Julien	RESUME D'ARTICLE	MS IA
GUEGAN		29/04/2020

"Show me the cup": Reference with Continuous Representations

1. Introduction

L'une des fonctions basiques du langage est de se référer à des objets d'une scène, cette capacité à faire des références est fondamentale puisqu'elle permet de passer de la symbolique du langage au monde réel. Pour qu'une référence soit réussie et comprise, le locuteur doit être précis et choisir une expression qui respecte certaines règles s'il veut que l'auditeur soit capable de le comprendre. L'exemple ci-dessous permet de comprendre que les références impliquent les mécanismes de ***caractérisation*** qui capture les propriétés des objets (« mug » vs « pencil ») et ***d'inviduation*** qui nous permet de distinguer les objets (« the » vs « some »)

(1) *Adam*: Can you please give me…
a. …the mug?
Barbara: Sure.
b. …the pencil?
Barbara (searching): Ahem, I can't see any pencil here…
c. …the book?
Barbara: Sorry, which one?

Figure 1 - Exemple "Show me the cup"

L'article propose un modèle de réseau de neurones qui a pour objectif de modéliser ces 2 mécanismes et qui peut être directement entraîner sur des actes de références. Ce modèle appelé Point-or-Protest (PoP) joue le rôle de Barbara dans l'exemple précédent : étant donnée une expression linguistique, il identifie l'ensemble de l'image correspondant ou proteste en cas d'échec.

2. Modèles

Point-or-Protest est un réseau de neurone feedforward. Dans la figure 2 utilisée ci-dessous, l'ensemble d'entrée utilise un oiseau et 2 tasses, le modèle doit relevé une anomalie du fait de l'ambigüité des 2 tasses. Dans un premier temps, on cherche à obtenir une représentation multimodale partagée de nos images et notre mot, l'expression textuelle est envoyée sur un espace dense en utilisant la représentation `cbow` (pré-entrainé) et les images sont envoyées sur des représentations vectorielles avec un `cnn`. On concatène simplement les 2 représentations pour obtenir la représentation d'entrée du modèle pour ensuite obtenir un vecteur qui encode des valeurs de similarité entre le texte et les images, plus la valeur obtenue est élevée plus l'image est un bon référent pour la requête. Ensuite, le système doit vérifier s'il y a une anomalie en regardant si la similarité cumulée est trop élevée (plus d'une image correspond au mot) ou trop basse (aucune image correspond au mot). Ce module de détection d'anomalie retourne en fait un score (par plusieurs combinaisons linéaires et non linéaire) qui est concaténé au vecteur de similarité qui via une non-linéarité `softmax` retourne une distribution de probabilité, l'indice de la valeur maximale donne la sortie du modèle (si le dernier indice est choisi alors une anomalie est détectée).

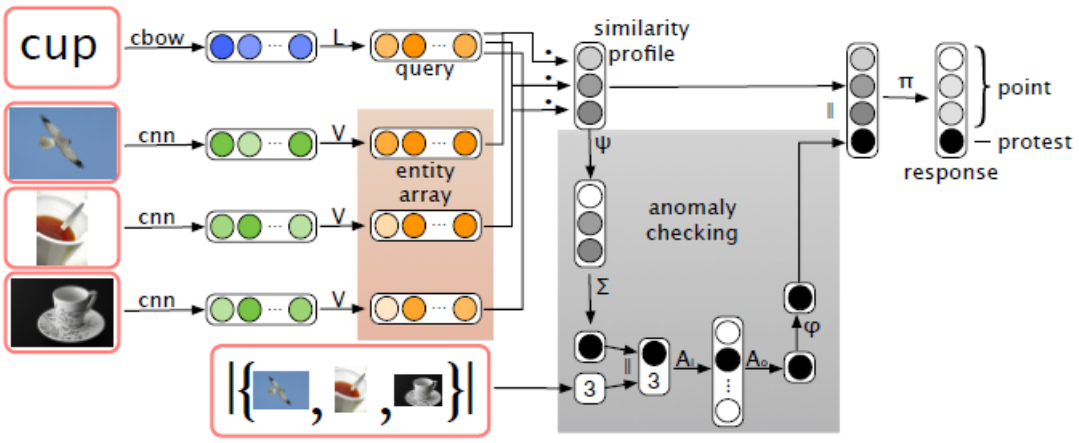

Figure 2 - Illustration du modèle PoP avec un exemple d'anomalie

Les auteurs de l'article proposent également un modèle **Pipeline** en reliant plusieurs modules entrainés et paramétrés séparément. Cette Pipeline utilise par exemple des représentations multimodales en optimisant les similarités entre les paires requête (mot) et objet (image). Pour ce qui est de la détection d'anomalie, des heuristiques utilisant un seuil sont créées.

Un 3ème modèle a été testé en utilisant un **CNN** qui à partir des images d'entrée retournent des labels à matcher avec le mot requête. Par exemple, le CNN serait performant s'il prédisait un synonyme de *cup* pour les images 2 et 3 de l'exemple de la figure 2.

Finalement, un dernier modèle **TRPoP** a été développé en constatant que la représentation `cbow` utilisé dans le modèle PoP repose sur des similarités de mots apprises grâce à des textes. L'hypothèse selon laquelle les significations des mots sont d'abord apprises séparément, uniquement à partir des statistiques linguistiques, puis affinées dans la configuration du modèle PoP, est irréaliste. Idéalement, l'objectif serait d'avoir un modèle qui apprenne les représentations de mots en parallèle à partir des actes de référence et des statistiques linguistiques. Pour le moment, les auteurs proposent une représentation des mots qui serait apprise des actes de référence pendant l'entrainement et le reste du modèle reste identique à PoP.

3. Expériences

Les modèles cités exploitent pour la représentation des mots un embedding `cbow` de 400 dimensions entrainé sur 2.8 milliards textes et un vecteur de dimension 4096 pour la représentation visuelle produit par le CNN VGG-19 pré-entrainé. Ce même modèle VGG sera utilisé pour générer les labels du modèle. Les paramètres des modèles PoP, TRPoP et Pipeline sont estimés par une descente de gradient stochastique. Pour comparer les performances des modèles, 3 baselines sont considérés : **random** assigne les étiquettes aléatoirement, **majority** attribue l'étiquette de sortie la plus fréquente, à savoir l'anomalie qui représente 30% des séquences et **probability** assigne au hasard les étiquettes en fonction de leur fréquence relative dans les données d'entraînement. Les résultats présentent la précision globale mais également détaillent la précision pour les références réussies (*Pointing*), les références loupées (*Miss-Ref*) et les anomalies (*MultRef*).

La 1ère expérience consiste à utiliser pour les données de la requête d'entrée comme expression textuelle un seul nom comme dans l'exemple de la figure 2. On remarque que PoP et TRPoP réalisent des performances comparable ce qui suggère qu'il n'est pas nécessaire d'utiliser une représentation textuelle pré-entrainée puisqu'elle peut être apprise directement des actes de références. D'autre part, on voit que

le simple CNN ne parvient pas à répondre aussi bien au problème que PoP qui, lui, généralise bien au-delà des connaissances du réseau pré-entrainé puisqu'il raisonne dans un espace multimodal.

La 2ème expérience se base sur un second jeu de donnée plus complexe que l'expérience 1, il n'utilise pas seulement un nom pour la requête d'entrée mais associe un nom avec un verbe. Les modèles Pipeline et TRPoP obtiennent de moins résultat que PoP pour ce jeu de donnée. Comparé à TRPoP, le modèle PoP a un important a priori dans la sémantique encodée dans l'embedding des mots pré-entrainé ce qui l'aide à découvrir des relations entre les mots et les images tout en gardant l'information apportée par le verbe à part du nom

	Exp 1: Object Only				Exp 2: Object+Attribute			
	Total	Pointing	MissRef	MultRef	Total	Pointing	MissRef	MultRef
PoP	66	71	57	51	69	77	57	46
TRPoP	65	70	58	44	62	70	38	48
Pipeline	67	75	51	45	65	74	37	55
CNN	35	9	100	94	-	-	-	-
Random	17	17	17	17	17	17	17	17
Majority	30	0	100	100	30	0	100	100
Probability	22	18	30	30	22	18	30	30

Figure 3 - Résultats obtenus (précision en %)

4. Conclusion

Dans ce travail, PoP est un modèle de réseau de neurone qui permet de, à partir d'un ensemble d'image, pointer une image correspondant à un mot donné ou signale la tâche impossible si le mot donné n'est pas adéquat. Le modèle proposé est décrit comme une architecture end-to-end générique qui peut directement être entrainé sur les données. PoP représente avec succès le phénomène de caractérisation parce qu'i est capable de relier les propriétés visuelles aux expressions textuelles. Actuellement, le modèle se concentre seulement sur des requêtes singulières simples (« show me THE cup »), l'architecture devrait être suffisamment générale pour apprendre d'autres types de requêtes (« show me ALL/MANY cup ») ce qui pourrait être tester dans des travaux ultérieurs.