Portfolio

jinyyim@gmail.com / Jinyeong Yim

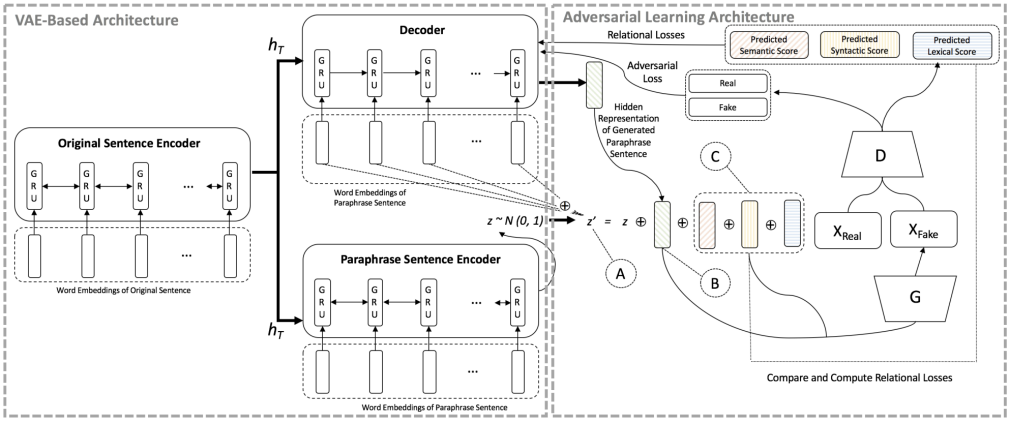
The portfolio consists of three main sections:

- Natural language processing / Machine learning
- Human computer interaction / Crowdsourcing
- Other research areas

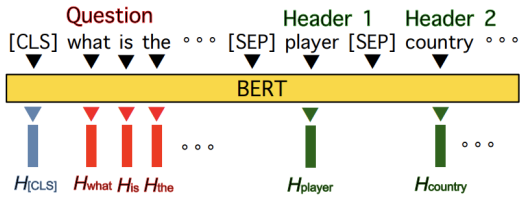
Natural Language Processing / Machine Learning (2018 ~ Present)

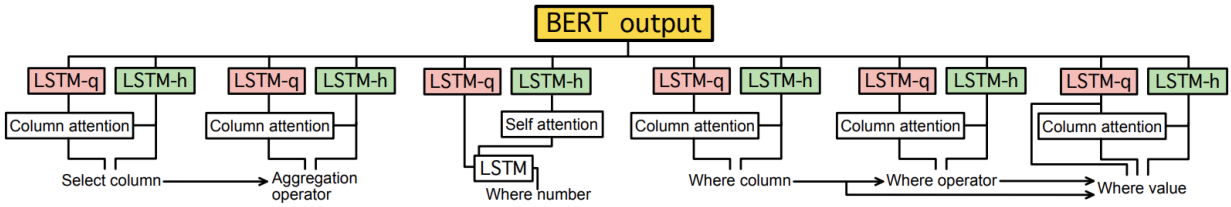
- [NLP] Paraphrase Generation (Jun. 2018 ~ Sep. 2018)
  - Paraphrase generation model with reinforcement learning objective to enhance the quality on seq2seq model
  - The total model learning objective consists of two objectives:
    - Autoregressive generation objective
    - Reinforcement learning objective: RL objective has a small contribution to gradient at the beginning, but its gradient contribution increases as the training procedure increases
  - (Dataset) Quora question pairs
  - (Evaluation metrics) ROUGE-1, ROUGE-2, BLEU, METEOR
  - (Role) everything - literature review, modeling, evaluation

  - Paraphrase diversification with biased data training techniques
  - (Publication)
    - Sunghyun Park, Seung-won Hwang, Fuxiang Chen, Jaegul Choo, Jung-Woo Ha, Sunghun Kim, and **Jinyeong Yim**. "Paraphrase diversification using counterfactual debiasing." In Proceedings of the AAAI Conference on Artificial Intelligence. 2019. (AAAI 2019)
  - (Dataset) Quora, MSCOCO, SNLI
  - (Evaluation metrics) ROUGE-1, ROUGE-2, BLEU, METEOR
  - (Role) Data processing and code review for reinforcement learning objective

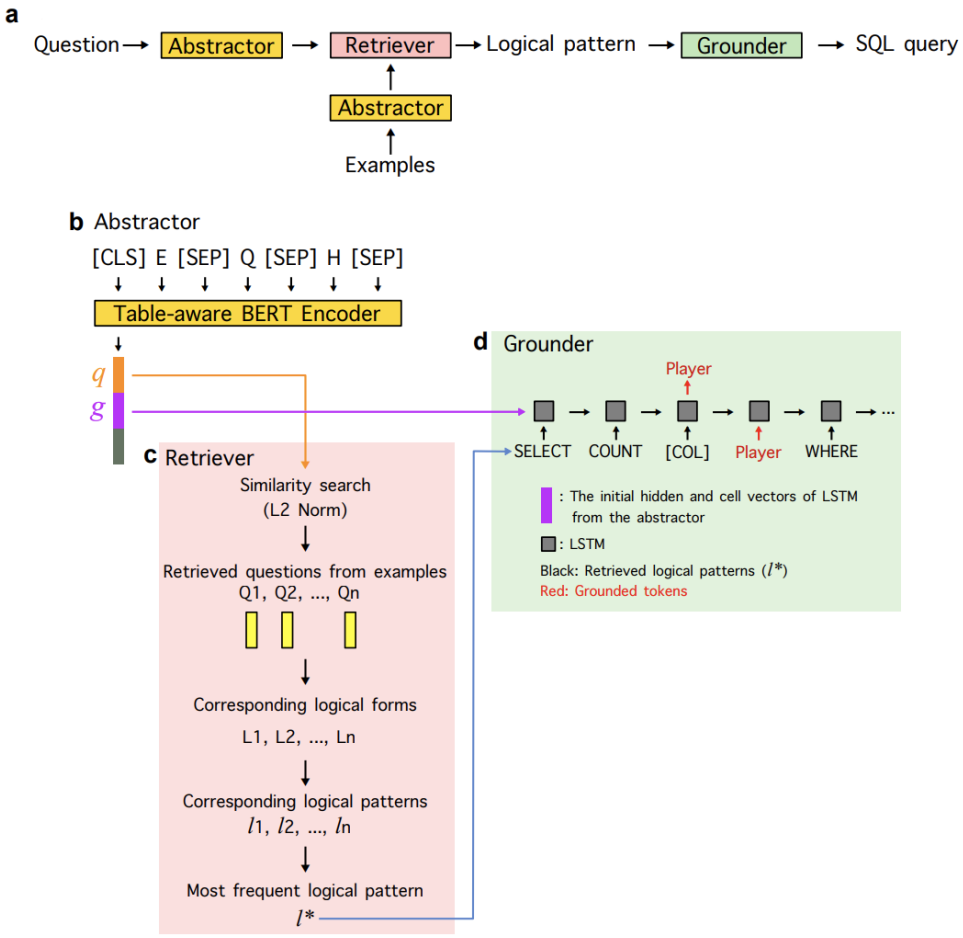


- [NLP] Semantic Parsing / Table QA (Jun. 2018 ~ Feb. 2019)
    - Converting a natural language into a specific logical form
        - Examples of the logical form: SQL query
    - Task types
        - Supervised: A natural language and logical form pair is given
            - Dataset: WikiSQL
        - Weakly-supervised: A natural language and an execution result of a logical form pair is given
            - Dataset: WikiTableQuestions
    - (Publication)
        - Wonseok Hwang, **Jinyeong Yim**, Seunghyun Park, and Minjoon Seo. "A comprehensive exploration on wikisql with table-aware word contextualization." arXiv preprint arXiv:1902.01069 (2019). (KR2ML Workshop at NeurIPS 2019)
    - The first work to apply BERT on the supervised semantic parsing task
        - Suggested how to make a BERT input with a natural language with a given table
    - (Role) Data processing / Model design / Human evaluation

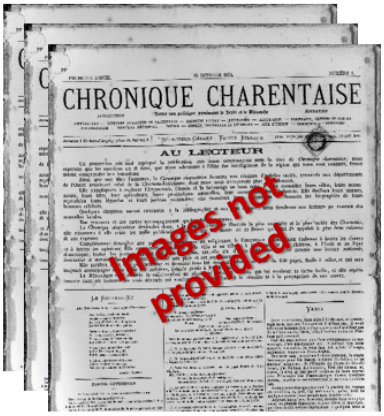




- [NLP] Retrieval-based Semantic Parsing (Feb. 2019 ~ Jun. 2019)
  - Generating a logical form is a heavy task for a parser in general
  - Especially, in a data-scarce environment, deep learning models often fails at successful inference
  - We suggest a retrieval-based method to find a target logical form based on the data we already have and match the input natural language sentence with the most similar natural language that we already have in the database.
  - (Data) WikiSQL
  - (Publication)
    - W. Hwang, **J. Yim**, S. Park, and M. Seo. "Syntactic Question Abstraction and Retrieval for Data-Scarce Semantic Parsing." Automated Knowledge Base Construction (AKBC 2020)
  - (Role) Data processing / Model design



- [NLP] Post-OCR erroneous text detection/correction (2019. 2. ~ 2019. 5.)
  - Participated in the Post-OCR correction competition
    - https://sites.google.com/view/icdar2019-postcorrectionocr
  - (Result) 1st ranked in both tracks in detection and correction
  - (Role) Solely conducted everything
    - Data analysis and data processing
    - Model design and implementation
    - Evaluation
  - Model: Context-based Character Correction (CCC) model
    - BERT-based character-level fine-tuning model
    - Model structure consists of context aggregation parts with convolutional neural networks and information merging parts with LSTM networks



Corpus counting up to 12 M characters

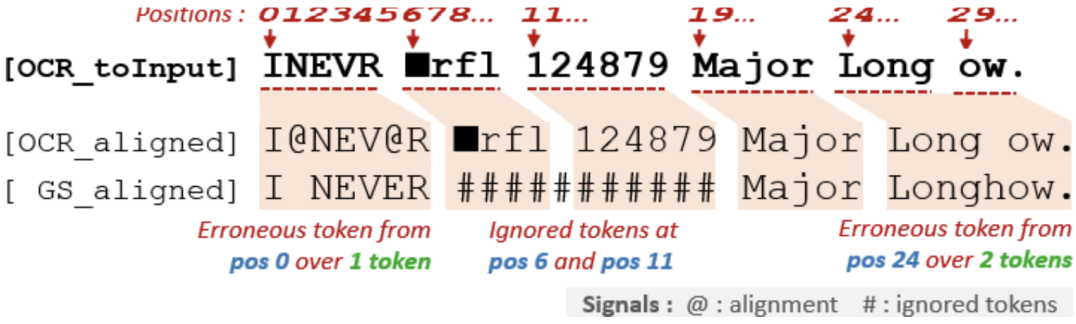
OCR-ed text

The law in that cafe was severe , for cowards and runaways were not omly degraded from a|H honors, but it was also a disgrace…

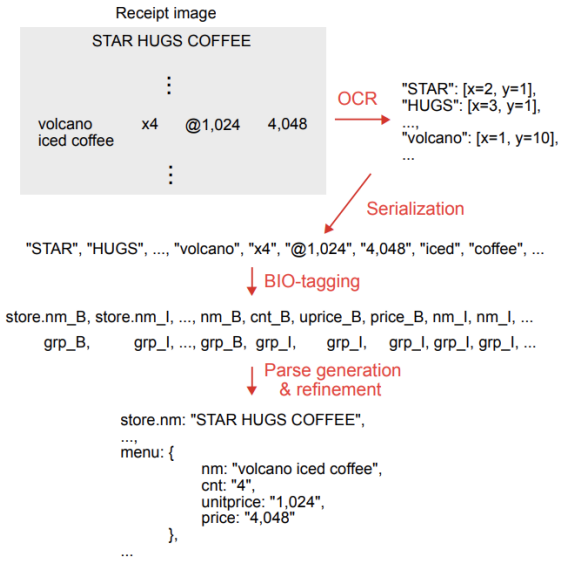
Aligned Gold Standard

The law in that cafe was severe , for cowards and runaways were not only degraded from all honors, but it was also a disgrace…

Challenge : **Find** and **correct** OCR errors

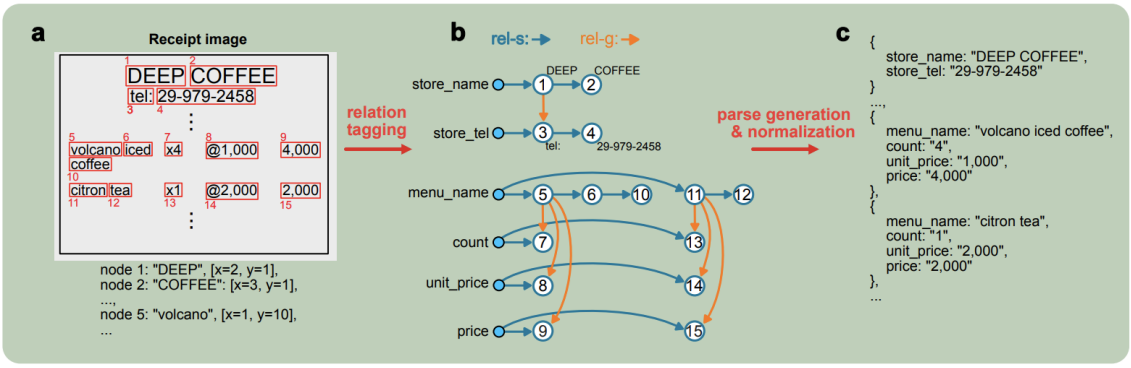


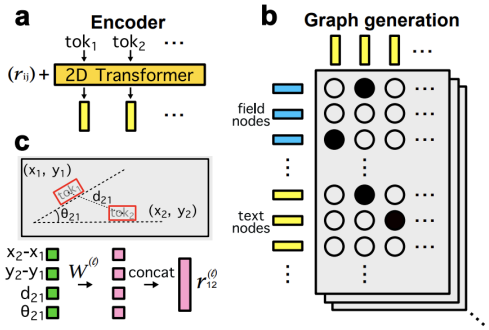
- [NLP] OCR Parsing (Feb. 2019 ~ Present)
  - OCR Parsing task takes OCR results of a given document image as an input and outputs the recognized information in a pre-defined structured form
  - There are 4 models developed for this task
  - This model is currently deployed in the real world business (CLOVA OCR)

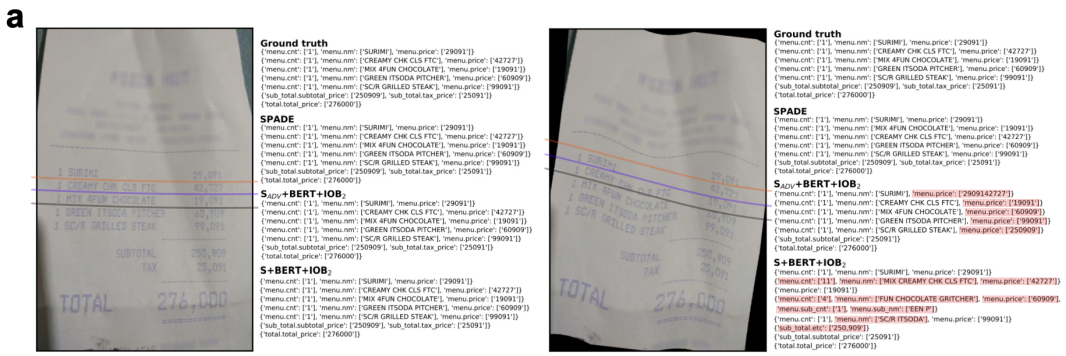


- (NLP, OCR Parsing) Tagging-based model (v1)
  - The model is a fine-tuning model based on BERT.
  - In order to use OCR outputs as input to the BERT model, the OCR outputs, which are spatially distributed, are serialized first
  - The serialized OCR outputs are converted into tokens and tagged as described above.
  - (Role) Model design / Evaluation system / Model training pipeline design
  - (Publication)
    - W. Hwang, S. Kim, M. Seo, **J. Yim**, S. Park, S. Park, J. Lee, B. Lee, H. Lee,"Post-OCR parsing: building simple and robust parser via BIO tagging." Workshop on Document Intelligence at the Conference on Neural Information Processing Systems
      (NeurIPS 2019 Workshop DI)

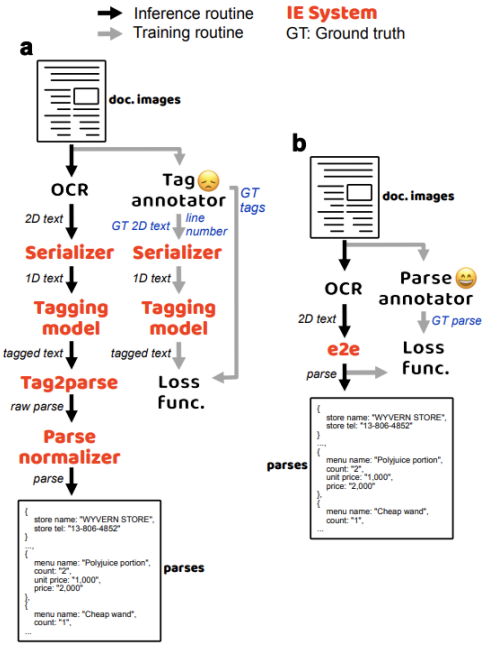
- (NLP, OCR Parsing) Graph-based model (v2)
  - The previous model (v1) is dependent on the token serialization so that the model cannot handle more noisy models
  - In order to make the model able to handle the spatial dependency of input tokens, we reformulate the task as a graph generation problem.
  - (Role) Data processing / Model design
  - (Publication)
    - W. Hwang, **J. Yim**, S. Park, S. Yang, and M. Seo. "Spatial Dependency Parsing for Semi-Structured Document Information Extraction." The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021)

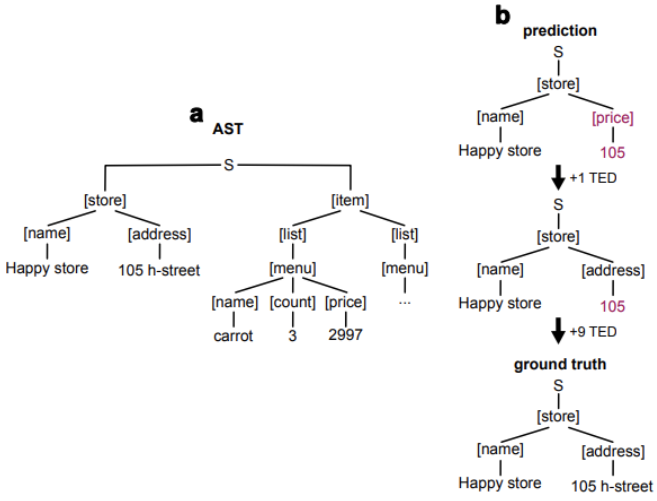




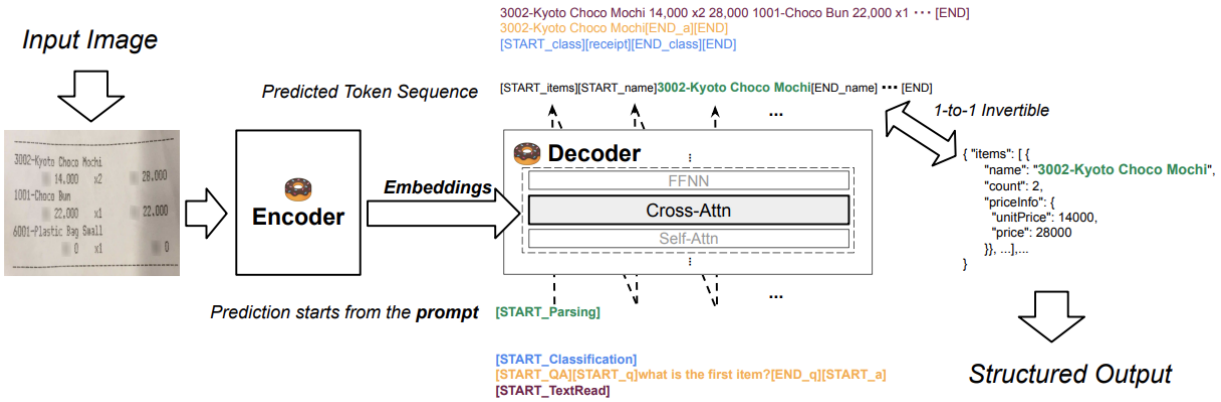


- (NLP, OCR Parsing) End-to-end parsing model (v3)
    - The previous model (v2) handles the noisy input data. However, the required data for model training is hard to be collected due to the difficult annotation
    - The model runs in an end-to-end manner. The input to the model is still the same, but the output is just a sequence, which is tree-structured, that can be converted into a desired information with a specific structure.
    - (Role) Model design / Human evaluation system setup
    - (Publication)
        - W. Hwang, H. Lee, **J. Yim**, G. Kim, M. Seo. "Cost-effective End-to-end Information Extraction for Semi-structured Document Images." The 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP 2021)





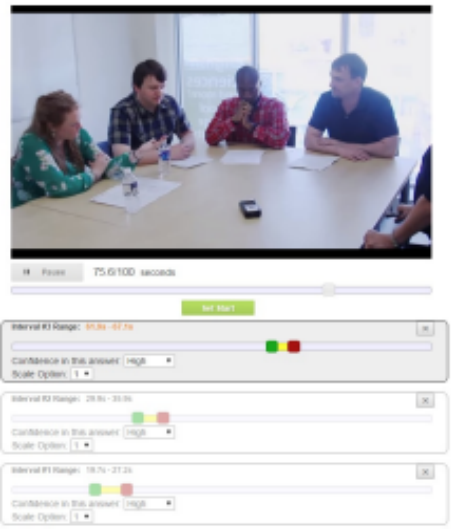
- (NLP, OCR Parsing) End-to-end parsing model without OCR (v4)
    - All previous models (v1, v2, and v3) are dependent on OCR outputs.
    - The new model does not use OCR and thus takes an image itself as an input while the generating output is still the same.
    - (Publication)
        - Geewook Kim, Teakgyu Hong, Moonbin Yim, Jinyoung Park, **Jinyeong Yim**, Wonseok Hwang, Sangdoo Yun, Dongyoon Han, Seunghyun Park, "Donut: Document Understanding Transformer without OCR," arXiv preprint arXiv:2111.15664 (2021).

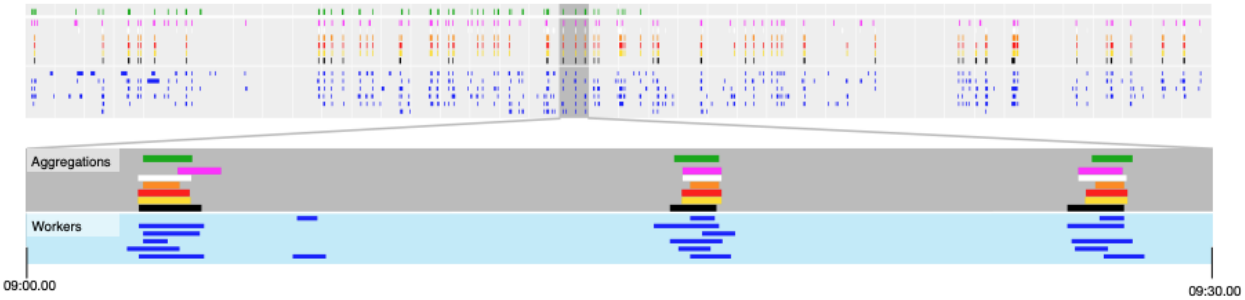


- [NLP] Other research topics that were either superficially covered or paused
    - Math word problems
        - The task is similar to Semantic parsing in a manner to convert an input natural sentence into a certain output form. In Semantic parsing, the output is a pre-defined logical form while in Math word problems, it is a mathematical expression with numbers
    - Semantic parsing - weakly supervised
        - Conducted some experiments but had to stop at that time due to limited computing resource for reinforcement learning
            - Many CPUs were required for multiprocessing
    - Clustering-based convergence from roughly annotated data (pause)
        - Starting with roughly annotated data, make the data have roughly all-annotated on newly added labels
            - Label feature-based clustering
    - Few-shot OCR parsing experiments (pause)
        - Approaches: one-pretrained model / double-pretrained model
        - data distribution convergence experiments based on labels

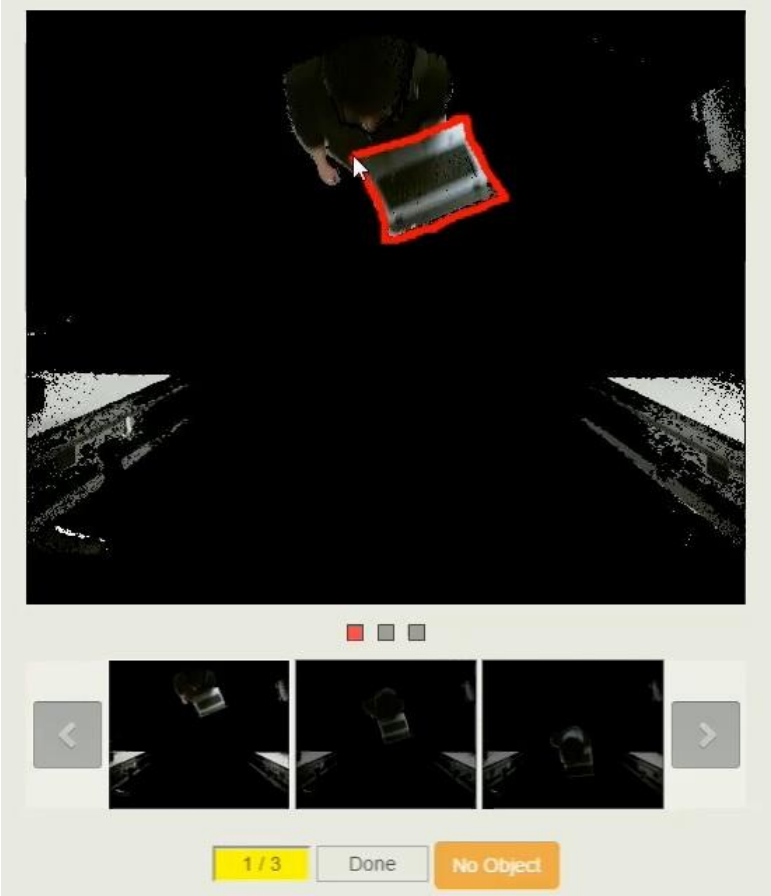
Crowdsourcing / HCI / Data annotation tools (2015~2016, 2018, 2020 ~)

- (HCI, Crowdsourcing) Video annotation tools (Sep. 2015 ~ Dec. 2015)
  - The video annotation management tool for the user
  - The user can generate microtasks for annotators by segmenting the target video
  - The segmented videos are annotated by a number of annotators and aggregated to be an annotation for the entire video.
  - (Role) Made the tool by myself
  - (Skills) Javascript, JQuery, PHP, MySQL, Amazon Mechanical Turk (AMT), AWS
  - (Publication)
    - J. Yim, W. Leung, J. Jasani, E. Lim, A.M. Henderson, M. Gordon, D. Koutra, J.P. Bigham, S.P. Dow, W.S. Lasecki. "Coding Varied Behavior Types Using the Crowd." (CSCW 2016 Demo)

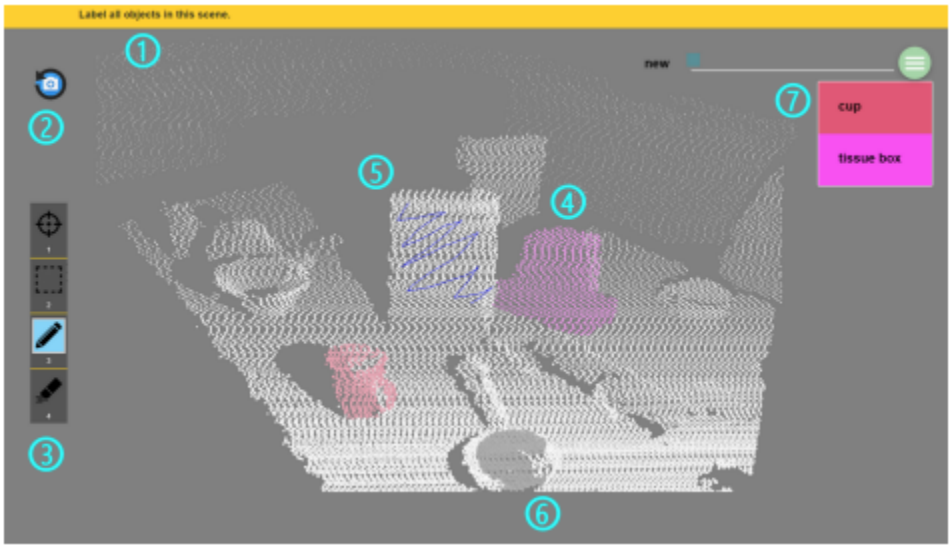




- (HCI, Crowdsourcing) Image object annotation tool for building an ML system
  (May. 2016 ~ Aug. 2016)
  - This project was done during the Internship at Bosch Research in Pittsburgh
  - Developed an image object annotation (free-drawing sketch) tool to collect data for training a machine learning object detection system
  - Microsoft Kinect based 3D point cloud data is hard for annotators to understand straightforwardly.
  - We instead provide normal pictures and ask annotators to annotate on normal corresponding images.
  - The annotations on normal images will be used for 3D point clouds to have pseudo-annotations and those will be used for constructing a recognition ML model.
  - (Role) Made the tool by myself
  - (Skills) Javascript, AngularJS, Sketch.JS, MongoDB, Amazon Mechanical Turk, AWS



- (HCI, Crowdsourcing) Object annotation tool for 3D point cloud in a real-time in multi-annotator cooperation (Apr. 2016 ~ Dec. 2016)
    - 3D Point cloud data is complex. 3D-rendered data annotation tools are still complicated to learn.
    - To overcome the limitations in annotating 3D point clouds, we recruit multiple crowdworkers in real-time and throw them to the cooperating environment.
    - We also investigated how to efficiently manage multi-users
    - (Role) Made the tool almost by myself
    - (Skills) Javascript, Three.JS, MongoDB, Amazon Mechanical Turk, AWS
    - (Publication) S. Gouravajhala, **J. Yim**, K. Desingh, Y. Huang, O.C. Jenkins, W.S. Lasecki. "EURECA: Enhanced Understanding of Real Environments via Crowd Assistance." (HCOMP 2018)

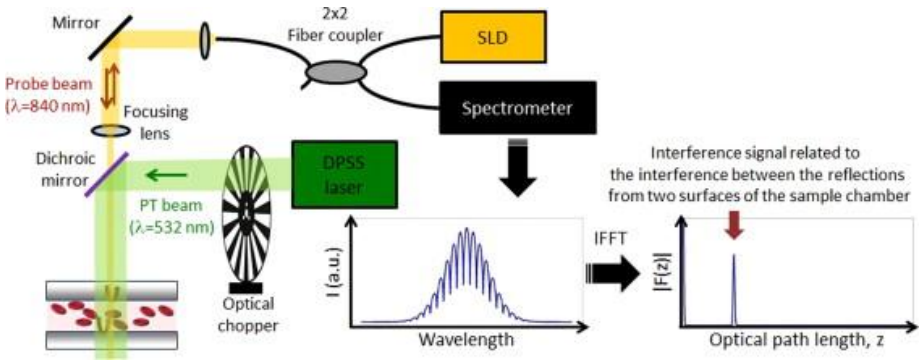


- (HCI, Crowdsourcing) Data acquisition process design for Table QA system (Jun. 2018 ~ Sep. 2018)
    - How to overcome challenges for collecting annotations for semantic parsing data
    - Collected (natural language - logical form) pairs for each table to train a model for semantic parsing on Relational Database (RDB) data

- A human-AI hybrid workflow for automated ML model creation (Jun. 2020 ~ Present)
  - Some tasks that we want to solve through training ML models require a complex form of data annotation that is even hard for task designers.
  - We want to leverage crowd power in a sense that ML models can mimic how humans think.
  - Also, we provide a semi-trained AI model output for non-experts to better make annotation rules and annotation itself.
  - (Skills) Vue.js, PostgreSQL, Amazon Mechanical Turk, AWS

Other research topics - Biomedical Optics (2013 ~ 2014)

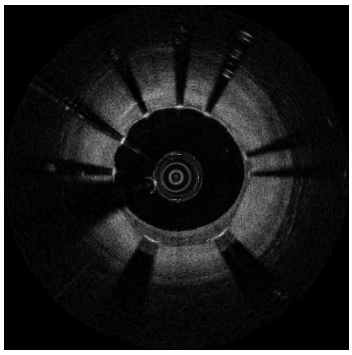
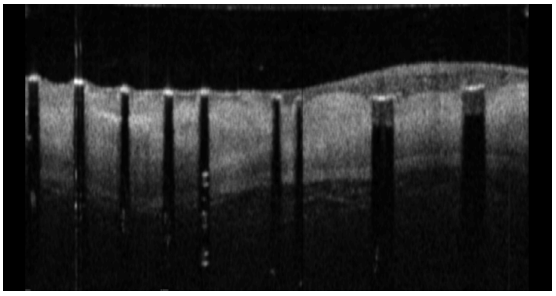
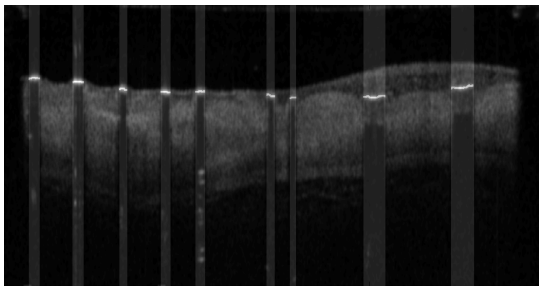
- Photothermal spectral-domain optical coherence reflectometry for direct measurement of hemoglobin concentration of erythrocytes
    - (Publication) - My first published journal paper as the first author (I wrote it when I was an undergrad.)
        - **Jinyeong Yim**, Hun Kim, Suho Ryu, Sungwook Song, Hyun Ok Kim, Kyung-A. Hyun, Hyo-Il Jung, and Chulmin Joo. "Photothermal spectral-domain optical coherence reflectometry for direct measurement of hemoglobin concentration of erythrocytes." Biosensors and bioelectronics 57 (2014): 59-64.
    - Developed a biosensor that can detect hemoglobin concentration without *in vivo* process
    - (Role) Designing experiments / Experiments / Data analysis

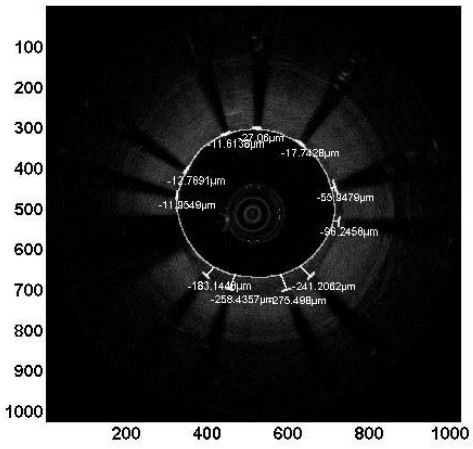


- Conducted clinical experiments about hemoglobin assay in anemic patients with a photothermal spectral-domain optical coherence reflectometric sensor (the one above)
    - Clinical paper using the previous optical method for hemoglobin assay
    - S. Song, H. Kim, **J. Yim**, C. Joo, H. O. Kim. "Evaluation of The New Developed Photothermal Spectral Domain Optical Detection Method for Hemoglobin Concentration for Pre-donation Screening In Blood Donors." Vox Sanguinis, 107 (2014) 90-91
    - H. Kim, S. Song, **J. Yim**, H. O. Kim, C. Joo. "Hemoglobin Assay in anemic patients with a photothermal spectral-domain optical coherence reflectometric sensor." Clinica Chimica Acta, 439 (2015) 71–76

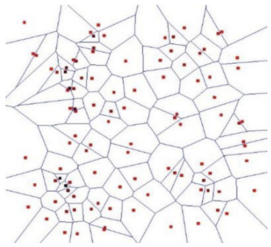
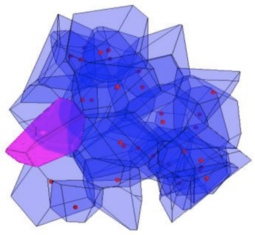
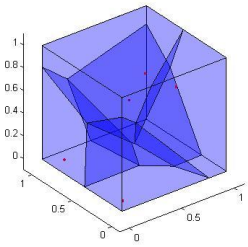
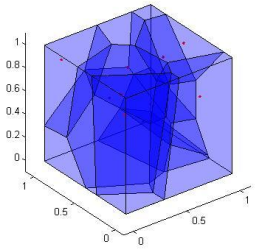
Other research topics - Image processing and simulation (2012 ~ 2013)

- Stent Detection in Intravascular OCT images (Optical Coherence Tomography)
  (Jun. 2012 ~ Sep. 2012)
  - Image processing techniques based on the gradient-method
  - We want to detect implanted stents in a blood vessel
  - Developed a novel image processing method (multi-directional gradient-method with channel conversion) for stent detection
    - Tried to proceed the project to make a publication but failed due to the conflict with the hospital that provided these medical images

- Optimizing agent distribution in 3-D space (Jan. 2013 ~ Mar. 2013)
    - Agent modeling based on Voronoi diagram
    - Provided a simulation tool for real-world multi-sensor networks
    - Simulation with Matlab

- Nano particle separation image processing (Apr. 2013 ~ Jun. 2013)
    - Internship at La Jolla Bioengineering Institute
    - Gaussian filter, gradient method
    - Building an automated system to count nanoparticles
    - Challenging parts were overlapped particles → convex hull method