**Your Diffusion Model is Secretly a Zero-Shot Classifier**

**1. Introduction**

- generative model은 데이터 분포를 정확하게 모델링하는 목표를 갖기 때문에, 본 논문은 이러한 장점을 활용해 diffusion model을 image classification에 적용한 zero-shot classifier인 Diffusion Classifier를 제안함

- diffusion model과 같은 conditional generative model로 모든 클래스 c에 대한 p(xlc)의 ELBO를 계산할 수 있고, 적절한 prior p(c)를 선택했을 때, 베이즈 이론에 따라 p(clx)를 얻어 classification을 수행할 수 있음

**2. Related Work**

- discriminative task의 경우, discriminative model은 주어진 task의 decision boundary p(ylx)를 학습하지만, generative model은 데이터 분포 p(x) 또는 p(xly)를 학습하며, robustness와 calibration에 장점을 보임

- 이전 연구들은 두 모델을 함께 사용해 p(x,y)를 학습하거나, generative representation을 fine-tuning하는 반면, 본 논문은 diffusion model을 직접 image classifier로 사용한다는 차이점이 있음

**3. Method**

**3.1. Diffusion Model Preliminaries**

- diffusion model은 가우시안 분포 q(xt-1lxt)에 따라 noise를 더하는 markov chain 형태의 forward process를 바탕으로, reverse process의 분포 q(xtlxt-1)를 근사하는 p(xt-1lxt)를 학습함

- training objective는 p(x0)를 MLE하는 문제가 되고, 최종 loss는 다음과 같이 정리됨

- 이 중 Lt-1항을 minimize해야 하며, 정리하면 ELBO를 두 분포의 mean 차이가 되고, 최종적으로 timestep t의 gaussian noise (epsilon)을 학습하는 문제가 됨

**3.2. Classification with diffusion models**- 1)conditional generative model로 logp(xlc) ELBO 사용

2)p(c)로 uniform prior를 사용

3)t:uniform, epsilon:N(0,1)로 샘플링해 기댓값의 Monte Carlo 추정 값 계산

- class 중 가장 낮은 epsilon error 갖는 class가 정답 class가 됨

**3.3. Variance Reduction via Difference Testing**

- 각 clss의 error 기댓값을 정확히 예측하기 위해 너무 많은 sample이 필요하므로, 그 대신 class간 error 차이를 구해 classification에 필요한 상대적 차이를 계산함

- 따라서 모든 class에 대해 고정된 sample set을 사용해 error를 예측

- (Fig2) 실제로 epsilon에 따라 절대적인 error값에 큰 차이를 보이지만, class 간 상대적인 차이는 epsilon 값이 달라도 일관적임

- algorithm: image sample x를 입력 받아, T번의 trial마다 timestep t와 epsilon 샘플링해 forward process로 xt 계산

- 각 class c에 대해, model에 x와 c 입력해 epsilon 구한 뒤 error 계산

- 모든 trial 끝난 뒤, mean error가 최소인 class 선택

- pretrained diffusion model만을 이용해, 학습 없이 classifier를 추출함

- stable diffusion model(TTI model): text, zero-shot classifier,

class-conditional diffusion model(DiT): class-index, standard classifier

**4. Practical Considerations**

- (Fig3) 하나의 timestep만을 사용한 evaluation 결과, t가 500 근처에서 가장 높은 정확도를 보임

- (Fig4) 이를 고려해, t=500 근처 값을 자주 샘플링하는 전략을 포함한 다양한 전략을 비교한 결과, (1,1000) 사이의 uniform 샘플링이 가장 높은 정확도를 보였음

- 각 image에 대해 (class 개수 C) x (trial 횟수 N) 만큼의 총 trial이 필요하므로, 아니라고 확신할 수 있는 class에 대해 error 계산을 중단하도록 했음

- (alg2) stage를 나눠, 각 stage마다 정해진 trial을 진행함 / stage가 지날수록, 작은 mean error를 갖는 몇 개의 class만 남게 됨

**6. Experimental Results**

- TTI diffusion model인 stable diffusion을 이용한 zero-shot classification 결과, SD feature와 label로 supervised하게 학습한 경우에 비해 전체적으로 높은 성능을 나타냄

- 다른 dataset으로 학습되었지만, CLIP과도 유사한 성능을 보였음

- stable diffusion 학습에 사용된 dataset을 고려했을 때, 대부분의 test data가 OOD data이므로, train set diversity를 증가시켰을 때의 성능 향상이 예상됨

- Compositional reasoning 결과, 모든 category에서 CLIP에 비해 높은 성능을 나타냄

- DiT를 사용한 OOD detection 경우, 다른 dataset은 ViT-B/16보다 못 미치는 성능을 보였지만, ResNet이 예측하기 어려워하는 image로 구성된 ImageNet-A에서 높은 effective robustness를 보임

- 또한 DiT 학습에 간단한 augmentation만 사용됐고, regularization은 사용되지 않은 점을 감안했을 때, overfitting 문제에 강함

- objective로 timestep에 따라 weight가 변하는 exact variational lower bound를 사용해 variance도 학습한 경우보다, fixed variance를 사용한 uniform weight 경우가 더 높은 성능을 보였음

**Generative Classifiers Avoid Shortcut Solutions**

**1. Introduction**

- shortcut learning: 의도하지 않은 (spurious feature) decision rule (ex: image background, text 특정 단어)에 따라 모델이 판단하도록 학습해, 작은 distribution shift에도 generalize하지 못하는 문제

- shortcut learning에 취약한 discriminative model과 달리, generative model은 input의 causal, spurious feature 전부 모델링하기 때문에 shortcut solution을 피할 수 있다고 본 논문은 가정함

- 여러 가설을 test해, 이런 가정이 실제로 맞다는 것을 보임

- distribution shift benchmark에서 generative classifier를 실험해, ood dataset에서의 높은 성능과 overfitting 문제에 강하다는 것을 증명함

-

**2. Related Work**

- 이전 연구들에서 hybrid generative-discriminative classifier를 test-time adaptation에 사용하거나, diffusion model로 zero-shot classification을 수행해 synthetic corruption, shape feature에 대한 robustness를 보였음

- 본 논문은 이와 다르게 realistic distribution shift에서 나타나는 spurious correlation에 대한 robustness를 다룸

**3. Preliminaries**

- 본 논문은 subpopulation shift와 domain shift 두 종류의 distribution shift를 고려함

- subpopulation shift: train, test set의 subpopulation 비율 차이 / subpopulation을 표현하는 spurious feature를 모델이 잘못 학습할 경우 (celebA), 다른 spurious feature를 갖는 minority group 예측 정확도에 문제가 생김

- domain shift: train, test set의 분포 차이로 인해, train set 예측에 사용되었던 spurious feature가 test set에서 도움을 주지 못함

- overparameterized discriminative classifier의 경우, 모델이 spurious feature를 활용해 majority group loss를 minimize하도록 fitting 된 이후, spurious feature가 도움되지 않는 minority example는 memorize하면서 overfitting되는 문제가 발생함

- 이를 해결하기 위해 data rebalancing, loss reweighting과 같은 방법이 제안되었지만, 학습이 복잡해지며, distribution shift에 대한 정보를 알아야 하기 때문에 realistic dataset에 적용하기 어려움

**4. Generative Classifiers**

- image classification의 경우, class-conditional diffusion model을 학습한 후, Diffusion Classifier 연구처럼 denoising error가 가장 작은 class를 정답으로 예측함

- text classification의 경우, autoregressive Transformer model을 이용함

- training: text sequence x에 대해 text modeling을 수행하도록 학습하고, 이 때 BOS token 자리에 label에 해당하는 class token을 입력함

- inference: text 앞에 각 class token을 넣어, class 개수 C회의 forward pass를 진행해, 전체 예측 sequence의 cross-entropy loss가 최소가 되는 class를 정답으로 예측함

- 따라서 추가적인 architecture, hyperparameter, data augmentation 없이 간단하게 학습 가능함

**5. Experiments**

- spurious feature 문제 개선 목적의 discriminative classifier baseline들과 비교한 결과, domain shift dataset에 대한 OOD accuracy, subpopulation shift dataset에 대한 worst-group accuracy 둘 다 더 높은 정확도를 나타냈고, ID accuracy 또한 더 높았음

- 따라서 causal feature 학습과 overfitting 문제에서 더 뛰어나다는 것을 알 수 있음

- generative, discriminative classifier family 간 비교를 했을 때, generative classifier의 ID accuracy가 낮더라도 OOD accuracy는 더 높았음

- effective robustness를 보이므로, scale 했을 때 generative model이 더 효과적일 것이라 예상 가능함

- epoch에 따른 majority, minority group의 평균 gradient norm의 변화를 비교했음

- discriminative model의 majority group gradient는 학습 초반부터 빠르게 줄어들고, minority group gradient는 초반에는 높다가 후반에 사라짐

- 반면 generative model은 두 group 모두 비슷한 gradient를 가지며, 조금씩 증가함

- 따라서 discriminative model은 spurious feature를 학습해 train set에 confident한 예측을 하므로 causal feature를 학습하기 위한 gradient를 잃게 되고, generative model은 input의 모든 feature를 학습하므로 일관적인 gradient를 가진다는 가설을 뒷받침함

- autoregressive model의 objective로 discriminative objective p(ylx)에 unconditional generative objective p(x)를 추가해도 성능이 차이가 없는 것으로 보아, generative classifier가 일반적으로 더 나은 feature p(x)를 학습한다는 가설은 기각함

- (Fig5) classifier는 class 간 p(xly)만 필요하기 때문에, sample quality가 낮더라도 높은 classification accuracy를 가질 수 있음

- 하지만 validation diffusion loss는 class-balanced accuracy와 높은 관련성을 보임

- causal feature인 hair color와 shortcut feature인 gender에 대해, 학습이 진행되면서 minority group인 blond men sample quality가 높아지는 것으로 보아, 두 feature가 낮은 상관관계를 가지는 방향으로 모델링되는 것을 알 수 있음

**6. Illustrative Setting**

- generative classifier의 OOD behavior에 대한 인사이트를 얻기 위해 synthetic data를 설정함

- spurious, causal feature가 같은 위치가 majority group이 됨

- spurious feature scale beta와 noisy feature variance sigma 값의 변화에 따른 generative, discriminative classifier의 ID, OOD accuracy 성능을 비교해, 네 개의 generalization phase로 구분함

- 1) sigma 클 때, discriminative model은 noise feature에 overfitting됨

- 2) sigma 작음

- 3) overfitting을 발생시킬 만한 noise가 있지만, 그보다 spurious feature 영향이 더 커 discriminator가 ID accuracy는 높지만, spurious feature에 의존해 OOD accuracy는 낮음

- 모든 상황에서 특정 모델이 항상 우수한 성능을 보이지는 않지만, shortcut과 noise의 강도가 모두 높은 복잡한 데이터셋의 경우, generative model이 더 높은 성능을 발휘할 것으로 예상됨

- 주어진 dataset을 고려했을 때 어떤 method가 적절한 지에 대한 인사이트를 제공함