Part 1. Disentanglement and Compositionality in Computer Vision

- Disentangled Representation Learning은 ML model이 feature space를 semantic meaning을 갖는 independent underlying factor들로 나눠 이해하도록 학습하는 방식임

- DRL은 model의 explanability 외에도 task에 맞는 representation만 분리 가능한 generalizability와, generation 시에 factor를 조작할 수 있는 controllability도 향상시킬 수 있음

- DRL method는 우선 representation structure에 따라 dimension-wise DRL과 vector-wise DRL로 분류할 수 있음

- dimension-wise method는 각 generative factor를 latent vector Z의 한 개의 차원으로 표현하며, VAE와 GAN model 구조에 의존하는 초기 DRL 연구에서 자주 사용됨

- 또한 fine-grained factor를 갖는 synthetic, simple dataset에만 적용할 수 있음

- vector-wise method는 각 factor를 하나의 latent vector로 표현하기 때문에, 높은 정보량의 coarse-grained factor로 이루어진 복잡한 real-world dataset에 적용되고, 최근 더 주목받고 있는 방식임

-DRL method를 model type에 따라 VAE, GAN, diffusion based method로 나눌 수 있음

- VAE-based model 중 dimension-wise method인 beta-VAE는 KL divergence loss 항에 beta coefficient를 추가해, beta가 클 때 disentangled representation 학습에 높은 가중치를 두게 됨

- 또한 reconstruction과 disentanglement 간 trade-off를 고려해 KL term에 information capacity를 학습에 따라 늘어나도록 한 method 이외에도, 기존 VAE에 regularizer를 추가한 여러 변형 method들이 제안됨

- GAN-based model dimension-wise DRL method인 InfoGAN은 생성자 input에 noise z와 함께 latent variable c를 넣고, 생성자가 c의 정보를 잃지 않도록 mutual information term I(c;G(z,c))를 추가함

- 더 나아가, InfoGAN-CR은 InfoGAN에 contrastive regularizer를 추가해 생성자에서 샘플링한 image들로부터 fixed dimension을 찾아내도록 했고, PS-SG GAN은 각 latent dimension의 focused area를 가리키는 spatial mask를 학습함

- DR-GAN은 vector-wise DRL을 적용했고, pose-invariant face recognition을 위해 identity와 pose를 각각 나타내는 두 개의 vector를 사용함

- diffusion-based model 또한 제안되었는데, 그 중 하나인 Disenbooth는 denoising 과정에 identity-preserved part, identity-irrelevant part로 분류하는 disentanglement task를 추가함

- 지금까지 언급한 DRL methods는 generative factor들을 parallel하게 가정하는 flat DRL method지만, hierarchical DRL은 factor를 계층적 abstraction level로 나눠 분류함

- FineGAN은 hierarchical level마다 하나의 latent representation을 사용해 background, parent, child code로 나눠 fine-grained object generation을 수행함

- Compositional learning은 model이 data를 여러 concept 간의 composition으로 이해하도록 학습하는 것을 의미함

- scene graph는 compositionality를 적용한 대표적인 방식이고, image를 object, attribute, relation으로 표현해 image retrieval, zero-shot recognition, object localization 등의 task에 도움을 줌

- disentanglement와 compositionality를 함께 활용해 다양한 task에 적용하기도 함

- image captioning으로 disentangled text를 얻어, image feature와 text feature alignment를 통해 각 text에 맞는 visual concept을 매칭하는 방식으로 image editing을 수행할 수 있음

- hybrid-distortion을 feature disentangle module를 이용해 distortion마다 다른 feature로 분리되도록 하고, content information을 추출해 aggregation하는 방식으로 image restoration에 적용 가능함

Part 2. Disentangled Model-based Visual Concept Learning

(1)

- group theory를 적용해 DRL 학습을 unsupervised한 방식으로 진행한 Groupified VAE 연구를 진행함

- VAE와 유사한 구조의 model이 모든 vertices로 이루어진 기존 group에서 시작해 rotation, flipping motion을 적용한 permutation group을 출력하는 과정으로 이루어짐

- 기존 group에 mapping을 적용한 permutation group 또한 element가 같아야 한다는 structure constraint를 만족시키는 isomorphism loss를 VAE loss 이외에 추가함

- isomorphism loss는 commutative 특성을 갖도록 하는 Abel loss와 cyclic 특성을 만족시키도록 하는 Order loss로 구성됨

- representation space를 original VAE와 비교했을 때, groupified VAE가 더 높은 disentanglement performance를 보여줌

- dihedral group에서 영감을 받아 group action을 permutation으로 대체,

(2)

- 문제점: 1) disentanglement, generation 간 trade-off / 2) disentanglement constraint 불충분

-> 해결: pretrained generative model의 generation ability 유지하면서, disentanglement 학습

- latent space에서 각 factor variation의 방향을 찾을 수 있다면, DRL 학습을 할 수 있다 가정함

- navigator가 latent space에서 추출한 query direction을 fixed pretrained generator가 입력 받아image pair 생성함

- 그 후, disentangled encoder에서 추출한 pair 간의 variation을 통해 contrastive한 방식으로 variation space 학습함

- 결과적으로 navigator는 disentangled direction을 찾고, disentangled encoder는 disentangled representation을 찾는 방향으로 학습됨

(3)

- 다음 연구는 transformer를 이용해 pixel 단위로 visual concept을 학습함

- meta concept을 의미하는 concept prototype과 각 image에 해당하는 concept을 의미하는 concept token 두 embedding을 이용해 factor를 표현함

- concept tokenizer는 image token과 concept prototype 간 cross-attention을 거쳐 input data에 해당하는 concept token을 얻음

- 그 후, concept token과 image query를 detokenizer에 입력해 기존 image token으로 복원하게 됨

- 또한 disentangling loss는 임의의 concept token에 변화를 주었을 때, detokenizer와 tokenizer를 거쳐 그 변화를 예측할 수 있도록 만들기 때문에 disentanglement representation을 얻게 됨

- CLIP text encoder를 사용해 text editing, text decoding도 수행할 수 있음

(4)

- conditional diffusion model을 이용해 text-to-image generation을 할 때, cross attention을 통해각 text에 맞는 spatial region을 추출해내는 것에서 영감을 받아, diffusion model의 cross attention을 DRL을 위한 inductive bias로 활용한 연구를 제시함

- word token 대신 image encoder로부터 추출한 concept token이 conditional input의 역할을 함

- diffusion model의 중간 spatial feature를 query로 사용하고, concept token을 key, value로 사용해 cross attention을 진행하는 과정을 통해 disentangled representation을 학습함

- 또한 reverse diffusion process에서 time step이 작아짐에 따라 KL divergence는 점차 증가하는데, 이것은 AnnealVAE처럼 information bottleneck의 capacity가 증가하는 것을 의미함

- 이것도 추가적인 inductive bias로 작용해, concept token의 disentangled representation 학습에 기여함

Part 3. Equivariant and Disentangled Representation Learning

- equivariance는 어떤 function이 symmetry transformation에 대해 commute하는 성질을 갖는다는 의미로, transformation과 function의 적용 순서를 바꿔도 동일한 결과를 산출하게 됨

- disentangling으로 학습한 factors of variation을 바탕으로 latent code를 traversing하는 것 또한symmetry transformation으로 볼 수 있음

- 이를 바탕으로 Flow Factorized Representation Learning 연구는 disentanglement와 equivariance를 동시에 학습함

- image space에서의 symmetry transformation은 왼쪽 식의 우항과 같이 encoding, latent flow path, decoding의 과정으로 표현할 수 있고, 좌항과 우항이 같을 때 equivariance를 만족하게 됨

- 여기서 disentanglement 또한 만족시키려면, latent flow가 gradient의 distinct set에 의해 생성되어야 하며, image space의 flow로의 이동 비용을 최소화하는 optimal transport를 따라야 함

- 이러한 constraint를 고려한 VAE의 lower bound를 구해 model을 학습하게 되고, 그 결과 각 transformation type k에 해당하는 latent flow를 학습함

- 또한 OOD data에 대해서도 equivariance generalization 능력을 보여줌

Part 4. Disentanglement and Composition for AGI

- traditional generalization bound에 따르면, data가 많아야 하고, model이 간단해야 하며, test 분포와 train 분포가 같을 때 generalization 성능을 높일 수 있음

- compositionality와 disentanglement는 이 세 가지 측면 전부에 대해 도움을 줄 수 있음

- compositionality를 이용하면 복잡한 분포를 각 factor에 해당하는 간단한 분포 간의 곱으로 나타낼 수 있으므로, 적은 data만으로도 학습이 가능해짐

- 또한 model 단위 composition을 통해, detection model과 object-to-paragraph model을 합쳐 image captioning task를 수행하는 것처럼 flexible하게 test 분포에 적용할 수 있음

-