Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis

링크	https://arxiv.org/pdf/2111.14791.pdf
분야	Medical
세미나	False
연도	2022
자료	
코드	https://github.com/Project-MONAI/research-contributions/tree/main/SwinUNETR/BTCV
학회	Computer Vision and Pattern Recognition

이 논문은 3d medical data를 다루는 Swin UNETR에서 Transformer Encoder의 사전 학습 방법에 대해 다룬다.

ImageNet 따위로 사전학습된 Swin Transformer를 사용할 수도 있지만, 이 방법은 Medical Data를 처리하는 데 적합하지 않다고 한다. 왜냐하면 일반적인 이미지와 메디컬 이미지(CT,MRI 등)의 갭 차이가 크므로 메디컬 데이터에 맞게 사전학습을 해야된다는 것이다.

따라서 저자는 자기지도학습(Self-Supervised Learning)을 통한 사전 학습 방법에 대해 다룬다. 자기지도학습은 라벨링이 없는 데이터에 대해 정답을 임의로 만든 뒤 학습하는 방법이다.

Swin Transformer를 사전학습시키는 구조는 다음과 같다.

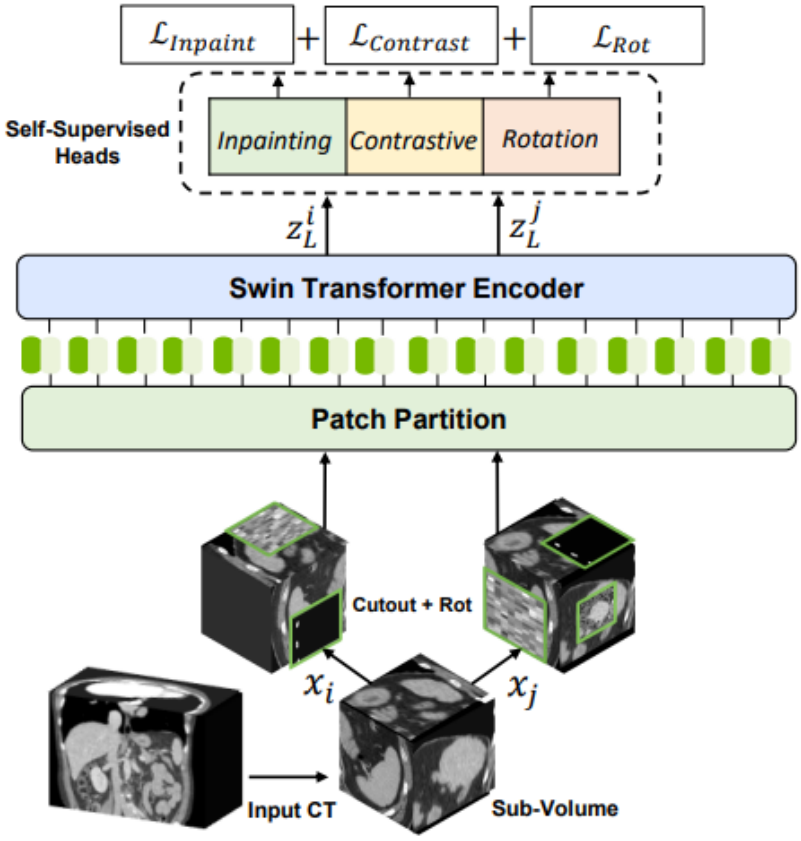


Figure 1. Overview of our proposed pre-training framework. Input CT images are randomly cropped into sub-volumes and augmented with random inner cutout and rotation, then fed to the Swin UNETR encoder as input. We use masked volume inpainting, contrastive learning and rotation prediction as proxy tasks for learning contextual representations of input images.

CT 영상을 랜덤하게 Crop을 해서 Sub-Volume을 여러 개 만든다.
그리고 각 Sub-Volume마다 cutout과 rotation을 두 번 거쳐서 xi와 xj를 제작한다.

xi와 xj는 Swin의 입력으로 사용된다.
이는 Patch Partition과 Encoder를 거친 뒤 Feature에 해당하는 ZLi와 ZLj로 출력된다.

두 output을 이용해서 **Inpainting, Contrastive Coding, Rotation**이라는 학습을 하게 된다.

- Inpainting은 cutout된 영역을 채우는 학습이며 정답은 원본 데이터가 된다.

- Contrastive Coding은 비교를 통한 학습이며, 동일한 Sub-volume으로부터 나온 pair는 유사도가 높고, 다른 Sub-volume으로부터 나온 pair는 유사도가 낮도록 학습시키는 것이다.

- Rotation은 회전된 입력에 대한 각도를 예측하는 것이다. 각도는 0도, 90도, 180도, 270도로 선정했다고 한다.

4.1. Masked Volume Inpainting

The cutout augmentation masks out ROIs in the sub-volume $\mathcal{X} \in \mathbb{R}^{H \times W \times D \times C}$ randomly with volume ratio of s. We attach a transpose convolution layer to the encoder as the reconstruction head and denote its output as $\hat{\mathcal{X}}^{\mathcal{M}}$. The reconstruction objective is defined by an $L1$ loss between $\mathcal{X}$ and $\hat{\mathcal{X}}^{\mathcal{M}}$

$$\mathcal{L}_{inpaint} = \|\mathcal{X} - \hat{\mathcal{X}}^{\mathcal{M}}\|_1, \tag{3}$$

The masked volume inpainting is motivated by prior work which focused on 2D images [43]. We extend it to 3D domain to showcase its effectiveness on representation learning of volumetric medical images.

먼저 Inpainting에 대한 설명이다.
Swin Encoder를 거치면 H,W,D가 2^4배로 작아지기 때문에 encoder에 transpose convolution layer를 부착했다고 한다.
이 layer를 거친 결과가 x^M이 되는데, 원본 X와의 L1 loss가 작아지는 방향으로 모델을 학습시키는 것이다.

4.2. Image Rotation

The rotation prediction task predicts the angle categories by which the input sub-volume is rotated. For simplicity, we employ R classes of $0°$, $90°$, $180°$, $270°$ rotations along the z-axis. An MLP classification head is used for predicting the softmax probabilities $\hat{y}_r$ of rotation categories. Given the ground truth y_r, a cross-entropy loss is used for rotation prediction task:

$$\mathcal{L}_{rot} = -\sum_{r=1}^{R} y_r \log(\hat{y}_r), \tag{4}$$

다음은 Rotation에 대한 설명이다.
Rotation이 0도인지, 90도인지, 180도인지, 270도인지를 예측하기 위해 Encoder이후 MLP Classifiaction head를 부착했다고 한다.
출력층에는 4개의 노드가 있을 것이고, 각각은 softmax가 취해진 확률 값이 저장될 것이다. 이 값들을 y^라고 할 때, y^과 정답인 y간의 Cross Entropy Loss가 작아지는 방향으로 모델을 학습시키는 것이다.

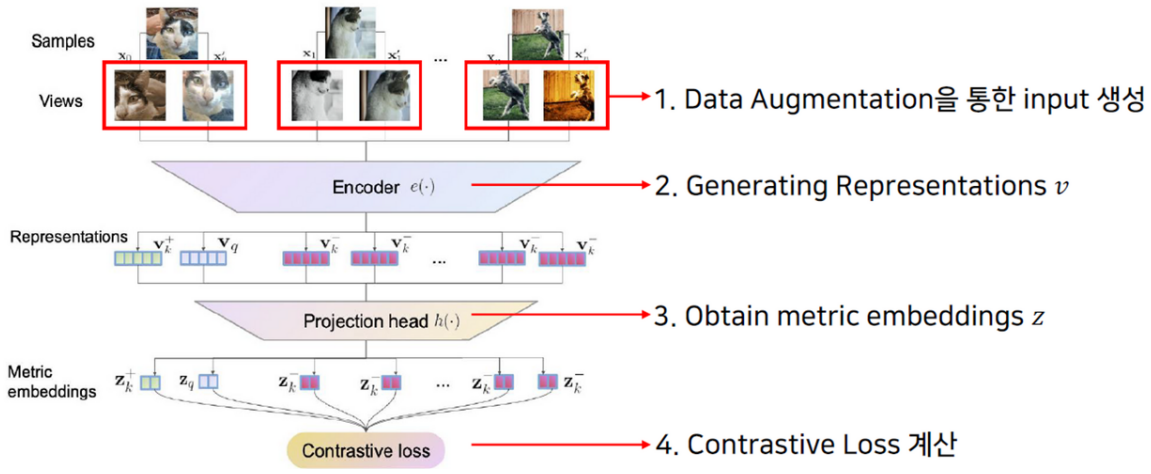
4.3. Contrastive Coding

The self-supervised contrastive coding presents promising performance on visual representation learning when transferred to downstream tasks [12, 42]. Given a batch of augmented sub-volumes, the contrastive coding allows for a better representation learning by maximizing the mutual information between positive pairs (augmented samples from same sub-volume), while minimizing that between negative pairs (views from different sub-volumes). The contrastive coding is obtained by attaching a linear layer to the Swin UNETR encoder, which maps each augmented sub-volume to a latent representation v. We use cosine similarity as the distance measurement of the encoded representations as defined in [12]. Formally, the 3D contrastive coding loss between a pair v_i and v_j is defined as:

$$\mathcal{L}_{contrast} = -\log \frac{\exp(sim(v_i, v_j)/t)}{\sum_k^{2N} 1_{k \neq i} \exp(sim(v_i, v_k)/t)}, \quad (5)$$

where t is the measurement of normalized temperature scale. 1 is the indicator function evaluating to 1 iff $k \neq i$. sim denotes the dot product between normalized embeddings. The contrastive learning loss function strengthens the intra-class compactness as well as the inter-class separability.

다음은 Contrastive Coding에 대한 설명이다.
latent representation(v)에 대한 pair를 입력받았을 때, 동일한 sub-volume으로부터 augmented된 v들(positive pair)이면 similarity가 높고, 동일하지 않으면(negative pair) similarity가 낮도록 학습시키는 것이 목표이다.
similarity는 cosine 유사도를 이용하는데, cosine 유사도는 두 벡터간의 유사도를 0과 1사이의 값으로 나타내는 방법이다. 따라서 v를 제작하기 위해 encoder에 한 개의 linear layer를 부착했다고 한다. ZLi와 ZLj는 linear layer를 거쳐 vi과 vj가 되는 것이다.
Loss는 Negative Log Likelihood에 대한 Softmax 함수를 사용한다. (확률이 1이면 loss는 0, 확률이 0이면 loss는 infinity)
softmax는 주로 딥러닝에서 출력층 노드 값을 표현하는 데 사용이 되지만, 여기서는 similarity를 확률로 나타냈다. 왜냐하면 단순히 vi와 vj간(positive pair)의 simiarity가 높은 것만이 중요한 게 아니라, negative pairs의 simarity도 낮도록 해야되기 때문이다.
예를 들어 분모의 시그마 항이 2개만 있다고 가정하고(positive pair와 negative pair 두 쌍만 있다고 가정하고), 두 쌍의 simiarity가 0.7, 0.5이면(negative pair는 학습이 잘 안 돼있는 상태) softmax는 0.7/1.2가 된다. 그러나 0.7, 0.1이면(negative pair는 학습이 잘 돼있는 상태) softmax는 0.7/0.8로 더 높아진다. 따라서 확률로 나타냄으로써 contrastive coding이 잘 되는가를 확인할 수 있다.



위 그림은 일반적인 Contrastive Learning에 대한 Architecture이다.
각 샘플마다 두 개씩 augmented돼서 미니배치가 모델의 입력으로 들어가게 된다.
따라서 latent vector(z)는 (batch size(N) x 2)개만큼 생성이 된다.
<https://daebaq27.tistory.com/97>

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

loss는 4-3에 적혀있는 수식을 봐도 되지만, 위 수식이 더 이해하기 쉽다. (완전히 같은 수식임)
softmax를 loss로 사용하기 위해 NLL(Negative Log Likelihood)을 적용했다고 앞에서 언급했다.
그런데 분모에 대해서 좀 더 설명하자면, k가 바뀌면서 sim(zi,zk)를 계산하는 것이고, (zi,zk)는 negative가 될 수도 있수도 있고, positive가 될 수도 있다.
positive가 되는 경우, zk는 zj랑 같고, negative가 되는 경우, zk는 미니배치에 있는 다른 augmented latent vector가 된다.
즉 분모에서 2N-1개의 simiarity를 계산하는데, 1개는 positive pair, 2(n-1)개는 negative pair라는 것이다.
이 내용은 아래 논문을 보면 더 자세히 이해가 가능하다.
<https://arxiv.org/pdf/2002.05709.pdf>
<we treat the other 2(N-1) augmented examples within a minibatch as negative examples.>

4.4. Loss Function

Formally, we minimize the total loss function by training Swin UNETR's encoder with multiple pre-training objectives of masked volume inpainting, 3D image rotation & contrastive coding as follows:

$$\mathcal{L}_{tot} = \lambda_1 \mathcal{L}_{inpaint} + \lambda_2 \mathcal{L}_{contrast} + \lambda_3 \mathcal{L}_{rot}. \quad (6)$$

A grid-search hyper-parameter optimization was performed which estimated the optimal values of $\lambda_1 = \lambda_2 = \lambda_3 = 1$.

앞에서 세 개의 자기지도학습 방법에 대해 소개했다.
loss를 각각 설명했지만 훈련을 할 때는 total loss를 이용한다고 한다.
하이퍼 파라미터인 lambda는 모두 1일 때가 적합했다고 한다.

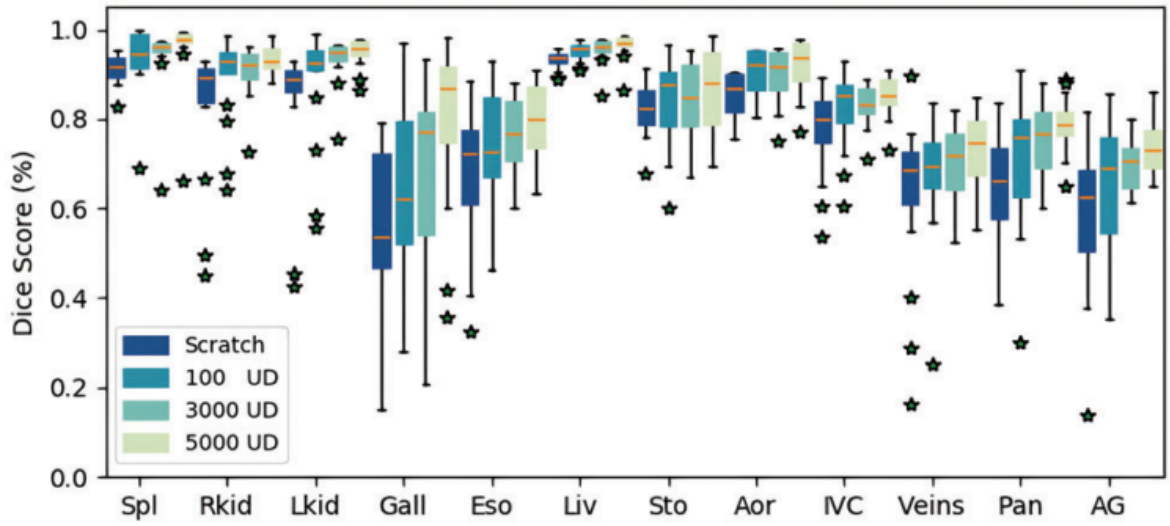


Figure 8. Pre-trained weights using 100, 3000 and 5000 scans are compared for fine-tuning on the BTCV dataset for each organ.

사전학습에 사용된 데이터가 많을수록 실제 Segmentation Task에서 Dice Score도 높게 나왔다. 자기지도학습을 많이 할수록 효과가 있다는 자료이다.