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 (ecrit/parle) a propos d'une image**
- Compréhension multimodale
  - Récupère des informations complémentaires
  - **Amelioration de la prediction dans le cas ou une même information est presente dans des di erents 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 donnees multimodales**
  - ...

Traitement Automatique des Langues / Analyse d'image

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

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*

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

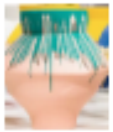
Le contenu du Corpus GuessWhat?!

Partie

[xmin,ymin,xmax,ymax,xcenter,ycenter,wbox,hbow]

Vase

Vase



Vase

Is it a vase? Yes

Is it partially visible? No

Is it the turquoise and purple one? Yes

Partie

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

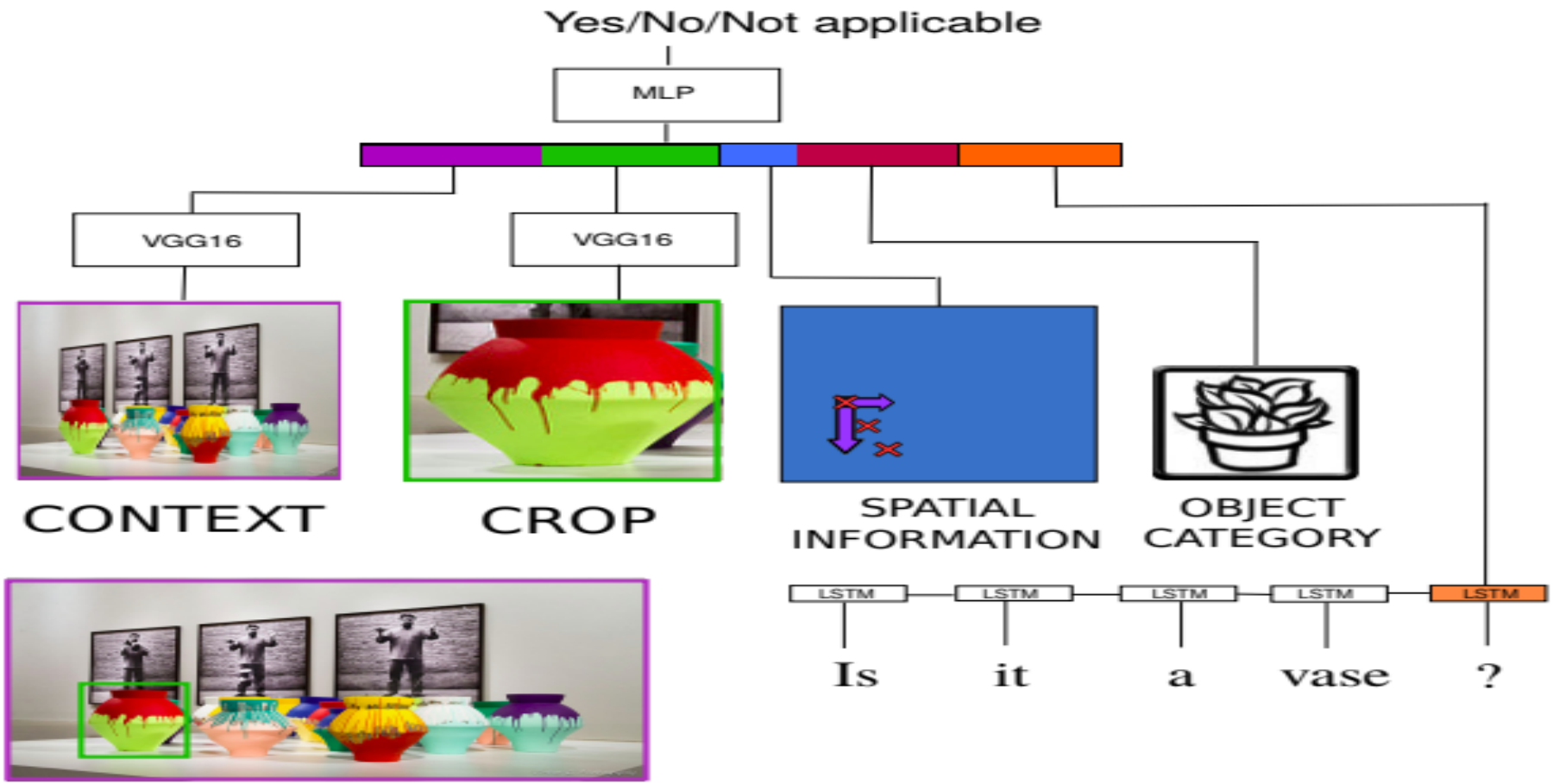


Figure 1: Base de référence de l'Oracle

Traits utilises par le modele

LSTM pour encoder les questions

- Dimension d'embedding de 300
- Cellule de type LSTM , avec 512 unites cachees

Utilise un MLP pour faire la prediction a partir de vecteurs qui contiennent tous les traits concatenes

- Avec 512 unites cachees

optimiseur adam

- 20 epoques
- 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 exploitée

Côté text

Utiliser Simplement les questions .
Utiliser des embedding aléatoire pour les questions.

Amelioration de l'Oracle

Utilisation des donnees VQA (Visual Question Answer)

- Contenant des questions plus ouvertes que GuessWhat?!

Architecture Resnet

Normalisation conditionne par lot

Figure 3: Modulating early visual processing by language

Normalisation conditionnée par lot

Normalisation conditionnée par lot

Ajout des donnees complementaires

Est-ce que l'ajout des donnees supplementaires permettra d'ameliorer le resultat ?

- Tout les categories
  - 90 categories 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 textue

Ajout des donnees complementaires

Classi cation des questions

- 4 categories (Spatial reasoning , Visual properties ,Object taxonomy ,Interaction) .
  Generer par un simple algorithme.

L'historique des questions

- Tout les questions ayons une reponse oui.
  Tout les questions ayons pour categories (Object taxonomy , Interaction).

L'ajout des autres Crop dans l'images

- Chaque image contient 3 a 20 objets .

Le mecanisme d'attention

Comment faire une analyse jointe des donnees ?

Le mecanisme d'attention est inspire de l'humain.

Tendances de la RD en 2017
Exemple d'utilisation (traduction automatique , description d'image)

Figure 4: Exemple de mecanisme d'attention

Modèle de co-attention

Dans le papier Are You Talking to Me?.

- Corpus VQA
- Resultat plus meilleur avec l'utisation Co-attention

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

Modèle de co-attention

Figure 6: Base de référence de l'Oracle

Generation d'une description pour le crop

Comment avoir des donnees descriptif du crop ?

Dans le papier Generation and Comprehension of Unambiguous Object Descriptions

- Jeu ReferIt Deux joeur
  - L'un pour decrire l'object
  - L'autre deviner l'object a partir de la description

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

Result

Model	Result
Baseline	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 + yeshist	13.3
Co_his_crops + categorisedquestion	13.3
Co_his_crops + yeshist + Ger_desobj	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 signi cative 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