www.onera.fr

ONERA
THE FRENCH AEROSPACE LAB

PROPOSITION DE STAGE EN COURS D'ETUDES

Référence : **DTIS-2022-61**
(à rappeler dans toute correspondance)

Lieu : Palaiseau

Département/Dir./Serv. : DTIS/IVA

Tél. : 01 80 38 65 69

Responsable(s) du stage : Stéphane Herbin

Email. : stephane.herbin@onera.fr

DESCRIPTION DU STAGE

Thématique(s) : Intelligence Artificielle et Décision

Type de stage : ☒ Fin d'études bac+5 ☒ Master 2 ☐ Bac+2 à bac+4 ☐ Autres

Intitulé : Analyse de la qualité d'une légende d'image par dialogue

Sujet :

Fried meat is served along side a salad, a few potatoes and chopsticks.

Questioner/Client		Answerer/Service
Are-you sure that there are chopsticks on the table ?	⇄	No. You are right, I don't see any.
Is there silverware by the plate?	⇄	There is a fork and a knife.
Isn't there a drink served with the plate?	⇄	Yes. I see a glass.
What is the color of the salad?	⇄	It's yellow and orange, I think.
OK. It should be cole slaw.		

Contexte

Le stage se situe dans les domaines de la vision artificielle et du traitement du langage naturel. Les techniques d'apprentissage profond ("deep learning") ont permis des avancées spectaculaires de ces deux domaines, indépendamment, mais ont également proposé des schémas algorithmiques de plus en plus sophistiqués capables de les associer. Il est possible maintenant de construire des fonctions de légende automatique d'image ("captioning"), de questions-réponses visuelles, de génération conditionnelle d'images à partir de texte, de recherche dans des bases multimédia, etc. [Mogadala, 2019]

Objectif

L'objectif du stage s'inscrit dans cette orientation de travaux associant vision et langage. Il est de concevoir et d'évaluer un schéma algorithmique capable de décider si la description d'une scène ou d'un environnement est à la fois juste (sans erreur) et suffisamment informative.

Le principe de l'approche à développer est de faire dialoguer deux agents ayant des rôles et des capacités complémentaires [Lee, 2018] : par exemple un répondeur/service, qui est le

seul à avoir accès à l'environnement pour fournir des éléments de description de l'environnement, et un questionneur/client, qui a besoin d'une description finale et sera juge de l'information qu'elle contient (cf. illustration).

La raison de la formulation de l'approche sous la forme d'un dialogue est double:

- permettre de rendre compte de manière explicite des étapes de décision intermédiaires, de prise d'information et de raisonnement conduisant à l'évaluation de la qualité de la description.
- découpler clairement la prise d'information de son utilisation, et permettre ainsi de satisfaire certaines contraintes liées à des besoins de confidentialité ("privacy") ou de propriété intellectuelle des données, ou de conditions d'acquisition particulières (caméra déportée).

Analyser la qualité d'une description peut être utilisé par exemple pour la détection de "fake news" [Atanasova, 2020] [Nakamura, 2020], pour la détection de changement dans un environnement [Jhamtani, 2018] [Park, 2019] [Forbes, 2019] [Hosseinzadeh, 2021] par exemple lorsqu'il est parcouru par un robot qui en possède une certaine description textuelle ou comme critère de récompense pour apprendre par renforcement à améliorer itérativement une description [Seo, 2020].

Contenu des travaux

Les travaux du stage consisteront à adapter une proposition de l'état de l'art à la formulation sous la forme d'un dialogue d'un l'algorithme de justification de la qualité d'une légende d'image. Plusieurs pistes possibles sont envisageables et seront définies au début du stage.

Le stage sera réalisé en collaboration avec l'équipe VERTIGO du laboratoire CEDRIC du CNAM (M. Crucianu et N. Thome).

Etat de l'art

Une première gamme de travaux d'intérêt concerne le dialogue visuel pour la navigation d'un mobile [Chen, 2019], l'identification d'une image [Das, 2017a] ou d'un objet dans une collection [De Vries, 2017] [Mazuecos, 2020] ou la description du contenu visuel d'une image [Das, 2017b] [Murahari, 2019]. Ces travaux ne portent pas cependant sur la construction collaborative d'une évaluation de la qualité de la description et insistent plutôt sur la dimension séquentielle de manipulation de l'information.

Un autre ensemble de travaux s'intéressent aux approches dites neuro-symboliques [2018, Mao] [Yi, 2018] [Stammer, 2021] [Sarker, 2021], en général utilisées pour implémenter un raisonnement visuel [Kottur, 2019] [Hudson, 2019] [Gan, 2019] [Shi, 2019] [Zellers, 2019] [Amizadeh, 2020] [Chen, W., 2021] [Hong, 2021], et pourraient être exploitées pour structurer les échanges d'information sous la forme d'un dialogue.

La construction de références visuelles ("visual grounding") mettant en relation des éléments de la description et leur correspondant dans la scène ou l'image est une question clé pour bâtir un indicateur de qualité et ont donné lieu à de nombreux travaux [Gupta, 2020] [Chen F., 2021] [Chen, Y., 2021] [Ding, 2021] [Gonzalez, 2021] [Arbelle, 2021] [Diomataris, 2021] [Wang, 2021] [Deng, 2021]. Ces problématiques sont à rapprocher de travaux sur l'explicabilité des processus de décision [Ray, 2019] [Kayser, 2021] [Camburu, 2017].

Bibliographie

- Amizadeh, S., Palangi, H., Polozov, A., Huang, Y., & Koishida, K. (2020, November). Neuro-Symbolic Visual Reasoning: Disentangling. In *International Conference on Machine Learning* (pp. 279-290). PMLR.
- Arbelle, A., Doveh, S., Alfassy, A., Shtok, J., Lev, G., Schwartz, E., ... & Karlinsky, L. (2021). Detector-Free Weakly Supervised Grounding by Separation. *arXiv preprint arXiv:2104.09829*.
- Atanasova, P., Simonsen, J. G., Lioma, C., & Augenstein, I. (2020, July). Generating Fact Checking Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 7352-7364).
- Camburu, O. M., Rocktäschel, T., Lukasiewicz, T., & Blunsom, P. (2018). e-snli: Natural language inference with natural language explanations. *arXiv preprint arXiv:1812.01193*.

- Chen, F., Meng, F., Chen, X., Li, P., & Zhou, J. (2021). Multimodal Incremental Transformer with Visual Grounding for Visual Dialogue Generation. *arXiv preprint arXiv:2109.08478*.
- Chen, Howard, et al. "Touchdown: Natural language navigation and spatial reasoning in visual street environments." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- Chen, Wenhu, et al. "Meta module network for compositional visual reasoning." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2021.
- Chen, Y., Li, Q., Kong, D., Kei, Y. L., Zhu, S. C., Gao, T., ... & Huang, S. (2021). YouRefIt: Embodied Reference Understanding with Language and Gesture. *arXiv preprint arXiv:2109.03413*. ICCV.
- Das, A., Kottur, S., Moura, J. M., Lee, S., & Batra, D. (2017). Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision* (pp. 2951-2960).
- Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Moura, J. M., ... & Batra, D. (2017). Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 326-335).
- Deng, J., Yang, Z., Chen, T., Zhou, W., & Li, H. (2021). TransVG: End-to-End Visual Grounding with Transformers. *arXiv preprint arXiv:2104.08541*.
- De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., & Courville, A. (2017). Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5503-5512).
- Ding, H., Liu, C., Wang, S., & Jiang, X. (2021). Vision-Language Transformer and Query Generation for Referring Segmentation. *arXiv preprint arXiv:2108.05565*.
- Markos Diomataris; Nikolaos Gkanatsios; Vassilis Pitsikalis; Petros Maragos (2021) Grounding Consistency: Distilling Spatial Common Sense for Precise Visual Relationship Detection. ICCV.
- Forbes, M., Kaeser-Chen, C., Sharma, P., & Belongie, S. (2019, November). Neural Naturalist: Generating Fine-Grained Image Comparisons. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 708-717).
- Gan, Z., Cheng, Y., Kholy, A., Li, L., Liu, J., & Gao, J. (2019, July). Multi-step Reasoning via Recurrent Dual Attention for Visual Dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 6463-6474).
- González, C., Ayobi, N., Hernández, I., Hernández, J., Pont-Tuset, J., & Arbeláez, P. (2021). Panoptic Narrative Grounding. *arXiv preprint arXiv:2109.04988*.
- Guo, X., Wu, H., Cheng, Y., Rennie, S., Tesauro, G., & Feris, R. S. (2018, December). Dialog-based interactive image retrieval. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (pp. 676-686).
- Gupta, T., Vahdat, A., Chechik, G., Yang, X., Kautz, J., & Hoiem, D. (2020). Contrastive learning for weakly supervised phrase grounding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16* (pp. 752-768). Springer International Publishing.
- Hong, X., Lan, Y., Pang, L., Guo, J., & Cheng, X. (2021). Transformation Driven Visual Reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6903-6912).
- Hosseinzadeh, M., & Wang, Y. (2021). Image Change Captioning by Learning From an Auxiliary Task. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2725-2734).
- Hudson, Drew A., and Christopher D. Manning. "Gqa: A new dataset for real-world visual reasoning and compositional question answering." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2019.
- Jhamtani, H., & Berg-Kirkpatrick, T. (2018). Learning to Describe Differences Between Pairs of Similar Images. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4024-4034).
- Kayser, M., Camburu, O. M., Salewski, L., Emde, C., Do, V., Akata, Z., & Lukasiewicz, T. (2021). e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. *arXiv preprint arXiv:2105.03761*.
- Kottur, S., Moura, J. M., Parikh, D., Batra, D., & Rohrbach, M. (2019, January). CLEVR-Dialog: A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog. In *NAACL-HLT (1)*.
- Lee, Sang-Woo, Yu-Jung Heo, and Byoung-Tak Zhang. "Answerer in Questioner's Mind: Information Theoretic Approach to Goal-Oriented Visual Dialog." (2018).

- Luo, R., Price, B., Cohen, S., & Shakhnarovich, G. (2018). Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6964-6974).
- Luo, R. (2021). Goal-driven text descriptions for images. PhD Thesis. TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO. *arXiv preprint arXiv:2108.12575*.
- Mao, Jiayuan, et al. "The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision." *International Conference on Learning Representations*. 2018.
- Matsumori, S., Shingyouchi, K., Abe, Y., Fukuchi, Y., Sugiura, K., & Imai, M. (2021). Unified Questioner Transformer for Descriptive Question Generation in Goal-Oriented Visual Dialogue. *arXiv preprint arXiv:2106.15550*.
- Mazuecos, Mauricio, et al. "On the role of effective and referring questions in GuessWhat?!." *Proceedings of the first workshop on advances in language and vision research*. 2020.
- Mogadala, A., Kalimuthu, M., & Klakow, D. (2019). Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *arXiv preprint arXiv:1907.09358*.
- Murahari, Vishvak, et al. "Improving Generative Visual Dialog by Answering Diverse Questions." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.
- Nakamura, K., Levy, S., & Wang, W. Y. (2020, May). Fakeddit: A New Multimodal Benchmark Dataset for Fine-grained Fake News Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 6149-6157).
- Pang, W., & Wang, X. (2020, April). Visual dialogue state tracking for question generation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 07, pp. 11831-11838).
- Park, D. H., Darrell, T., & Rohrbach, A. (2019). Robust change captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 4624-4633).
- Ray, A., Yao, Y., Kumar, R., Divakaran, A., & Burachas, G. (2019, October). Can you explain that? Lucid explanations help human-AI collaborative image retrieval. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (Vol. 7, No. 1, pp. 153-161).
- Seo, P. H., Sharma, P., Levinboim, T., Han, B., & Soricut, R. (2020, April). Reinforcing an image caption generator using off-line human feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 03, pp. 2693-2700).
- Shi, Jiaxin, Hanwang Zhang, and Juanzi Li. "Explainable and explicit visual reasoning over scene graphs." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2019.
- Suhr, Alane, et al. "A Corpus for Reasoning about Natural Language Grounded in Photographs." *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.
- Sarker, M. K., Zhou, L., Eberhart, A., & Hitzler, P. (2021). Neuro-Symbolic Artificial Intelligence Current Trends. *arXiv preprint arXiv:2105.05330*.
- Stammer, Wolfgang, Patrick Schramowski, and Kristian Kersting. "Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting with their Explanations." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2021.
- Wang, Y., Joty, S., Lyu, M., King, I., Xiong, C., & Hoi, S. C. (2020, November). VD-BERT: A Unified Vision and Dialog Transformer with BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3325-3338).
- Wang, L., Huang, J., Li, Y., Xu, K., Yang, Z., & Yu, D. (2021). Improving Weakly Supervised Visual Grounding by Contrastive Knowledge Distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14090-14100).
- Yi, Kexin, et al. "Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding." *NeurIPS*. 2018.
- Zellers, R., Bisk, Y., Farhadi, A., & Choi, Y. (2019). From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 6720-6731).

Est-il possible d'envisager un travail en binôme ? **Non**

Méthodes à mettre en oeuvre :

☐ Recherche théorique ☐ Travail de synthèse

☒ Recherche appliquée ☐ Travail de documentation

☐ Recherche expérimentale	☐ Participation à une réalisation
Possibilité de prolongation en thèse :	**Oui**
Durée du stage : Minimum : 4 mois	Maximum : 6 mois
Période souhaitée : mars - septembre	

PROFIL DU STAGIAIRE

Connaissances et niveau requis : Cursus en intelligence artificielle, vision par ordinateur ou traitement du langage naturel. Expérience des environnements Deep Learning (Pytorch, Tensor Flow)	Ecoles ou établissements souhaités : Grande école d'ingénieur ou M2R