**1. Introduction**

- Auto-encoding-based method는 Unsupervised object-centric learning의 대표적인 방법 중 하나임 / bottleneck module에서 얻어진 inductive bias로부터 중요한 object feature를 우선적으로 encoding할 수 있음

- 이후, 기존 방식의 bottleneck에 slot을 추가한 구조로 발전함

- 우선, encoder가 input으로부터 ‘slot’이라 불리는 latent vector set을 출력함 / gaussian 분포로부터 얻은 initial slot vector에서 시작해, 여러 번의 iteration을 거쳐 최종 slot vector를 얻게 됨 / 한 번의 iteration은 다음과 같이 진행됨 / slot과 input feature 간의 attention을 바탕으로 input feature의 weighted mean을 구함 / 이 값을 input, slot vector를 초기 state로 사용해, GRU와 같은 recurrent function과 MLP를 거쳐 slot vector를 update해 줌

- 이후, decoder가 slot이 내포한 정보로부터 input을 복원하는 task를 수행하게 되며, 이 과정에서 object-centric representations를 학습할 수 있음

**2. Related work**

**Unsupervised object-centric learning**

- 초기 slot-attention bottleneck은 complex real-world scene 다루기 어려운 문제가 있었음

- 이를 해결하기 위해 다양한 측면의 개선 방법이 제안됨 / 1) encoder slot-attention module의 precision, stability를 향상시키는 방법(bi-level optimization, model 구조 변화, regularization losses) / 2) decoder 개선해 decomposition을 향상시키는 방식((SLATE) autoregressive transformer, diffusion-based methods) / 3) (DINOSAUR) reconstruction target으로 input image 대신 self-supervised pre-trained feature 사용해, 효과적으로 complex scene을 다룸

**3. Method**

- 본 논문에서는 encoder로 slot-based auto-encoder를 사용하며, image encoder와 slot-attention module로 구성됨 / image encoder는 DINO와 같이 self-supervised pre-trained feature encoder를 사용하고, 출력된 feature는 reconstruction target으로도 쓰임 / high-level feature 사용하기 때문에, input을 그대로 사용하는 것에 비해 object-centric representation 학습에 도움이 됨

- decoder로는 autoregressive transformer를 사용함 / yi-1까지의 feature와 모든 slot vector를 입력 받아, feature yi를 예측하는 과정이 진행됨

**3.1. Self-training via slot-attention distillation**

- n\*k 크기의 matrix A는 slot attention mask를 의미하고, k개 slot마다의 n개 feature patch의 assignment를 나타냄 / slot-attention module, decoder module 두 부분에서 slot attention mask를 얻을 수 있음

- slot-attention module에서는 임의의 initial slot vector에서 시작해, slot-to-patch cross-attention을 반복해 output slot vector를 얻는 과정이 진행됨 / slot attention mask Aslot은 마지막 cross-attention iteration에서 계산할 수 있음 / 이 때 key는 patch-wise feature로부터 얻고, query는 이전 iteration의 slot vector로부터 얻음

- decoder module은 slot vector에서 patch를 예측하기 위한 patch-to-slot cross-attention layer들로 이루어져 있음 / slot vector를 key로 사용하고, 이전 layer의 decoder input를 query로 사용해, 마지막 layer에서 decoder attention mask Adec를 계산할 수 있음

- 두 attention mask 중 decoder mask가 slot mask보다 object decomposition에 있어 더 좋은 결과를 보인 것을 이용해, 본 논문에서는 Two-stage training with slot-attention distillation 학습을 제안함

- first stage에서는 reconstruction loss만을 이용해 teacher model을 학습함

- second stage에서는 student의 slot mask가 teacher의 decoder mask와 같아지도록 하는 slot-attention distillation loss와 reconstruction loss를 사용해 student를 학습함

- teacher의 decoder mask At의 각 row에 argmax와 one-hot 연산을 적용해, 각 slot마다 하나의 patch에 해당하도록 변형된 A’t를 계산함 / 그 후, student mask As와 A’t를 Hungarian matching으로 k개 slot끼리 matching하고, 이 때 mask 간 IoU를 cost function으로 이용함 / 마지막으로, 매칭된 결과로 얻은 teacher mask A’’t와 student mask As 사이의 cross-entropy loss가 distillation loss가 됨

- 본 연구는 이러한 self-training을 통해 student slot-attention module의 slot generation을 향상시키는 것을 목적으로 함

**3.2. Autoregressive transformer decoder with sequence permutations**

- autoregressive transformer decoder는 spatial broadcasting을 적용한 MLP decoder보다 global consistency 측면에서 장점을 가짐

- patch 간 independence를 가정하는 MLP와 달리, autoregressive decoder는 patch 간의 joint distribution을 표현할 수 있음

- 하지만, 높은 model capacity로 인해 decoding이 진행될수록 slot vector 정보보다, 이전 patch token에 의존하는 경향이 있음 / 그림과 같이 token 위치에 따라 encoder에게 전해지는 reconstruction loss gradient에 큰 차이가 생겨, slot vector 학습에 문제가 생김

- 따라서 본 연구에서는 다양한 sequence permutation을 적용해 patch order에 variation을 줌

- 매 training step마다 permutation을 랜덤하게 선택해 token 순서를 변화시키기 때문에, slot에 대한 gradient가 token 위치에 영향을 적게 받게 됨

**4. Experiments**

- self-training, sequence-permutation의 유무를 비교한 실험 결과, 두 technique 모두 성능 향상에 기여하는 것을 알 수 있음

- 특히 self-training을 적용한 경우 sequence-permutation의 유무가 decoder 성능에 큰 차이를 보임 / self-training으로 인해 slot vector 학습은 극대화된 반면, decoder는 앞서 언급했듯 slot 정보는 무시하게 되는 경향을 가지기 때문에 이를 활용하지 못하기 때문에 이러한 차이가 나타남

- self-training을 적용했을 때, second stage에서 image encoder를 fine-tuning한 경우 추가적인 성능 향상을 보여줌 / 이 때 self-training loss가 reconstruction loss와 다른 방향으로 image encoder를 학습시켜, object-specific slot을 생성하는 데에 기여해 training stabilization 효과를 줌

- 다른 object-centric methods와 비교했을 때, 특히 real-world dataset에 대해 높은 성능을 보임

- 또한 그림과 같이 over-segmentation을 줄이면서도, detailed segmentation을 유지함