes

Success

# Traitement Automatique des Langues / Analyse d'image

- *Guesser*
  - Génération de questions
  - Gestion du dialogue
- *Oracle*
  - Analyse/compréhension des questions
  - Analyse multimodale pour répondre *oui/non*




#29043

is it electronic? Yes

is it a charger? No

is it the laptop? Yes

Success

# Traitement Automatique des Langues / Analyse d'image

- *Guesser*
  - Génération de questions
  - Gestion du dialogue
- *Oracle*
  - Analyse/compréhension des questions
  - Analyse multimodale pour répondre *oui/non*

Ce Stage s'intéressera à l'amélioration de l'Oracle


#29043

is it electronic? Yes

is it a charger? No

is it the laptop? Yes

Success

# Le contenu du Corpus GuessWhat?!

# Quelques chiffres

| Type de données | Quantité |
|---|---|
| Image | 66 537 |
| Dialogue | 155 280 |
| Question / réponse | 821 889 |
| Taux Réponse (Non,Oui,N/A) | (5.2%,4.6%,2.2%) |
| Moyenne du nombre des questions par partie | 5.2 |
| Taille du vocabulaire | 3 986 192 mot |

Table 1: Description des données

# Ressources existantes

- Code disponible sur le git du projet
  - Réalisé avec le Framework Tensorflow
  - Documentation très partielle

- 3 Papiers
  1. La baseline de l'oracle et du Guesser
  2. L'amelioration de Guesser
  3. La multimodalité pour les données VQA

# Oracle - base de référence

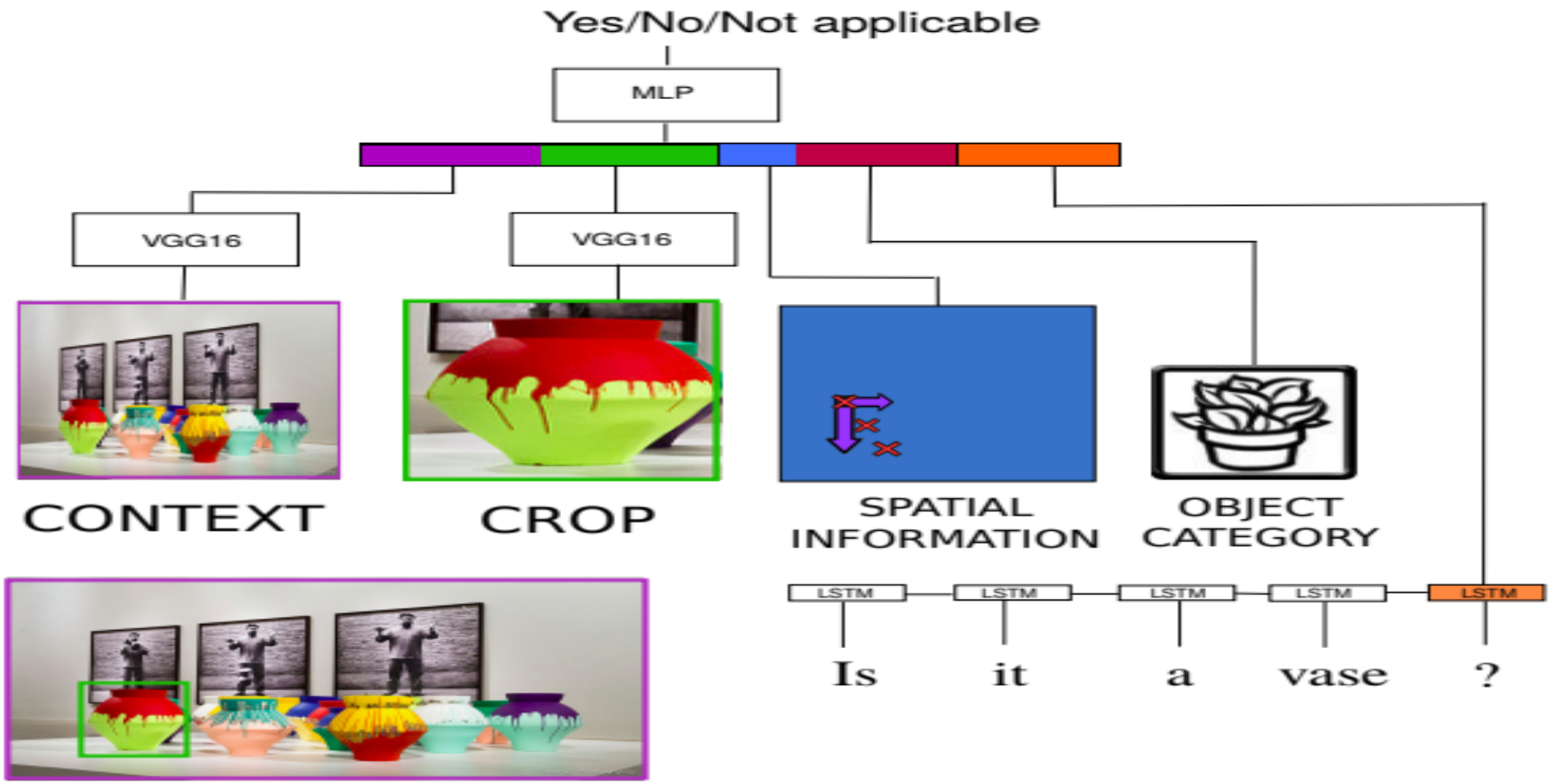



Figure 1: Base de référence de l'Oracle

# Traits utilisés par le modèle

- LSTM pour encoder les questions
  - Dimension d'embedding de 300
  - Cellule de type LSTM , avec 512 unités cachées
- Utilise un MLP pour faire la prédiction à partir de vecteurs qui contiennent tous les traits concaténés
  - Avec 512 unités cachées
- optimiseur adam
  - 20 époques
  - 64 taille lot (batch size)

| Modèle | Train err % | Val err% | Test err% |
|---|---|---|---|
| Question | 40.2 | 41.7 | 41.2 |
| Image | 45.7 | 46.7 | 46.7 |
| Crop | 40.9 | 42.7 | 43.0 |
| Category | 43.0 | 42.8 | 43.1 |
| Question + Crop | 22.3 | 29.1 | 29.2 |
| Question + Image | 37.9 | 40.2 | 39.8 |
| Question + Category | 23.2 | 25.8 | 25.7 |
| Question + Spatial | 28.1 | 31.2 | 31.1 |
| **Question + Category + Spatial** | **17.0** | **21.2** | **21.5** |
| Question + Category + Crop | 20.4 | 24.4 | 24.6 |
| Question + Spatial + Crop | 19.4 | 26.2 | 26.2 |
| Question + Category + Spatial + Crop | 16.1 | 21.7 | 22.1 |
| Question + Spatial + Crop + Image | 20.7 | 27.7 | 27.9 |

# Les défauts de la Baseline

- Image non exploité
- Côte text
  - Utiliser Simplement les questions .
  - Utiliser des embedding aléatoire pour les questions.

# Amélioration de l'Oracle

arXiv:1707.00683v3 [cs.CV] 18 Dec 2017

**Modulating early visual processing by language**

**Harm de Vries***
University of Montreal
mail@harmdevries.com

**Florian Strub***
Univ. Lille, CNRS, Centrale Lille, Inria, UMR 9189 CRIStAL
florian.strub@inria.fr

**Jérémie Mary†**
Univ. Lille, CNRS, Centrale Lille, Inria, UMR 9189 CRIStAL
jeremie.mary@univ-lille3.fr

**Hugo Larochelle**
Google Brain
hugolarochelle@google.com

**Olivier Pietquin**
DeepMind
pietquin@google.com

**Aaron Courville**
University of Montreal, CIFAR Fellow
aaron.courville@gmail.com

**Abstract**

It is commonly assumed that language refers to high-level visual concepts while leaving low-level visual processing unaffected. This view dominates the current literature in computational models for language-vision tasks, where visual and linguistic input are mostly combined late in deep neural networks. However, in this paper, we show that early modulation of visual processing by language can significantly improve a strong baseline on two visual question answering tasks.

**1 Introduction**

Human beings combine the processing of language and vision with apparent ease. For example, we can use natural language to describe perceived objects and we are able to imagine a visual scene given a given textual description. Developing intelligent machines with such impressive capabilities remains a long-standing research challenge with many practical applications.

Towards this grand goal, we have witnessed an increased interest in tasks at the intersection of computer vision and natural language processing. In particular, image captioning [16], visual question answering (VQA)[1, 23] and visually grounded dialogue systems[5, 6] constitute a popular set of example tasks for which large-scale datasets are now available. Developing computational models for these tasks raises many interesting questions. For example, a fundamental question in all these tasks: how to fuse/integrate visual and textual representations? To what extent should we process visual and linguistic input separately, and at which stage should we fuse them? And equally important, what fusion mechanism to use?

In this paper, we restrict our attention to the domain of visual question answering which is a natural testbed for fusing language and vision. The VQA task concerns answering open-ended questions about images and has received significant attention from the research community [1, 9, 17, 23]. Current state-of-the-art systems often use the following computational pipeline [2, 17, 20] illustrated in Fig 1. They first extract *high-level* image features from an ImageNet pretrained convolutional network (e.g. the activations from a ResNet network [12]), and obtain a language embedding using a

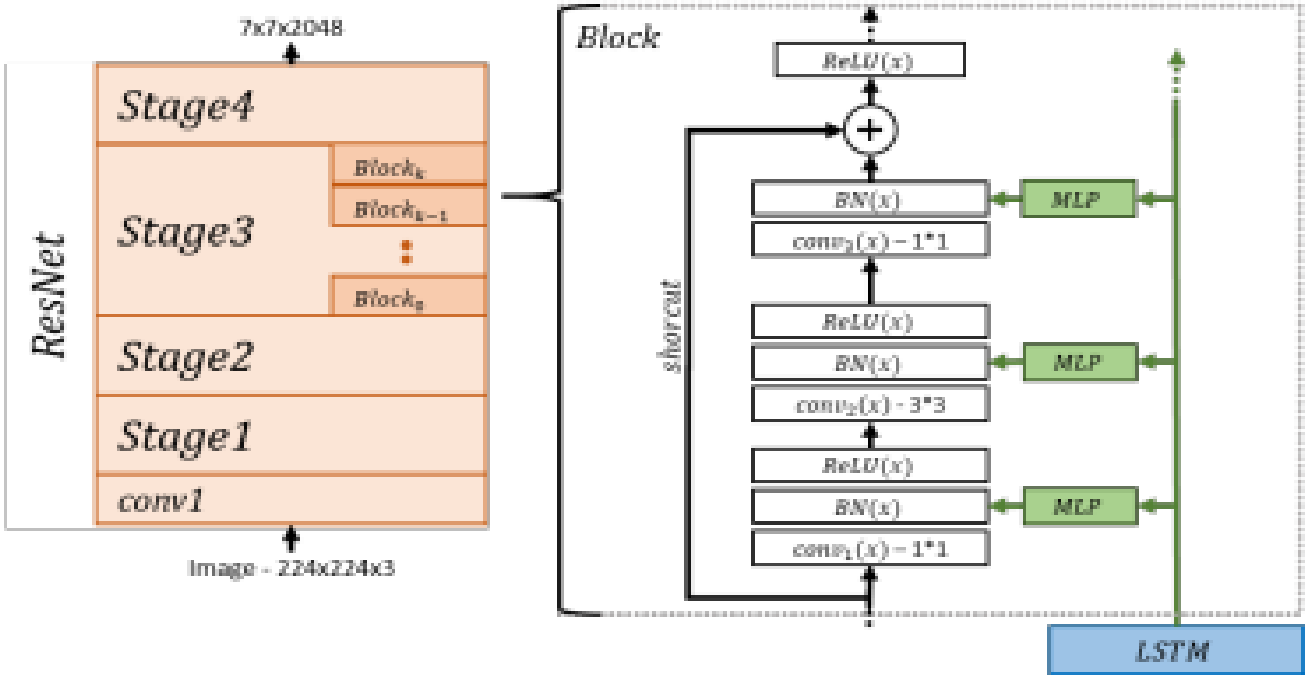
*The first two authors contributed equally
†Now at Criteo

Figure 3: Modulating early visual processing by language

- Utilisation des données VQA (Visual Question Answer)
  - Contenant des questions plus ouvertes que *GuessWhat?!*
- Architecture Resnet
- Normalisation conditionné par lot

# Normalisation conditionné par lot

- Normalisation conditionné par lot



# Ajout des données complémentaires

*Est-ce que l'ajout des données supplémentaires permettra d'améliorer le résultat ?*

- Tout les categories
  - 90 catégories possibles (humain, table, . . . )
  - Utilisation d'un vecteur 1-hot de taille 90
- Description de l'image
  - Le corpus MSCOCO contient les images avec des descriptions textuelles

# Ajout des données complémentaires

- Classification des questions
  - 4 catégories ( Spatial reasoning , Visual properties ,Object taxonomy ,Interaction) .
  - Générer par un simple algorithme.
- L'historique des questions
  - Tout les questions ayons une réponse oui.
  - Tout les questions ayons pour categories ( Object taxonomy , Interaction).
- L'ajout des autres Crop dans l'images
  - Chaque image contient 3 à 20 objets .

# Le mécanisme d'attention

*Comment faire une analyse jointe des données ?*


Figure 4: Exemple de mécanisme d'attention

- Le mécanisme d'attention est inspiré de l'humain.
  - Tendances de la RD en 2017
  - Exemple d'utilisation (traduction automatique , description d'image )

# Modèle de co-attention

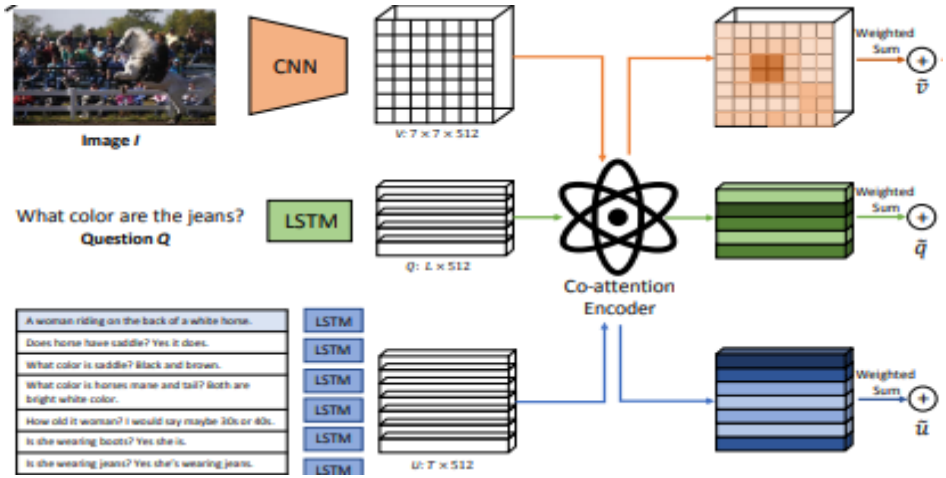



Figure 5: Exemple de modèle d'attention dans le papier *Are you Talking to Me ?*

- Dans le papier *Are You Talking to Me?*.
  - Corpus VQA
  - Résultat plus meilleur avec l'utisation Co-attention

# Modèle de co-attention

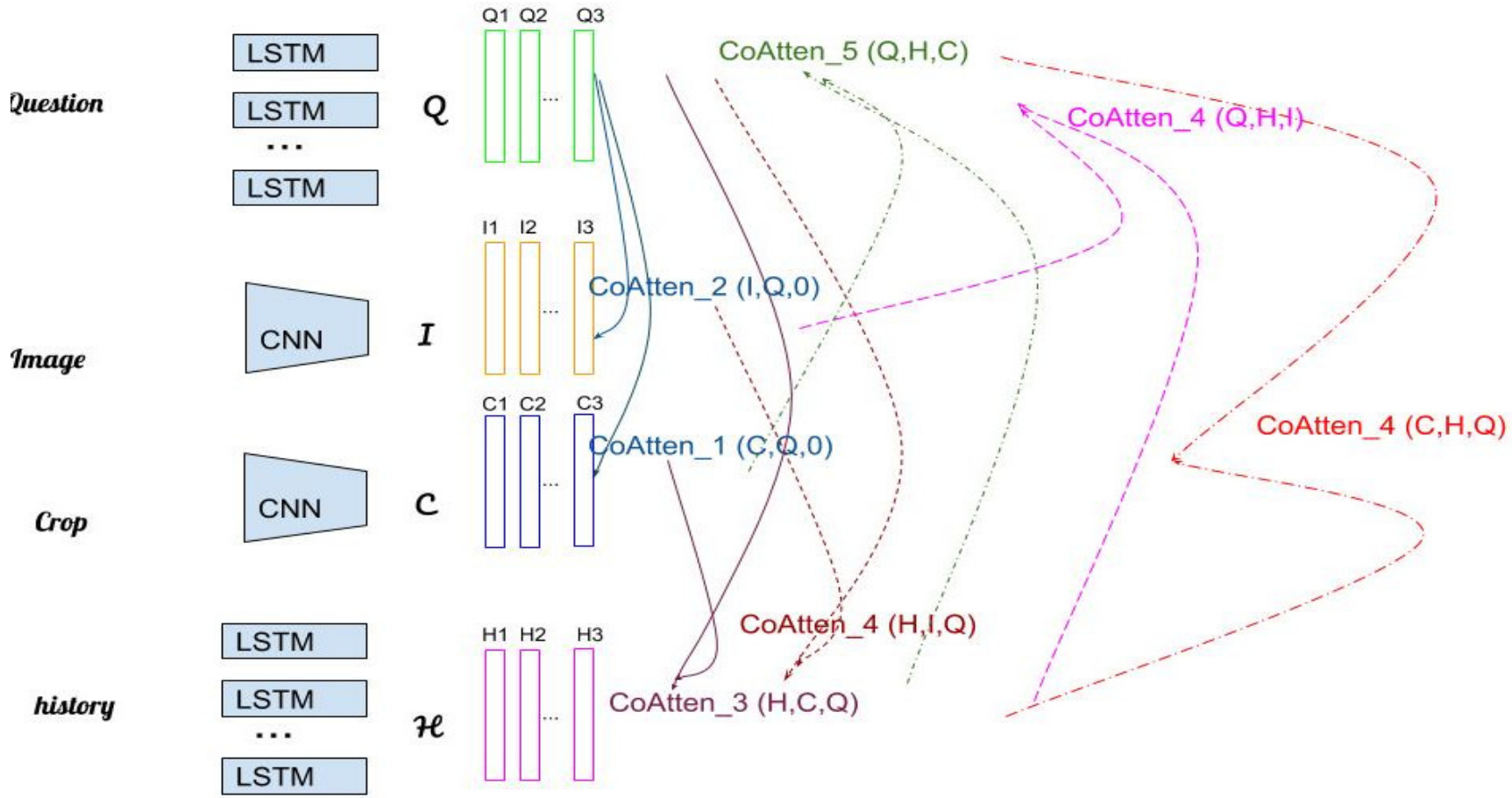



Figure 6: Base de référence de l'Oracle

# Génération d'une description pour le crop

*Comment avoir des données descriptif du crop ?*

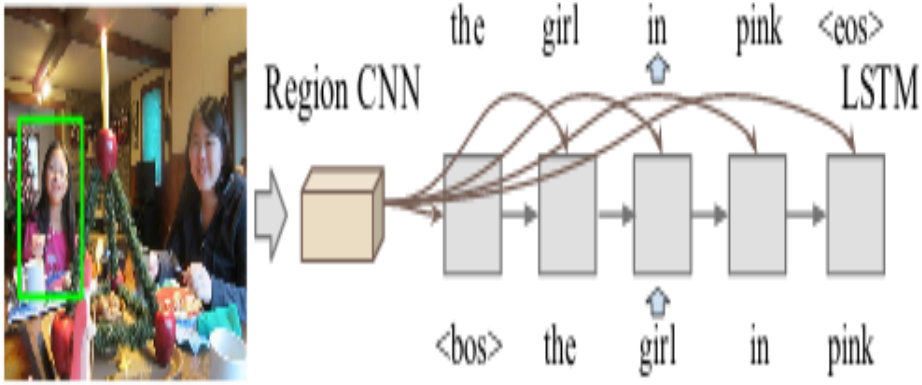



Figure 7: Baseline pour Générer d'une description du crop

- Dans le papier *Generation and Comprehension of Unambiguous Object Descriptions*.
  - Jeu ReferIt Deux joeur
    1. L'un pour decrire l'object
    2. L'autre deviner l'object a partir de la description

# Résult

| Model | Result |
|---|---|
| Base_line | 21.5 |
| ICQ | 21.25 |
| ICQ+Description | 22.4 |
| ICQ+Batch_Normalisation | 19.0 |
| ICQ+Emb_Fasttext+Batch_Normalisation | 17.5 |
| Co_his | 14 |
| Co_his_crops | 13.55 |
| Co_his_crops + yes_hist | 13.3 |
| Co_his_crops + categorised_question | 13.3 |
| Co_his_crops + yes_hist + Ger_des_obj | 12.8 |

Table 2: Result of all experiences .

# Exemple


Figure 8: Un exemple du model de co-attention

# Conclusion

- Oracle GuessWhat? : une tâche multimodale
- Amélioration significative par rapport au système de base
  - Normalisation par lot
  - Co-attention
  - Génération des description du crop

# Perspective

- Amélioration de l'historique des questions
- Utilisation plus poussée du génération du description du crop

# Références


- https://arxiv.org/pdf/1707.00683.pdf
- https://arxiv.org/pdf/1703.05423.pdf
- https://arxiv.org/pdf/1611.08481.pdf
https://arxiv.org/pdf/1707.00683.pdf

- Merci de votre attention