Expérimentation de la Visual Question Answering avec le dataset MSCOCO

Mulapi Tita Ketsia[1], Hanane Anir Chenaf[2], Minh Ha Van Nguyen[3]

[(1,2,3)] Master 2 - Science des Données
Université de Rouen Normandie

18 janvier 2023

1 Résumé

La visual question answering est une tâche d'Intelligence Artificielle qui consiste à répondre à des questions posées par rapport à des images, elle commence d'abord par comprendre l'image et la question, avant d'y répondre à l'aide des phrases en langage naturel. Ainsi une combinaison de compétences en traitement d'images et en traitement du langage naturel s'impose. Dans ce travail, nous avons étudié l'architecture que propose les auteurs [1] ayant abordés cette approche pour la première fois avant d'utiliser le dataset MSCOCO pour entraîner, valider et tester le modèle.

Mots-Clés :
Visual Question Answering (VQA), Apprentissage automatique (ML), Intelligence Artificielle (IA), MSCOCO (Microsoft Coco), Caractéristiques (features,variables,colonnes), données (dataset), instance (observation, ligne), label(classe, cible), Convolutional neural Network (CNN), Long Short-Term Memory (LSTM), Natural Language Processing (NLP).

2 Introduction

Ce résumé s'adresse aux étudiants en science des données et intelligence artificielle, en sciences économiques, en sciences mathématiques, en sciences de l'ingénieur, etc. Il est destiné aux personnes qui utilisent les mathématiques dans la gestion de la production. Il est nécessaire d'avoir des connaissances en apprentissage automatique, en traitement d'image, en optimisation non linéaire et en statistique pour comprendre les notions abordées.

Réalisé dans le cadre du cours de Deep Learning qui s'inscrit comme un champ de recherche très actif en intelligence artificielle, ce résumé est principalement axé sur la conciliation de la théorie et de la pratique des méthodes d'appentissages profonds, il consiste à faire une expérimentation d'une proposition de recherche en partant de l'état de l'art à son éxécution technique. Pour notre part, le choix s'est porté sur le sujet repris en marge du fait qu'il constitue une tâche d'IA-complète se caractérisant ainsi par la présence d'un modèle deep learning pour le traitement de texte (LSTM) et, d'un modèle deep learning pour le traitement des images (CNN) ce qui lui confère la propriété de tâche à connaissances multimodèles.

La VQA trouve son importance dans de nombreux domaines, parmis eux nous pouvons citer l'aide aux personnes malvoyantes, à l'analyste du renseignement, et tant d'autres. Pour y parvenir elle fait recourt à des connaissances de bon sens ainsi qu'à une compréhension visuelle de la scène qui lui est présenté, le tout dans le but de répondre à des questions dans des domaines variés tel que la détection d'objet, la reconnaissance d'activité, pour ne citer que ceux-là. Pour notre part, les résultats que nous avons obtenus ont été concluants, au regards de ce que nous donne l'évaluation à l'état de l'art, ce qui montrent effectivement que cette méthode se positionne à l'état de l'art, comme le "Saint Graal" de la compréhension des images. Ce modèle d'architecture est simple à comprendre d'autant plus que son évaluation se base sur le nombre de questions aux quelles il répond correctement.

Eu égard à ce qui précède nous pouvons donc formuler notre problématique de la manière suivante :

- Comment comprendre une image ?
- Comment comprendre une question (phrase) ?
- Comment répondre à une question posée en rapport à une image ?

Outre le résumé, l'introduction, la conclusion et les références, notre travail s'articule de la manière suivante :

1. Les datasets;
2. Les travaux connexes;
3. L'architecture;
4. L'évaluation.

3 Les Datasets

Afin de mieux comprendre le problème, nous retenons ce qui caractérisent les données de façon générale et de façon particulière à notre expérience :

1. Les datasets proviennent de la base MSCOCO disponible sur **visualqa.org**, elle comporte des images, des questions et des réponses à 0.25, 0.75 et 10 millions respectivement;
2. Deux(2) types d'images existent : **les images réelles** et **les images abstraites** qui sont des images synthétiques générées;
3. Une image est aussi appelé une scène et pour chacune d'entre elle, trois (3) questions ont été recueillies.
4. Une question est répondue par 10 personnes.
5. Il y a 760 000 Questions
6. Il y a 10 Millions de réponses ou annotations
7. Il existe Deux (2) types de réponses : ouverte et à choix multiple.
8. On appelle réponse ouverte (open-ended), une réponse donné par un humain
9. une tâche à choix multiple est à l'opposé d'une réponse ouverte, un algorithme qui choisit une réponse
10. On peut aussi retrouver des données adaptées à des situations particulier :
    (a) MSCOCO contient 5 légendes d'une seule phrase pour toutes les images réelles et 5 autres pour les images abstraites
    (b) les auteurs ont aussi utilisés des questions sans images pour évaluer la nécessité d'avoir une image.
11. Dans le cadre de notre expérience nous avons utilisé les datasets de la version 2 qui contient un totale de **204 721** images réelles associées à des réponses ouvertes;
12. **Train Data** : 443 757 questions, 82 783 images et 4 437 570 de réponses
13. **Val Data** : 214 354 questions, 40 504 images, 2 143 540 de réponses
14. **Test Data** : 447 793 questions, 81 434 images
15. Le fractionnement du dataset consiste en la création de plusieurs datasets parfois pour des raisons d'ingénierie et parfois par prudence : test-dev, test-standard, test-challenge, test-reserve, etc.

4 Les travaux connexes

Traditionnellement, pour répondre à une question sur les images, ce qui se faisait jusque là dans la littérature consistait à rechercher des informations spécifiques et, des réponses relativement simples d'un à trois mots qui suffisent à elles seule à répondre à de nombreux questions, ici nous retenons 3 propositions présente dans la littérature.

1. Effort VQA : Ici on a des questions qui provienennt d'un monde fermé où sont rpédéfinis 16 coulerus de base ou 894 catégorie d'objets.
2. text-based QA : C'est la question answring classique en NLP
3. Description du contenu visuel : ce sont des tâches de démarquages d'images qui servent à décrire le contenu visuel tel que telles que l'image captionning.

5 L'architecture

La VQA explore un principe dénommé la "fusion tardive" (late fusion) qui stipule que la représentation de laquestion avec un réseau LSTM et des caractéristques de l'image avec un CNN sont calculés indépendamment, et fusionnées à l'aide d'une multiplication élément par élément avant d'être posées à travers des couches entièrement connectées pour générer une distribution softmax sur les classes de réponses de sortie. Il s'agirait donc d'essayer de comprendre conjoitement le texte qui est la question posée et la vison qui est matérialisé par une image sachant que les questions sont données par un humain et qu'elles nécessitent des informations spécifiques détaillées afin de fournir un meilleur résultat.
Pour se faire, on dit que la VQA est un 2-channel Vi sion c'est à dire qu'il possède 2 cannaux qui sont ensuite culminé avce une softmax sur k sorties possibles, concrètement cela revien tà dire ce qui suit :

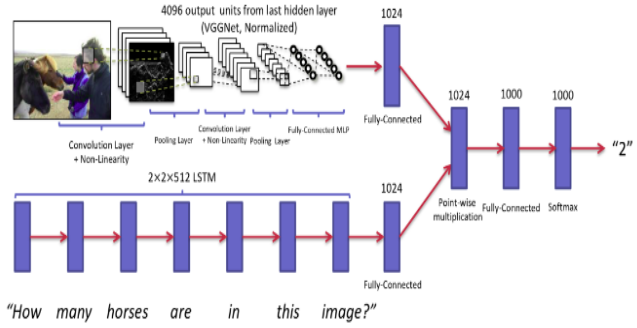
1. Pour le channel Image on veut obtneir l'embedding de l'image :
    (a) I : On active la dernière couche cachée de VggNet qui sont utilisées comme embedding d'image 4096-dim.

(b) Norm I : ensuite, on applique une l2 activation normalisées de a dernière couche cahcé de VggNet

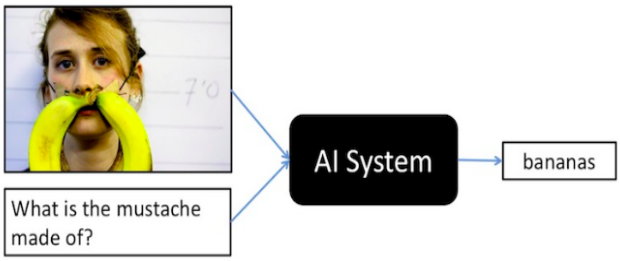
2. Pour le channel de la question on veut faire la même chose mais pour l'image : 3 possibilités s'offrent à nous telle que le Bag-of-Words, le LSTM Q et le Deeper LSTM Q, mais ce qui nous intéresse ici c'est le **LSTM Q** : On utilise un 1-LSTM avec une seule couche cachée pour obtenir un embedding de dimension 1024. cette embedding, est une concaténation des représentations du dernier état de cellule et du dernier état caché (chacun de dimension 512) de la couche cachée du LSTM. On encode chaque mot de la question en une incorporation (embedding de dimension 300) grâce à une couche entièrement connecté et à la fonction d'activation tanh qui est non linéaire. Ces encodeurs sont ensuite transmis au LSTM. A savoir que le vocabulaire d'entrée vers la cocuhe de l'embedding se compose de tous les mots présent dans les données d'apprentissage.

3. La prédiction (le channel réponse) : une fois que les représentations en embedding est faite, le Multi Layer Perceptron se charge de combiner les 2 embeddings (Image + Texte) pour obtenir un embedding unique qui est suivit d'un FCNN connecté suivit d'une couche softmax pour prédire la fameuse distribution sur les k réponses attendues.

4. autres informations : le modèle VQA est un End2End modèle avec une perte de l'entropie croisée et VggNet est à la fois pré-entrainé et non ajusté dans le cannal de l'image.



petite illustration :



6 L'évaluation

Les auteurs se basent sur une formule proposé :

$accuracy = min(\frac{Humanthatprovidedthatanswer}{3}, 1)$

On dit que une réponse est exacte à 100% si au moins 3 personne ont donné cette réponse exacte. Mais on ne se limite pas à ça, c'est à dire qu'on va aussi jusqu'à obeserver les rapports entre les réponses, less types de questions et les types de réponses.

Voici les différents résultats :

Accuracy	**Global**	**Oui/Non**	**Nombre**	**Autre**
Notre résultat	46.66%	66.87%	31.22%	35.38%
Résultat de base	49.15%	67.42%	32.44%	37.28%
Papier Original	41.38%	54.22%	73.46%	35.18%

7 Conclusion

Nous avons vu que l'article de Aishwarya Agrawal sur la réponse aux questions visuelles démontre l'importance de la compréhension de l'image pour résoudre des tâches de réponse aux questions visuelles. Les méthodes de traitement des images combinées à des modèles de traitement du langage naturel ont permis d'améliorer les résultats de ces tâches. Cependant, il reste encore des défis à relever pour améliorer la précision de la réponse aux questions visuelles. Les recherches futures devraient se concentrer sur des méthodes pour combiner encore plus efficacement les données visuelles et linguistiques pour résoudre des tâches de réponse aux questions visuelles plus complexes..

References

[1] Visual Question Answering - Aishwarya Agrawalÿ, Jiasen Luÿ, Stanislaw Antolÿ, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parikh : https://arxiv.org/pdf/1505.00468.pdf

[2] Site web : https://visualqa.org/download.html