Postdoctoral project proposal

Title: **Uncertainty in Vision Transformers**

Postdoc Advisors:

Franck Davoine, CNRS researcher, and Thierry Denoeux, Professor.
Heudiasyc Lab., CNRS, University of Technology of Compiègne, Alliance Sorbonne Université, **France**.
Franck.Davoine@hds.utc.fr / Thierry.Denoeux@hds.utc.fr

---> The closing date for applications is June 5, 2021.

Context of the study:

Since the seminal paper "Attention is all you need" of Vaswani et al. [1] published in 2017, self-attention techniques, and specifically Transformers have become the state of the art in natural language processing (NLP), with their ability to explicitly model all pairwise interactions between elements in a text. Transformers are networks developed as an alternative to recurrent or convolution layers for sequence modelling. They gave birth to huge but very powerful models like OpenAI's GPT (*Generative Pre-Training Transformer*) and Google AI's BERT (*Bidirectional Encoder Representations from Transformers*), achieving state-of-the-art performance on many language modelling benchmarks, and performing rudimentary reading comprehension, machine translation, question answering, and summarization.

Since 2019, Transformers are becoming increasingly popular in computer vision (much like convolutional networks became very popular in the 2012s). One reason is that, in their standard version, Transformers are free of convolutions and attain good results compared to state-of-the-art CNNs when pre-trained on very large image datasets. To handle 2D images for example, the standard vision Transformer receives as input a 1D sequence of linearly embedded fixed-sized image patches, treated as tokens (or words) in an NLP application. It relies on a so-called multi-head self-attention mechanism (with position-wise feed-forward network sub-layers) to model and capture long-range interactions between semantic concepts in images.

Several model adaptations and variants have been proposed to solve different tasks like image classification [2], semantic image segmentation [3], monocular depth estimation [4], object detection [5] and tracking [6], video-based prediction [7] or joint text and image or video encoding [8]. Transformers are shown able to address other notoriously difficult tasks such as, e.g., generative adversarial networks (GANs) [9], image view synthesis [10] or reinforcement learning. Surveys of efficient Transformer architectures are given in [11,12,13].

Postdoc description:

Recent work focuses on different questions such as: How to learn Transformers from smaller datasets [14], how to limit the computational complexity of the training stage [15], how to compute relevancy in Transformer networks [16], or should we incorporate convolutions in vision Transformers to yield the best of both designs [17].

But we notice that, despite the importance of Transformer models, the literature with regards to uncertainty in vision Transformers is sparse. A first method dedicated to estimating uncertainty in the

Transformer model, in the context of sequence prediction, has been proposed in December 2020 [18]. Authors use sequential Monte Carlo methods to approximate the observations distribution in the transformer architecture. Other work applying transformers to hate speech detection uses Monte Carlo dropout within the attention layers of the model to provide well-calibrated reliability estimates [19].

On the other hand, uncertainty estimation in deep learning based on convolutional networks has received a lot of interest from researchers over the last decade, mostly using frequentist or Bayesian approaches. At Heudiasyc, T. Denoeux has revisited logistic regression and its extensions, including multilayer feed-forward neural networks, by showing that these classifiers can be seen as converting (input or higher-level) features into mass functions and aggregating them by Dempster's rule of combination [20]. This research has been extended and applied to classification of objects in 3D point clouds using so-called evidential end-to-end deep neural networks [21], and more recently to semantic image segmentation using fully convolutional networks [22].

---> In connection with these research works carried out in the two CID and SyRI teams of Heudiasyc Lab., the postdoc project will aim at quantifying prediction uncertainty in vision Transformers using evidence theory. The research will focus on target applications, depending on the applicant interests: on-road driving scenes analysis (multi-object detection and/or tracking, or semantic image segmentation), or mimicking gesture recognition in collaboration with Sorbonne University.

Keywords: Self-attention mechanisms and vision Transformers, uncertainty, evidence theory and belief functions.

Postdoc duration: The duration of the post is 12 months in the first instance. Remuneration is based on the pay scale of the University of Technology of Compiègne.

Start date: Fall 2021 (not later).

Candidate's profile:
Applicants must hold a PhD and should demonstrate a consistently outstanding academic record including publications, and be highly proficient in spoken and written English. The ideal candidate will be experienced in one or more of the following areas: machine/deep learning, computer vision, management of uncertainty in intelligent systems and/or data science. Strong evidence for advanced software development experience (Python/C++ programming, active Github/Gitlab profile or similar) is desirable. The role holder may be asked to assist in the supervision of student projects, the development of student research skills, plan/deliver seminars relating to the research area.

Recruitment: Transformers are a hot topic today that should be of interest to many candidates that for example have already an experience on deep learning based on convolutional networks. The project is an exceptional opportunity to conduct ambitious research at the forefront of machine learning and artificial intelligence.

To apply, please submit to the postdoc advisors a single PDF file containing a full CV (inc. publication list), personal statement (describing your research interests and motivation for applying), strong evidence for advanced software development experience (Python/C++ programming, Github/Gitlab profile or similar), contact information of two or three referees.

Formal interviews of shortlisted candidates will take place likely via video conference in mid-June 2021.

References:

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin. Attention is All you Need. NIPS **2017**. arXiv:1706.03762

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. **2020**. arXiv:2010.11929

[3] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H.S. Torr, Li Zhang. Rethinking Semantic Segmentation from a Sequence-to-Sequence Perspective with Transformers. CVPR **2021**. arXiv:2012.15840

[4] René Ranftl, Alexey Bochkovskiy, Vladlen Koltun. Vision Transformers for Dense Prediction. **2021**. arXiv:2103.13413

[5] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, Sergey Zagoruyko. End-to-End Object Detection with Transformers. ECCV **2020**. arXiv:2005.12872

[6] Peize Sun, Yi Jiang, Rufeng Zhang, Enze Xie, Jinkun Cao, Xinting Hu, Tao Kong, Zehuan Yuan, Changhu Wang, Ping Luo. TransTrack: Multiple-Object Tracking with Transformer. **2020**. arXiv:2012.15460

[7] Tsung-Ming Tai, Giuseppe Fiameni, Cheng-Kuang Lee, Oswald Lanz. Higher Order Recurrent Space-Time Transformer. 2021. arXiv:2104.08665

[8] Max Bain, Arsha Nagrani, Gül Varol, Andrew Zisserman. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. **2021**. arXiv:2104.00650

[9] Yifan Jiang, Shiyu Chang, Zhangyang Wang. TransGAN: Two Transformers Can Make One Strong GAN. **2021**. arXiv:2102.07074

[10] Robin Rombach, Patrick Esser, Björn Ommer. Geometry-Free View Synthesis: Transformers and no 3D Priors. 2021. arXiv:2104.07652

[11] Yi Tay, Mostafa Dehghani, Dara Bahri, Donald Metzler. Efficient Transformers: A Survey. **2020**. arXiv:2009.06732

[12] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, Zhaohui Yang, Yiman Zhang, Dacheng Tao. A Survey on Visual Transformer. **2021**. arXiv:2012.12556

[13] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, Mubarak Shah. Transformers in Vision: A Survey. **2021**. arXiv:2101.01169

[14] Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, Humphrey Shi. Escaping the Big Data Paradigm with Compact Transformers. **2021**. arXiv:2104.05704

[15] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, Hervé Jégou. Training data-efficient image transformers & distillation through attention. **2021**. arXiv:2012.12877

[16] Hila Chefer, Shir Gur, Lior Wolf, Transformer Interpretability Beyond Attention Visualization, **2021**. arXiv:2012.09838

[17] Haiping Wu, Bin Xiao, Noel Codella, Mengchen Liu, Xiyang Dai, Lu Yuan, Lei Zhang. CvT: Introducing Convolutions to Vision Transformers. **2021**. arXiv:2103.15808

[18] Alice Martin, Charles Ollion, Florian Strub, Sylvain Le Corff, and Olivier Pietquin. The Monte Carlo Transformer: a stochastic self-attention model for sequence prediction. **2020**. arXiv:2007.08620

[19] Kristian Miok, Blaž Škrlj, Daniela Zaharie, Marko Robnik-Šikonja. To BAN or Not to BAN: Bayesian Attention Networks for Reliable Hate Speech Detection. Cognitive Computation, **2021**. arXiv:2007.05304

[20] Thierry Denoeux. Logistic Regression, Neural Networks and Dempster-Shafer Theory: A New Perspective. Knowledge-Based Systems, 176, **2019**. hal-01830389

[21] Édouard Capellier, Franck Davoine, Véronique Cherfaoui, You Li. Fusion of neural networks for LIDAR-based evidential road mapping. Journal of Field Robotics, **2021**. arXiv:2102.03326

[22] Zheng Tong, Philippe Xu and Thierry Denoeux. Evidential fully convolutional network for semantic segmentation. Applied Intelligence (to appear), **2021**. arXiv:2103.13544

Other related references:

[-] Zheng Tong, Philippe Xu and Thierry Denoeux. An evidential classifier based on Dempster-Shafer theory and deep learning. Neurocomputing (to appear), **2021**. arXiv:2103.13549

[-] Maxime Chaveroche, Franck Davoine, Véronique Cherfaoui. Focal points and their implications for Möbius Transforms and Dempster-Shafer Theory. Information Sciences. 555, **2021**. arXiv:2011.06549

[-] Xuhong Li, Yves Grandvalet, Franck Davoine. A Baseline Regularization Scheme for Transfer Learning with Convolutional Neural Networks. Pattern Recognition, 98, **2020**. hal-02315752