Part 1: Inference in Neural Networks

-deep learning model은 test time에 보지 못한 novel data에 대해서도 정확한 예측을 해야 함

-novel data를 training에 사용할 수도 없지만, 만약 사용하더라도 제대로 학습하기 어려움

-학습 초반에 novel data를 사용하게 되면, novel data가 높은 정보량을 가지기 때문에 잘못된 방향으로 학습을 하게 됨

-반대로 학습 후반에 novel data를 사용하더라도, catastrophic forgetting으로 인해 초반에 학습한 normal data는 잊고 novel data에만 집중하게 됨

-따라서 이번 tutorial에서는 novel data를 inference time에 다룰 때를 주제로 함

-Novel data에 대해서 robust한 예측을 하려면, 학습한 분포 함수로 novel data를 나타내야 하고, novel data에서 추출한 정보를 활용할 방법이 필요함

-inference 시에는 data 하나만 주어지고 NN에 관한 정보를 알 수 없기 때문에, NN가 나타내는 likelihood 분포 L(theta)의 manifold에 직접 접근할 수 없음

-대신 fisher information을 이용해, gradient를 구해 분포 함수에 관한 정보를 얻을 수 있음

-x가 novel data이더라도, gradient는 x 주변에 대한 분포의 local information을 알려줌

Part 2: Explainability at Inference

-Explanation이란 모델이 예측에 사용한 important characteristics를 나타냄

-주어진 여러 question에 따라 explanation도 달라지는 contextual하고 relevant한 특성을 가짐

-Grad-CAM은 task-specific layer 이전의 마지막 conv layer activation에 대한 likelihood의 gradient를 통해 계산하며, task에 대해 generalization이 가능함

-manifold의 local information을 제공해주며, 또한 class P에 대한 “Why P?” 질문에 대한 explanation을 의미함

-CounterfactualCAM은 negative gradient를 통해 구하고, “What if region P were absent in the image?”에 대한 explanation을 제공함

-ContrastCAM은 likelihood 대신 class P와 Q 사이의 loss의 gradient를 “Why P, rather than Q?”에 대한 explanation을 제공함

-예를 들어, CounterfactualCAM의 경우 No-Left 표지판 image의 “Why No-Left, rather than STOP?” 질문에 대한 activation map이 STOP 글자가 아닌 edge 부분을 가리키는데, 이는 train set에서 STOP 표지판만 edge를 갖기 때문임. 모델이 갖고 있는 이러한 bias를 inference 때 알 수 있음

-또한 ContrastCAM은 “Why P, rather than Q?” 질문에 대한 activation map을 통해, inference 시에 class P에 속하는 single image만으로도 class Q에 관한 정보를 얻을 수 있음

-이 방법들을 활용해 introspection learning을 NN에 적용한 연구가 진행되었음

-introspection은 자신의 생각을 관찰하는 방법을 뜻하며, visual sensing으로 input을 감지하고, 이후 logical, targeted question을 통한 reflection의 과정을 거쳐 결론을 도출해 냄

-technical definition으로 정의하면, input x에 대한 label이 모델이 예측한 class P가 아닌 다른 class Q로 주어졌을 때의 모델 분포 f(x)의 parameter 변화를 의미함

-마지막 fc layer weight에 대한 loss J(P,Q)의 gradient를 class Q에 대한 introspective feature로 생각할 수 있음

-well-trained network의 경우, class 확률이 서로 dependent하기 때문에 gradient가 robust함

-이를 이용해 input x에 대한 모든 class들의 feature를 얻어 MLP를 학습해, 모델 끝에 추가해 줌

-그 결과 OOD data accuracy를 뜻하는 generalization 성능과, prediction accuracy와 confidence 차이를 뜻하는 calibration loss가 낮아짐

Part 3: Uncertainty at Inference

-model uncertainty란 모델이 data를 모델링하지 못하는 경우를 의미하며, uncertainty estimation이란 모델 자신이 모른다는 사실을 얼마나 알고 있는 지 측정하는 일을 뜻함

-uncertainty estimation을 위해 model ensemble을 이용해 output variation을 구하는 방법이 있지만, trained model이 여러 개 필요함

-novel input data feature와 class centroid 간 거리를 구하는 방법은 single network, single pass만 이용하지만, 추가적인 dataset이 필요함

-Dropout을 통해, N번의 forward pass를 거쳐 나온 N개의 logits을 이용하는 방법도 있음

-이 때, 예측 평균값의 엔트로피가 predictive uncertainty를 의미하고, 각 예측값의 엔트로피의 평균값이 data 자체의 noise를 나타내는 aleatoric uncertainty를 의미함

-두 값의 차이를 계산해 모델의 uncertainty를 뜻하는 epistemic uncertainty를 구할 수 있음

-gradient로 train manifold와 novel data 사이의 거리를 표현할 수 있고, 거리가 클수록 uncertainty 또한 크다고 볼 수 있음

-앞서 언급한 introspective learning에서는 gradient가 learned manifold에 속한 local area 사이의 차이를 의미했다면, 여기에서는 gradient가 novel data가 주어졌을 때의 manifold의 local change를 의미하기 때문에 uncertainty를 표현할 수 있게 됨

-예측 logit과 multiple class label로 된 confounding label 간의 loss의 parameter에 대한 gradient를 구할 수 있음

-trained data와 untrained data에 대해 gradient의 L2 norm을 계산해보면, trained data는 0에 가까운 반면, untrained data는 큰 값을 보이며, 따라서 uncertainty가 높다는 것을 알려줌

-이 방법은 adversarial detection, corruption detection에서도 높은 성능을 보여줌

Part 4: Intervenability at Inference

-Intervenability란 모델이 human intervention을 얼마나 받아들일 수 있는지를 의미하며, causality, privacy, interpretability, prompting, benchmarking의 5가지 관점에 따라 다양하게 해석 가능함

-causality 관점에서 intervention이란 모델 예측의 causal factor를 test하기 위한 data manipulation으로 볼 수 있음

-Privacy 관점에서는 machine unlearning과 같이 trained model에서 privacy data에 관한 knowledge를 없애기 위한 사후 개입을 의미함

-Interpret 관점은 앞서 설명한 explanation을 평가하기 위한 data intervention을 의미함

-Prompting 관점에선 foundation model과 user prompt와의 상호작용을 의미하고,

-benchmarking 관점에서는 모델의 generalization 성능을 test하기 위한 data intervention을 뜻함

-대부분의 intervention은 user가 model을 이해하고, 상호작용을 하기 위한 목적에서 쓰이므로 positive하지만, negative intervention이 사용되기도 함

-예를 들어, 의료 데이터로 학습된 target model을 이용해 특정 data가 모델 학습에 사용됐는지 추론하는 attack model을 학습할 수도 있음

-intervention을 통한 causality 평가를 하려면 3단계 rule을 거쳐야 하지만, 여러 문제가 발생함

-먼저, rule 1에 따라 causal feature insertion / deletion을 통해 평가해야 하지만, causal feature를 정확한 픽셀 단위로 구분할 수 없고, 모델도 causal model이 아님

-이후 rule 2와 rule3에서는 causal factor 이외의 다른 factor들을 바꿔 test하는 과정을 거치지만, 이때 모든 pixel에 개입하기도 어렵고, model uncertainty 때문에 결과를 신뢰하기 어려움

-6000개의 random masking 처리한 image의 예측 확률의 weighted sum으로 explanation을 얻는 RISE explanation과, 예측 확률의 변화를 이용하는 SHAPE explanation 방법이 있음

-평가 metric은 SHAPE explanation이 더 높지만, 사람이 봤을 때는 더 해석하기 어렵기 때문에 causality 평가에 쓰기 힘듦. 또한 모든 6000개 mask만 사용하기 때문에, 모든 intervention을 다루지 못함

-대신, explanation evaluation 방법을 몇 가지 생각해 볼 수 있음

-첫 번째로, explanation map으로 masking한 image를 model에 넣어, 예측 score에 따라 explanation을 평가할 수 있지만, 넓은 evaluation map을 선호하게 되는 문제가 있음

-두 번째로, pixel-wise deletion으로 heatmap의 가장 중요한 pixel부터 하나씩 제거해가며 예측 확률을 얻을 수 있고, AUC가 낮을수록 좋은 explanation을 의미함

-하지만 사람이 pixel 단위로 인식하지 않기 때문에, structure-wise한 방법이 필요함

-따라서 good explanation일수록 최소한의 bit로 structure-rich feature를 인코딩할 것이라는 가정 하에 structure-wise deletion 방법을 적용해 볼 수 있음

- heatmap에서 threshold보다 낮은 pixel을 제거한 masked image와 original image 간의 Huffman code 비율에 따른 prediction 확률을 그래프로 나타낼 수 있음. 그래프가 왼쪽에 치우칠수록, 적은 bit로 많은 정보를 encoding했기 때문에 good explanation이라고 판단함

-하지만 이 방법 또한 직관에 반하는 결과를 나타내는 문제가 있음

-따라서 아직까진 explanation evaluation 전략 중 어떤 방법이 가장 좋다고 판단할 수 없음

-explanation의 uncertainty를 측정해 모델 판단을 해석하려는 연구도 진행됨

-explanation masked image를 Sx라 할 때, 모델의 predictive uncertainty를 V[y|Sx]로 표현 가능함

-variance decomposition을 하면, 이 값은 V[E(y|Sx)]와 E(V[y|Sx]) 두 값의 합으로 표현됨

-첫 번째 값은 masked image와 original image 간의 예측 확률 차이를 의미하며, 좋은 explanation일수록 첫 번째 값을 줄일 수 있음

-두 번째 값은 explanation 중 모델 예측에 사용되지 않은 residual explanation과 관련되어 있으며, uncertainty in explanation을 계산하기 위해 사용됨

-따라서 explanation 자체는 첫 번째 값에만 관련되므로 모델 uncertainty를 부분적으로만 줄일 수 있기 때문에, explanation의 uncertainty도 이해하는 것이 중요함

-예를 들어, 왼쪽 강아지 그림에서 explanation은 코 부분을 턱 부분만큼 강조하지 않지만, 코 부분은 다른 강아지 품종과 구분할 때는 중요하기 때문에, uncertainty of explanation을 나타냄

- 모든 class에 대해 앞서 언급한 contrastCAM처럼 loss J(P,Q)에 대해서 contrastive explanation map을 구하고, 분산을 계산해 uncertainty map을 얻을 수 있음

-이 때 uncertainty를 측정하기 위한 metric으로 explanation map과 uncertainty map 간의 IoU를 이용하거나, uncertainty map 자체의 signal-to-noise ratio를 이용할 수 있음

-IoU가 클수록 uncertainty가 높고, SNR은 작을수록 uncertainty가 높음을 의미함

-