Proposal

Zhiyu Xue

August 2020

1 Paper List

I wonder if it is feasible to do some interesting work by combining different areas, include but not limited to few-shot learning, image captioning and interpretability. I select three possible topics and their corresponding papers as follows.

- Interpretable Few-shot Learning
  According to my knowledge, rare of previous work focused on the issue of interpretable few-shot learning when I submitted my paper to ACM MM. However, some works about this issue have been released recently. I list the papers ranked by time ascendingly.

  1. Few-shot Learning by Exploiting Visual Concepts within CNNs [6]
     This is the earliest paper (2017) about interpretable few-shot learning that I can find. In this paper, the researcher explains the inner process by using additional visual concepts(VCs). VCs can be considered as the internal representations corresponding to mid-level semantic visual cues.
  2. Interpretable Few-Shot Learning via Linear Distillation [2]
     This paper presents a bidirectional distillation method to explain few-shot models on linear neural networks(LNNs), where LNNs are forward neural networks with no nonlinearities. (Personally, I think this paper is somehow weak because it only evaluates the model on some toy dataset like MNIST and Omniglot)
  3. Explanation-Guided Training for Cross-Domain Few-Shot Classification [11]
     In this paper, the writer applies layer-wise relevance propagation(LRP) method to few-shot learning, by developing a model-agnostic sample method to find the features which are important to predictions [3].
  4. Concept Learners for Generalizable Few-Shot Learning [4]
     This paper somehow uses the same motivation as my paper, which means it views interpretability as the question to quantify regions' contributions to the final decision. Honestly, it is much better than mine. This paper suggests an interpretable domain-agnostic meta-learning approach, which consists of different concept learners.

- (Few-shot/Interpretable) Text-to-Image Generation

  Your series of work about how to generate images from texts(text-to-image), as well as the AI creativity, is really impressive to me. I wonder if we can go much deeper in this area.

  For example (just a simple idea without thinking carefully, it may be incorrect), can we reveal the inner process between inputted text and outputted image? Can we control the visual items by altering the words in the text? Like I change the expression from 'black beak' to 'white beak', will it reflect this change in the generated images?

  Moreover, can our model generate meaningful images when it comes only a few text samples related to new categories?

  1. A Survey and Taxonomy of Adversarial Neural Networks for Text-to-Image Synthesis [1]
     This is a survey of text-to-image generation, and I still read on it.

2. Semantics Disentangling for Text-to-Image Generation [8]

   This paper presents a novel text-to-image generation model that implicitly disentangles semantics to fulfill high-level consistency and low-level diversity.

   I think this paper may inspire me to design a metric-based few-shot text-to-image model since it uses the Siamese mechanism in the discriminator.

3. Controllable Text-to-Image Generation [9]

   This paper presents a generator which is able to manipulate specific visual attributes without affecting the generation of other content. I consider it as a kind of interpretable model.

- Image Captioning

  This is new area for me, and I only study on this topic for several weeks. Personally, I think image captioning is one of the most difficult task in CV and NLP, since it need to bridge the gap between the representations of images and sentences, and generate the text word by word.

  1. Fast Parameter Adaptation for Few-shot Image Captioning [7]

     Up till now, it is the only paper I found about few-shot image captioning.

  2. Memory Transformer for Image Captioning [5]
  3. X-Linear Attention Networks for Image Captioning [10]

  The last paper is mainly about how to use transformer to achieve image captioning.

2 Plan for My Internship

Personally, the main purpose of my internship is to demonstrate that I am qualified for your graduate program in 2021 fall. Some of your research programs are highly matched with my previous works, and I'm highly interested in them.

Getting a PhD position is the best result, but I will also be glad to have a master's position if you think I need to learn more before taking the PhD program.

As we have discussed in the interview, I can take at least seventy-five percents of my spare time to catch up with this internship program.

From now, my main plan about the internship is to learn some new things, and then try to come up with a new idea by combining these new things with my original knowledge. It cannot be better if we can publish a paper that is creative and solid.

3 Agenda

Dear Prof. Elhoseiny, just for your reference, please check my schedule in Table 1.

Day / Time	Mon	Tue	Wen	Thu	Fri	Sat	Sun
8:00 - 10:00	Free	Course	Free	Free	Free	Free	Free
10:00 - 12:00	Free	Course	Free	Course	Free	Free	Free
12:00 - 14:00	Free	Free	Course	Free	Free	Free	Free
14:00 - 16:00	Course	Free	Free	Free	Free	Free	Free
16:00 - 18:00	Free	Free	Free	Free	Free	Free	Free
18:00 - 24:00	Free	Free	Free	Course(18:00 - 21:00)	Free	Free	Free

Table 1: My Schedule: Note that in China we always begin to work on Monday per week. *Time: China Time(GMT+8)

References

[1] Jorge Agnese, Jonathan Herrera, Haicheng Tao, and Xingquan Zhu. A survey and taxonomy of adversarial neural networks for text-to-image synthesis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1345, 2020.

[2] Arip Asadulaev, Igor Kuznetsov, and Andrey Filchenkov. Interpretable few-shot learning via linear distillation. *arXiv preprint arXiv:1906.05431*, 2019.

[3] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[4] Kaidi Cao, Maria Brbic, and Jure Leskovec. Concept learners for generalizable few-shot learning. *arXiv preprint arXiv:2007.07375*, 2020.

[5] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020.

[6] Boyang Deng, Qing Liu, Siyuan Qiao, and Alan Yuille. Few-shot learning by exploiting visual concepts within cnns. *arXiv preprint arXiv:1711.08277*, 2017.

[7] Xuanyi Dong, Linchao Zhu, De Zhang, Yi Yang, and Fei Wu. Fast parameter adaptation for few-shot image captioning and visual question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 54–62, 2018.

[8] Fei Fang, Fei Luo, Hong-Pan Zhang, Hua-Jian Zhou, Alix LH Chow, and Chun-Xia Xiao. A comprehensive pipeline for complex text-to-image synthesis. *Journal of Computer Science and Technology*, 35:522–537, 2020.

[9] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip Torr. Controllable text-to-image generation. In *Advances in Neural Information Processing Systems*, pages 2065–2075, 2019.

[10] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10971–10980, 2020.

[11] Jiamei Sun, Sebastian Lapuschkin, Wojciech Samek, Yunqing Zhao, Ngai-Man Cheung, and Alexander Binder. Explanation-guided training for cross-domain few-shot classification. *arXiv preprint arXiv:2007.08790*, 2020.