(NIPS 2020)
Contrastive learning of global and local features for medical image segmentation with limited annotations

Krishna Chaitanya Ertunc Erdil Neerav Karani Ender Konukoglu

Computer Vision Lab, ETH Zurich

Sternwartstrasse 7, Zurich 8092, Switzerland

Luoyao Kang
2021.8.18

Content

- Background
- Method
- Experiment and Results

背景和动机

有监督学习的成功依赖大量有标注的数据集，然而在医疗图像分析中，这种条件很难满足（很难得到大量高质量的标注图像）。自监督学习提供了一种利用无标签数据预训练网络的策略，随后利用少量有标签的数据针对下游任务进行微调。对比学习是自监督学习的一种变体，能够学习到图片级别的表示。

本文提出的方法，利用了domain-specific和problem-specific的线索，拓展了在少量标注数据情况下，针对3D医疗图像分割的对比学习框架。

对比学习（contrastive learning）

- 对比学习的一般范式：

  对任意数据x，对比学习的目标是学习一个编码器 f 使得

$$score(f(x), f(x^+)) >> score(f(x), f(x^-))$$

其中x^+是和x相似的正样本，x^-是和x不相似的负样本，score是一个度量函数来衡量样本间的相似度。

- 对比学习的损失函数：

$$L_N = -\mathbb{E}_X \left[\log \frac{\exp\left(f(x)^T f(x^+)\right)}{\exp(f(x)^T f(x^+)) + \sum_{j=1}^{N-1} \exp\left(f(x)^T f(x_j^-)\right)} \right]$$

对比学习范式存在的问题

1. 对比学习范式对于需要图像级别表示的下游任务（例如分类），此策略很有用。但涉及像素级预测（例如图像分割）的下游任务可能另外需要独特的局部表征来区分相邻区域。

2. 对比策略通常是基于数据增强中使用的转换来设计的，并且没有利用数据集中不同图像之间可能存在的相似性。

论文贡献

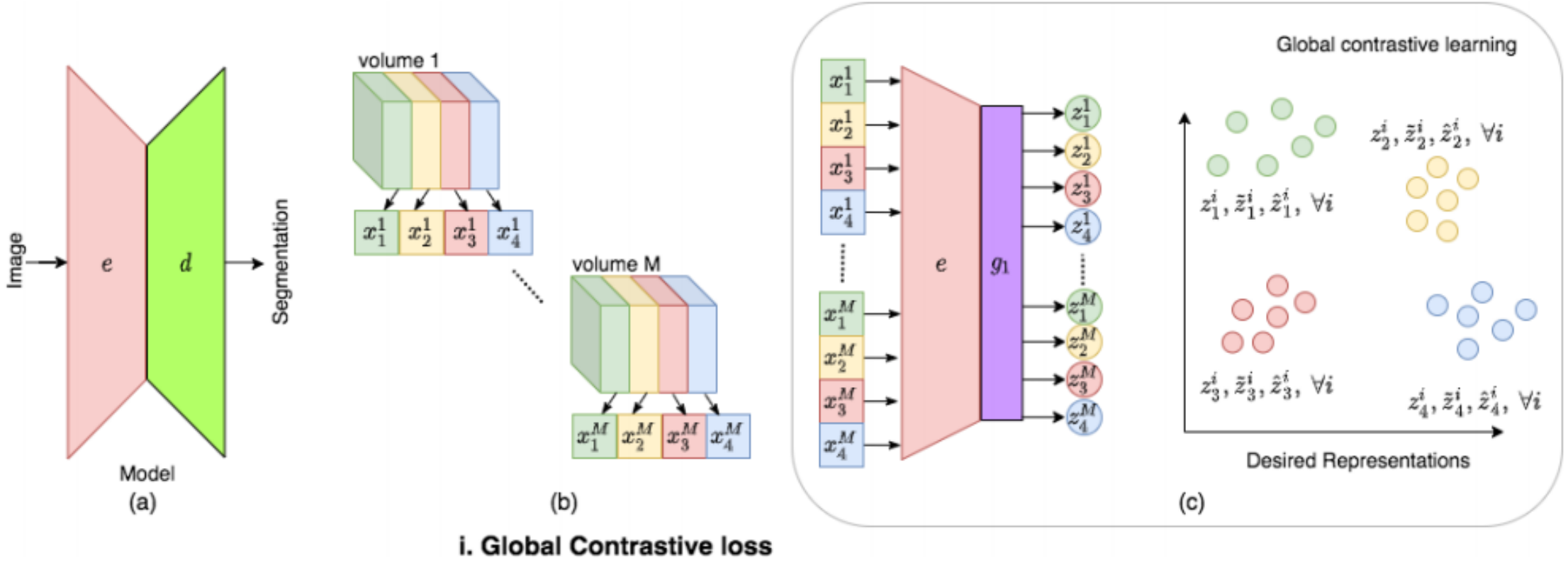
- 针对以上两个问题，本篇文章提出两个解决方法

  1. 提出了一种新的对比策略，利用了3D医疗图像中结构相似的特征
     (文中称之为domian-specific cue)。

  2. 提出了局部版本的对比损失函数，能够从局部区域学到特征表示，有助于像素级别的分割
     (文中称之为problem-specific cue)。

方法

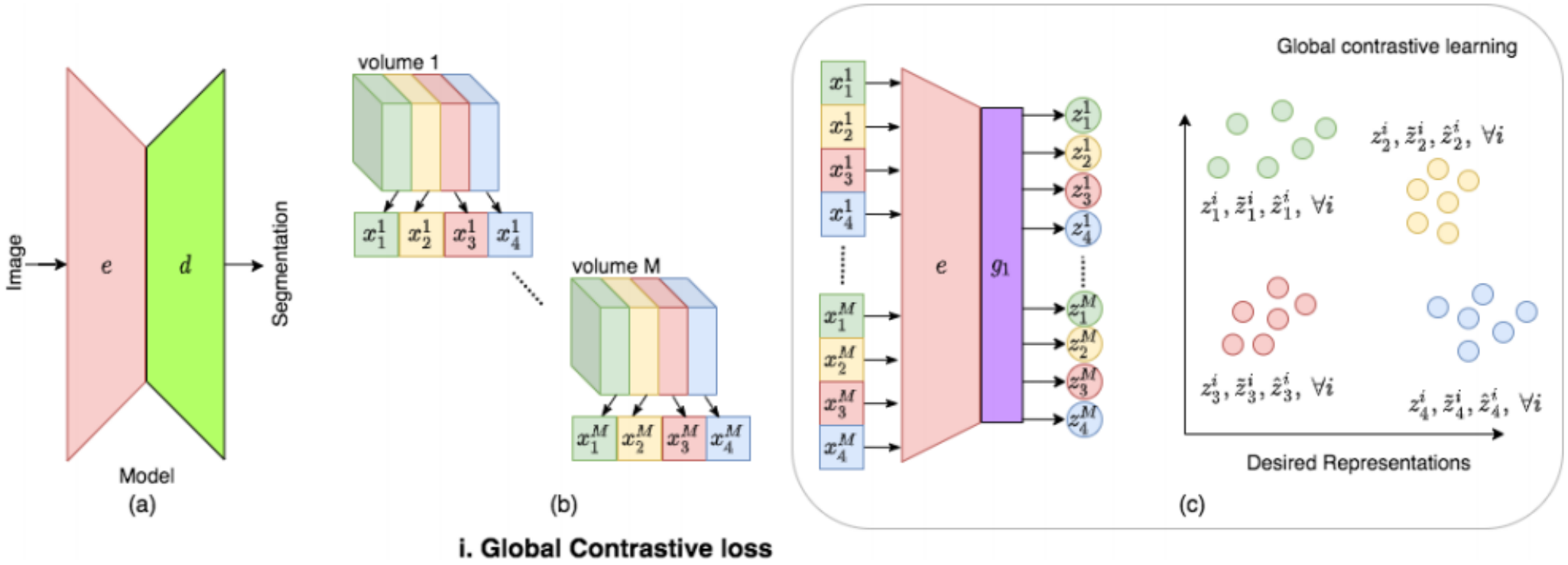
- Global Contrastive loss



i. Global Contrastive loss

Global Contrastive loss

$$l(\tilde{x}, \hat{x}) = -\log \frac{e^{\text{sim}(\tilde{z},\hat{z})/\tau}}{e^{\text{sim}(\tilde{z},\hat{z})/\tau} + \sum_{\bar{x}\in\Lambda^-} e^{\text{sim}(\tilde{z},g_1(e(\bar{x})))/\tau}}, \quad \tilde{z} = g_1(e(\tilde{x})), \quad \hat{z} = g_1(e(\hat{x})).$$



i. Global Contrastive loss

Encoder构建正负样本集的策略

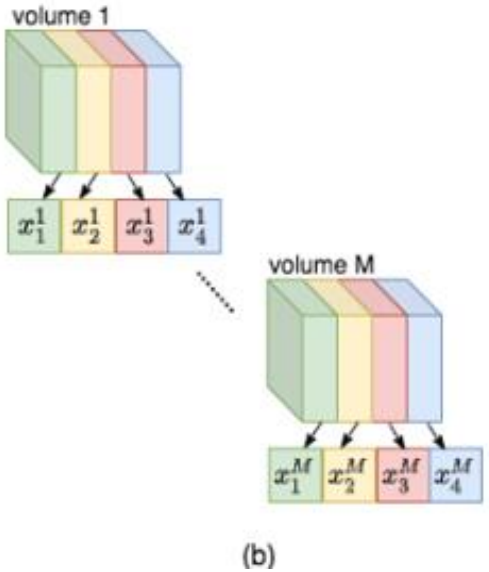
1. 按照对比学习范式(G^R):

在所有volumes上随机采样N个图像并对每个图像使用一对随机变换来得到相似对 $(\tilde{x_s^i}, \hat{x_s^i})$

负样本集 Λ^- 为所有剩余的2N-2个图像

2. 策略 G^{D-} ----没有认为不同的volume的相同位置相似

Λ^+ $(x_s^i, \hat{x_s^i})$ $(x_s^i, \tilde{x_s^i})$ $(\tilde{x_s^i}, \hat{x_s^i})$ Λ^- $\{x_k^l, \tilde{x_k^l}, \hat{x_k^l}\}, \forall k \neq s;$



3. 策略 G^D ----不同的volume的相同位置认为是相似的

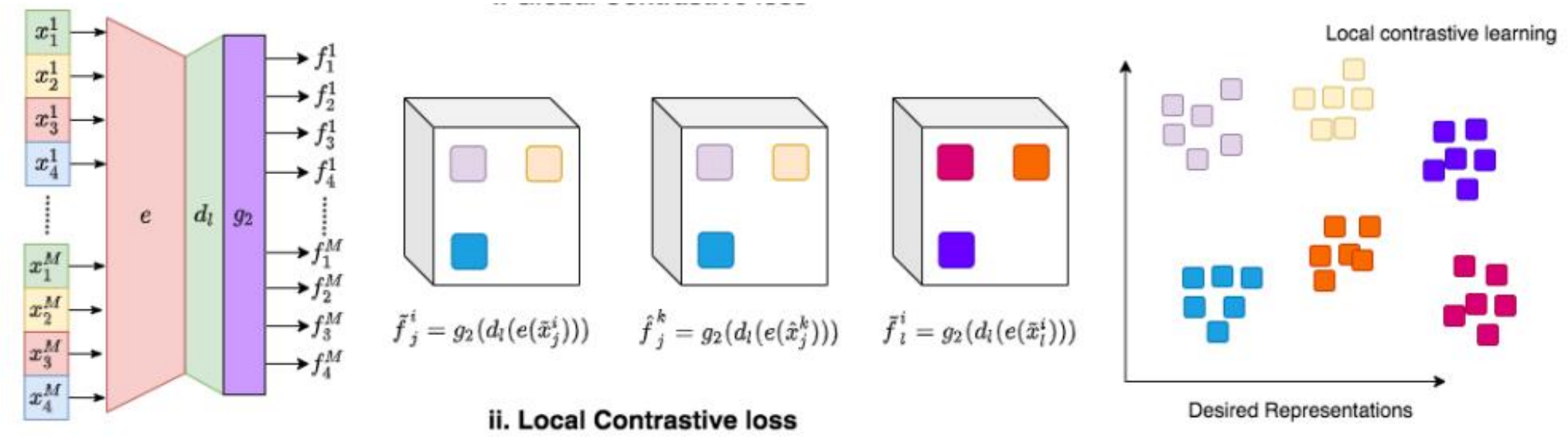
相似集合 Λ^+ 考虑了不同的volume的相同位置, 及其各种变换

Λ^+ (x_s^i, x_s^j) ..... $(x_s^i, \hat{x_s^i})$ $(x_s^i, \tilde{x_s^i})$ $(\tilde{x_s^i}, \hat{x_s^i})$ ...; Λ^- $\{x_k^l, \tilde{x_k^l}, \hat{x_k^l}\}, \forall k \neq s;$

方法

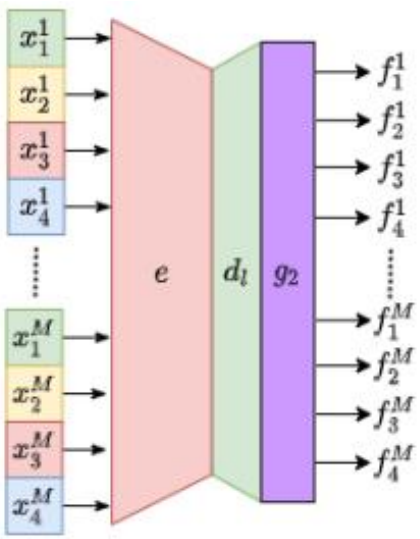
- Local Contrastive loss

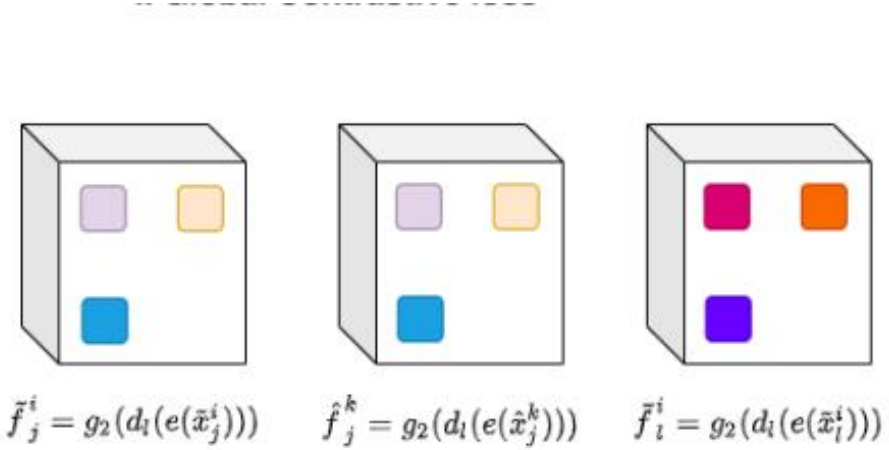
对于分割任务而言，需要像素级别的预测，好的局部特征表示可能对区分相邻区域有帮助。



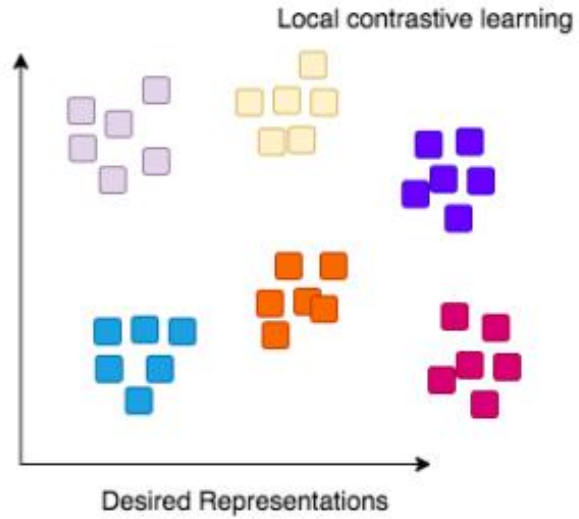
Local Contrastive Loss

$$l(\tilde{x}, \hat{x}, u, v) = -\log \frac{e^{\text{sim}(\tilde{f}(u,v), \hat{f}(u,v))/\tau}}{e^{\text{sim}(\tilde{f}(u,v), \hat{f}(u,v))/\tau} + \sum_{(u',v') \in \Omega^-} e^{\text{sim}(\tilde{f}(u,v), \hat{f}(u',v'))/\tau}},$$

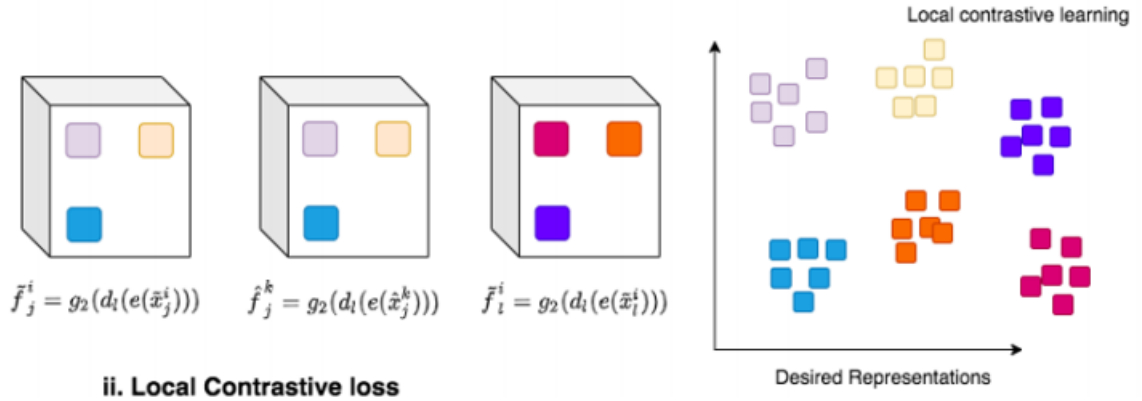




ii. Local Contrastive loss



Decoder的采样策略



ii. Local Contrastive loss

策略 L^R

先在所有volume中随机采样N个图像，每张图都进行随机变换来得到相似对 $(\tilde{x_s^i}, \hat{x_s^i})$ 然后解码得到$(\hat{f}, \tilde{f})$ 。

相似集合 Ω^+ 为对应位置的 $\hat{f}(\mu, \nu)$ 和 $\tilde{f}(\mu, \nu)$

负样本集 Ω^- 为不同位置的 $\hat{f}(\mu', \nu')$ 和 $\tilde{f}(\mu', \nu')$

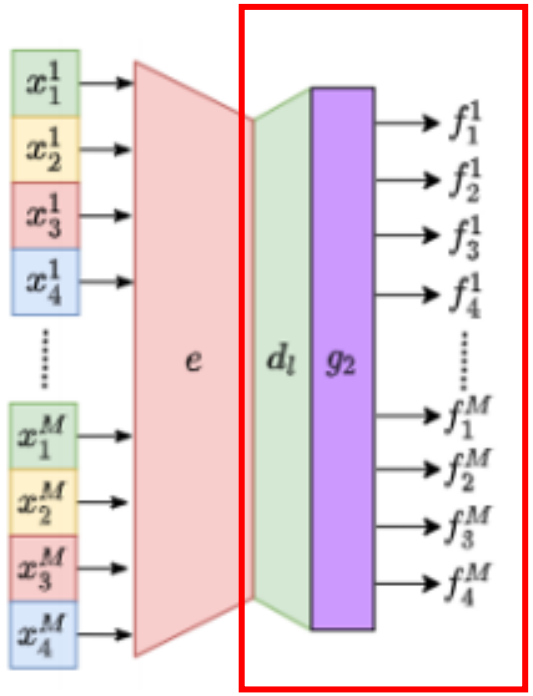
策略 L^D

不同volume的相同位置也有相似性。即 $\left(f_s^i(u, v), f_s^j(u, v)\right)$

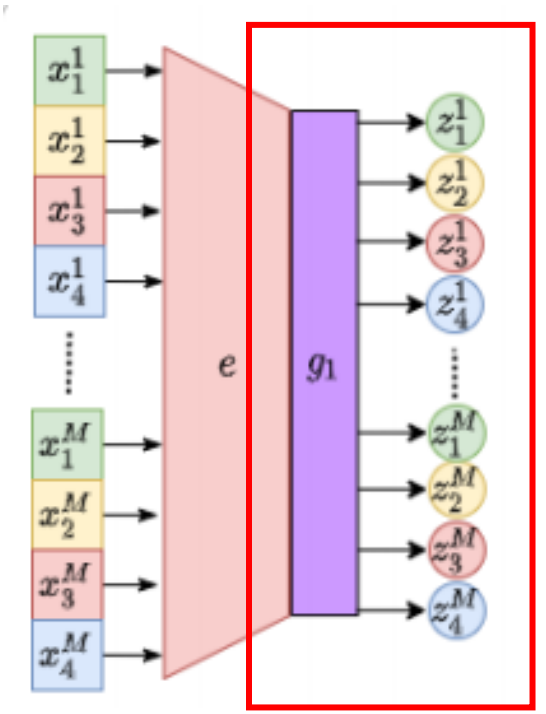
$$\Omega^+ \ (x_s^i, x_s^j, \tilde{x_s^i}, \hat{x_s^i}, \ \tilde{x_s^j}, \hat{x_s^j})$$

方法

- Local Contrastive loss。

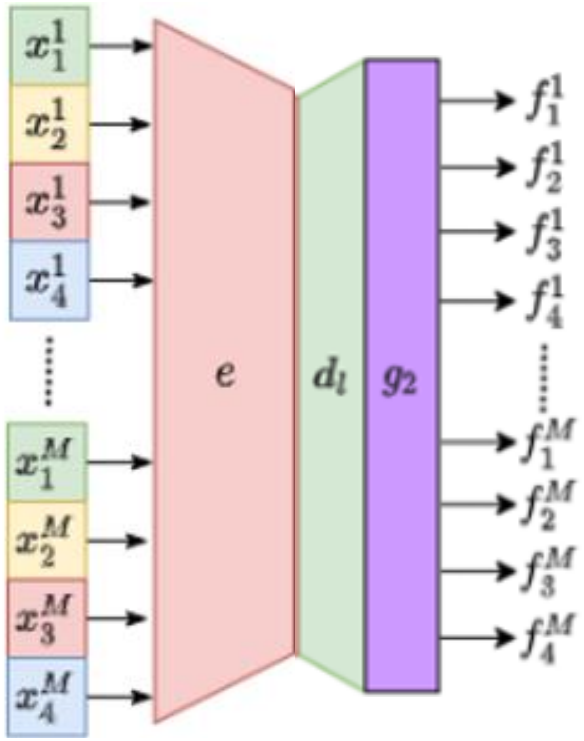


Local Model



Global Model

训练过程



$$l(\tilde{x}, \hat{x}, u, v) = -\log \frac{e^{\mathrm{sim}(\tilde{f}(u,v), \hat{f}(u,v))/\tau}}{e^{\mathrm{sim}(\tilde{f}(u,v), \hat{f}(u,v))/\tau} + \sum_{(u',v') \in \Omega^-} e^{\mathrm{sim}(\tilde{f}(u,v), \hat{f}(u',v'))/\tau}},$$

实验结果

Initialization of		ACDC			Prostate			MMWHS		
Encoder	Decoder	$\|X_{tr}\|$=1	$\|X_{tr}\|$=2	$\|X_{tr}\|$=8	$\|X_{tr}\|$=1	$\|X_{tr}\|$=2	$\|X_{tr}\|$=8	$\|X_{tr}\|$=1	$\|X_{tr}\|$=2	$\|X_{tr}\|$=8
random	random	0.614	0.702	0.844	0.489	0.550	0.636	0.451	0.637	0.787
Global contrasting strategies										
G^R	random	0.631	0.729	0.847	0.521	0.580	0.654	0.500	0.659	0.785
G^{D-}	random	0.683	0.774	0.864	0.553	0.616	0.681	0.529	0.684	0.796
G^D	random	0.691	0.784	0.870	0.579	0.600	0.677	0.553	0.686	0.793
Local contrasting strategies										
G^R	random	0.631	0.729	0.847	0.521	0.580	0.654	0.500	0.659	0.785
G^R	L^R	0.668	0.760	0.850	0.557	0.601	0.663	0.528	0.687	0.791
G^R	L^D	0.638	0.740	0.855	0.542	0.605	0.672	0.520	0.664	0.779
Proposed method										
G^D	L^R	**0.725**	**0.789**	**0.872**	**0.579**	**0.619**	**0.684**	**0.569**	**0.694**	**0.794**

实验结果

Method	ACDC			Prostate			MMWHS		
	$\|X_{tr}\|$=1	$\|X_{tr}\|$=2	$\|X_{tr}\|$=8	$\|X_{tr}\|$=1	$\|X_{tr}\|$=2	$\|X_{tr}\|$=8	$\|X_{tr}\|$=1	$\|X_{tr}\|$=2	$\|X_{tr}\|$=8
Baseline									
Random init.	0.614	0.702	0.844	0.489	0.550	0.636	0.451	0.637	0.787
Contrastive loss pre-training									
Global loss G^R [12]	0.631	0.729	0.847	0.521	0.580	0.654	0.500	0.659	0.785
Proposed init. (G^D + L^R)	0.725	<u>0.789</u>	<u>0.872</u>	0.579	<u>0.619</u>	<u>0.684</u>	<u>0.569</u>	<u>0.694</u>	0.794
Pretext task pre-training									
Rotation [21]	0.599	0.699	0.849	0.502	0.558	0.650	0.433	0.637	0.785
Inpainting [46]	0.612	0.697	0.837	0.490	0.551	0.647	0.441	0.653	0.770
Context Restoration [11]	0.625	0.714	0.851	0.552	0.570	0.651	0.482	0.654	0.783
Semi-supervised Methods									
Self-train [6]	0.690	0.749	0.860	0.551	0.598	0.680	0.563	0.691	<u>0.801</u>
Mixup [62]	0.695	0.785	0.863	0.543	0.593	0.661	0.561	0.690	0.796
Data Aug. [9]	<u>0.731</u>	0.786	0.865	<u>0.585</u>	0.597	0.667	0.529	0.661	0.785
Adversarial training [66]	0.536	0.654	0.791	0.487	0.544	0.586	0.482	0.655	0.779
Combination of Methods									
Data Aug. [9] + Mixup [62]	0.747	-	-	0.577	-	-	-	-	-
Proposed init. + Self-train [6]	0.745	0.802	0.881	**0.607**	**0.634**	**0.698**	**0.647**	**0.727**	**0.806**
Proposed init. + Mixup [62]	**0.757**	**0.826**	**0.886**	0.588	0.626	0.684	0.617	0.710	0.794
Benchmark									
Training with large $\|X_{tr}\|$	($\|X_{tr}\|$ = 78) 0.912			($\|X_{tr}\|$ = 20) 0.697			($\|X_{tr}\|$ = 8) 0.787		

总结

本文的工作在许多方面与现有的对比学习方法不同:

1. 先前的工作集中在用于图像方式预测任务的编码器类型架构上，而本篇专注于用于像素方式预测的编码器-解码器架构。
2. 提出了局部形式对比学习损失函数，也就是同一张图片经过不同变换以后，在局部区域的表征上相似，同一张图片不同位置的区域表征不相似。
3. 在学习全局表征时，将医学成像领域的知识整合在一起，以定义一组相似图像对而不是相同图像的不同变换。

(NIPS 2020)
Bootstrap Your Own Latent
A New Approach to Self-Supervised Learning

Jean-Bastien Grill∗,1 Florian Strub∗,1 Florent Altché∗,1 Corentin Tallec∗,1 Pierre H. Richemond∗,1,2

Elena Buchatskaya1 Carl Doersch1 Bernardo Avila Pires1 Zhaohan Daniel Guo1

Mohammad Gheshlaghi Azar1 Bilal Piot1 Koray Kavukcuoglu1 Rémi Munos1 Michal Valko1

1DeepMind

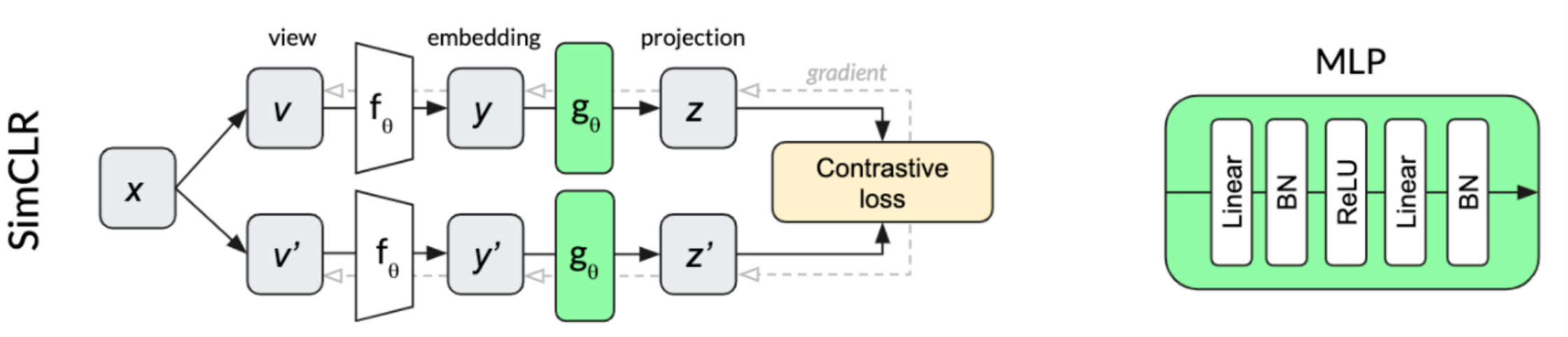
2Imperial College

Luoyao Kang
2021.8.18

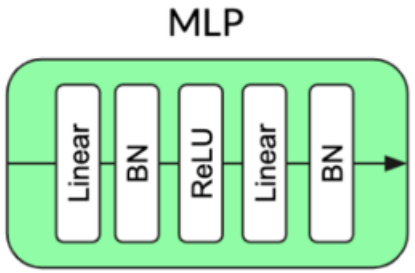
主要贡献和动机

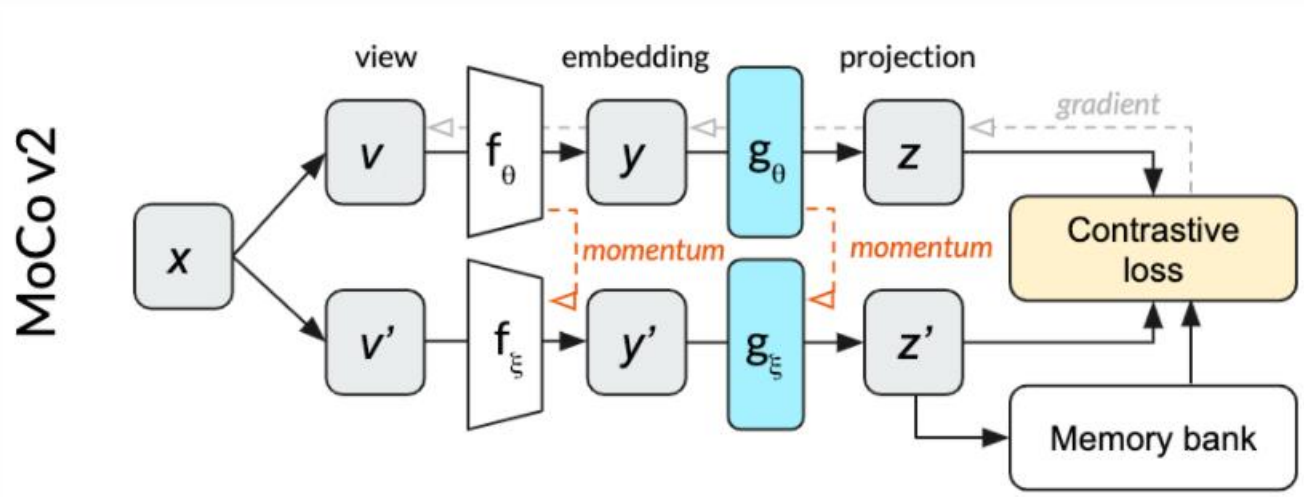
- 最近几年CV领域的大部分自监督表示学习的方法都是依赖于精心设计的pretext task。但在对比学习兴起之后使人们省去了设计pretext task的时间，通过利用Contrastive Predictive Coding这个通用的预训练任务，产生了不少优秀的文章，也提出了一些效果比较好的模型，最经典的当属SimCLR、MoCo。
- 但是他们都不可避免的使用正、负样本对来做对比，这就需要非常大的batch size，对存储要求比较高。这篇文章提出了一种不需要负样本即可超过SOTA的方法。

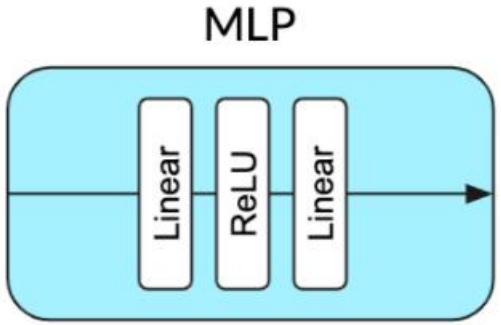
SimCLR


SimCLR
view
embedding
projection
gradient
x
v
v'
f_θ
f_θ
y
y'
g_θ
g_θ
z
z'
Contrastive loss
MLP
Linear
BN
ReLU
Linear
BN

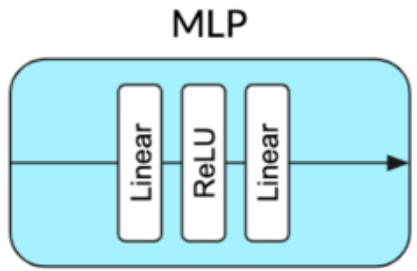
MoCo

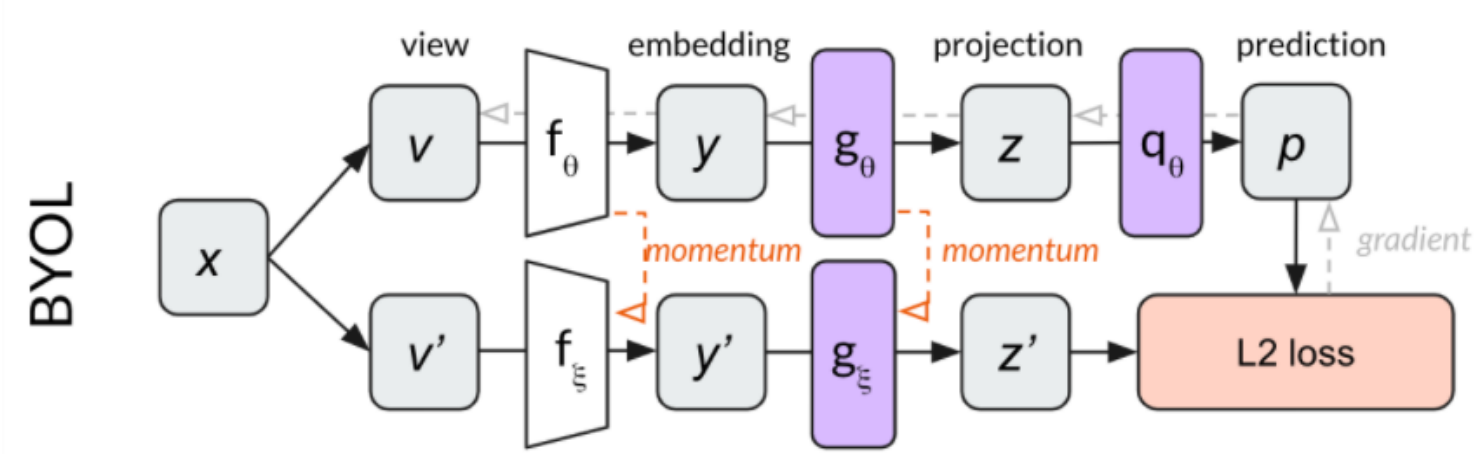

MLP
Linear
BN
ReLU
Linear
BN

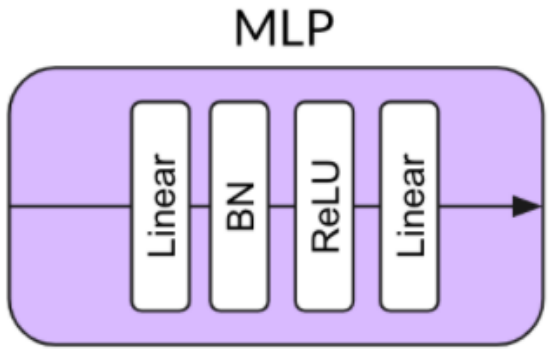

MoCo v2
view
embedding
projection
gradient
x
v
v'
f_θ
f_ξ
y
y'
g_θ
g_ξ
z
z'
momentum
momentum
Contrastive loss
Memory bank


MLP
Linear
ReLU
Linear

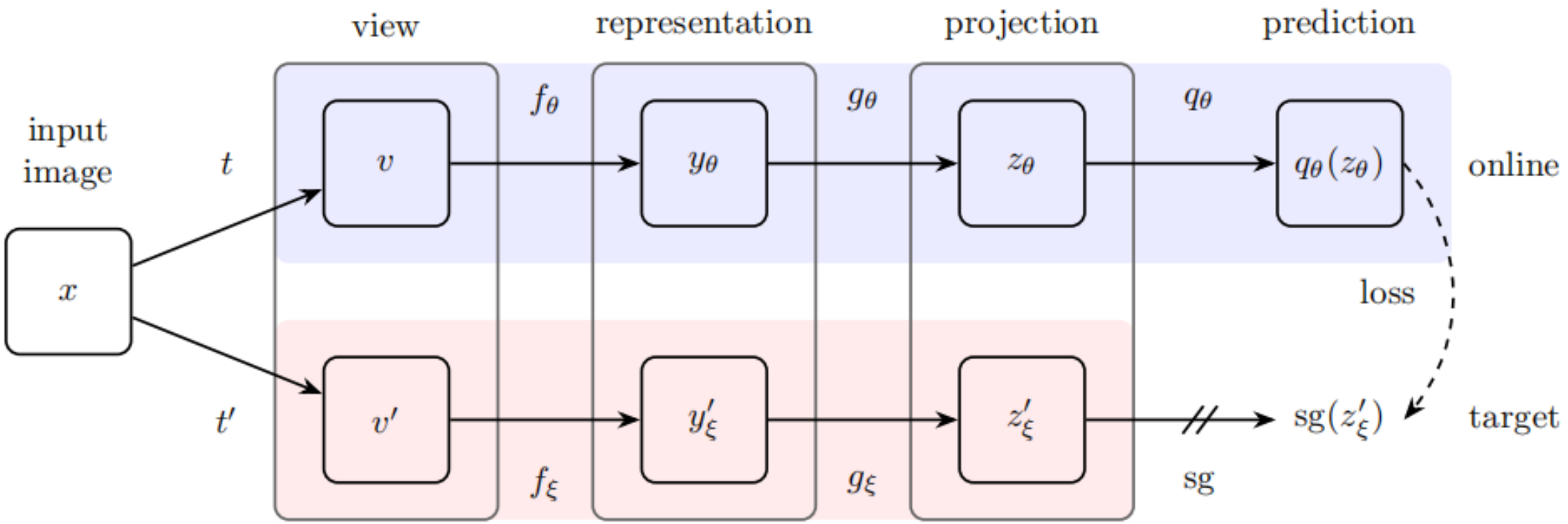
BYOL


MLP
Linear
ReLU
Linear

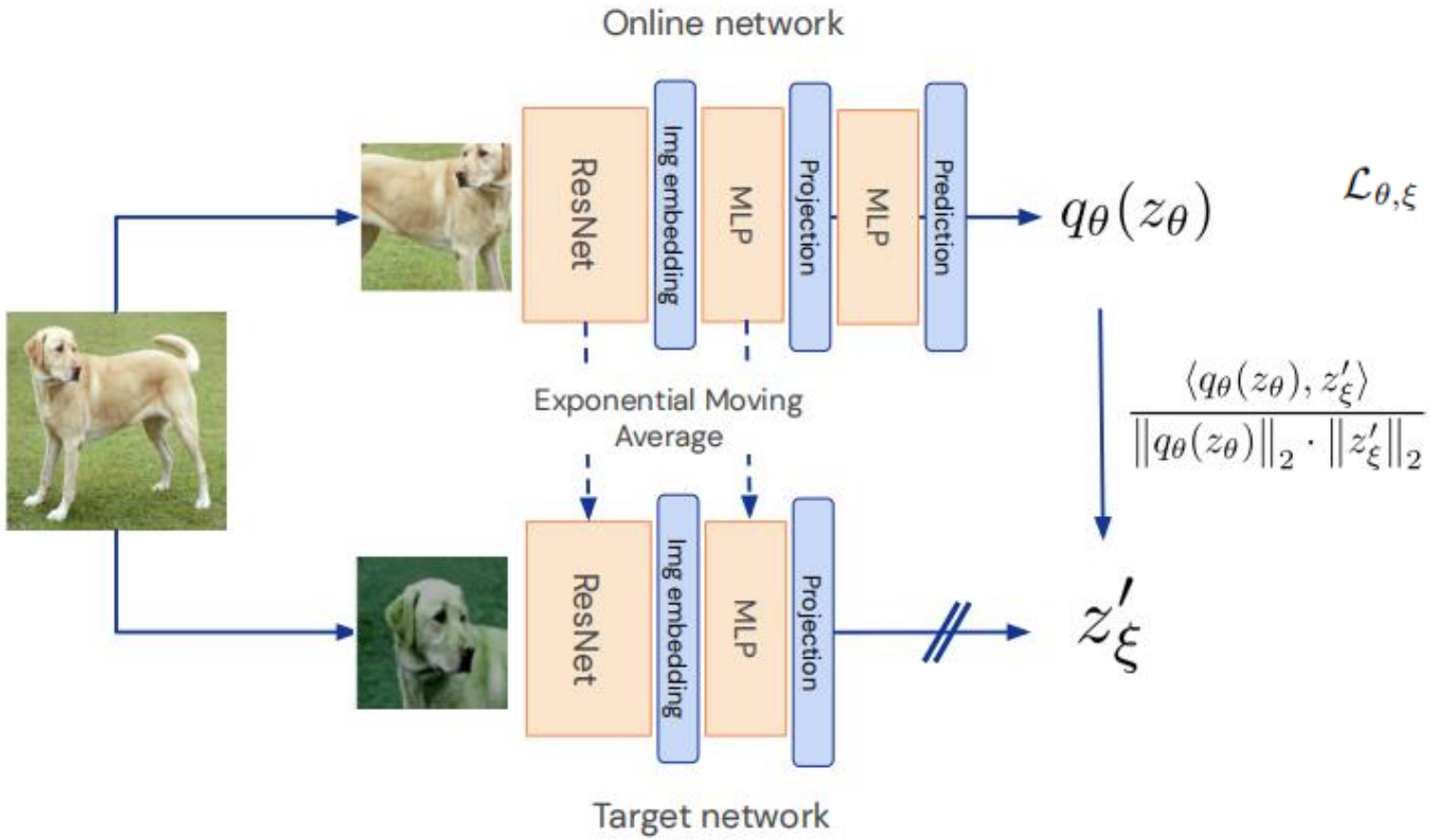

BYOL
view
embedding
projection
prediction
x
v
v'
f_θ
f_ξ
y
y'
g_θ
g_ξ
z
z'
q_θ
p
momentum
momentum
gradient
L2 loss


MLP
Linear
BN
ReLU
Linear

模型


view
representation
projection
prediction
input image
x
t
t'
v
v'
f_θ
f_ξ
y_θ
y'_ξ
g_θ
g_ξ
z_θ
z'_ξ
q_θ
$q_\theta(z_\theta)$
online
loss
$\mathrm{sg}(z'_\xi)$
target
sg

工作流程



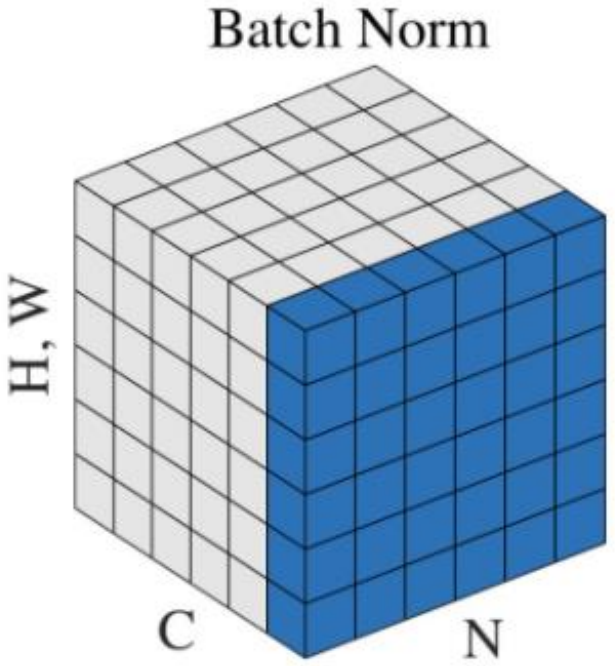
$$\frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 \cdot \|z'_\xi\|_2}$$

$$\mathcal{L}_{\theta,\xi} \triangleq \left\|\overline{q_\theta}(z_\theta) - \overline{z}'_\xi\right\|_2^2 = 2 - 2 \cdot \frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 \cdot \|z'_\xi\|_2}.$$

$$\theta \leftarrow \text{optimizer}\left(\theta, \nabla_\theta \mathcal{L}^{\text{BYOL}}_{\theta,\xi}, \eta\right),$$
$$\xi \leftarrow \tau\xi + (1-\tau)\theta,$$

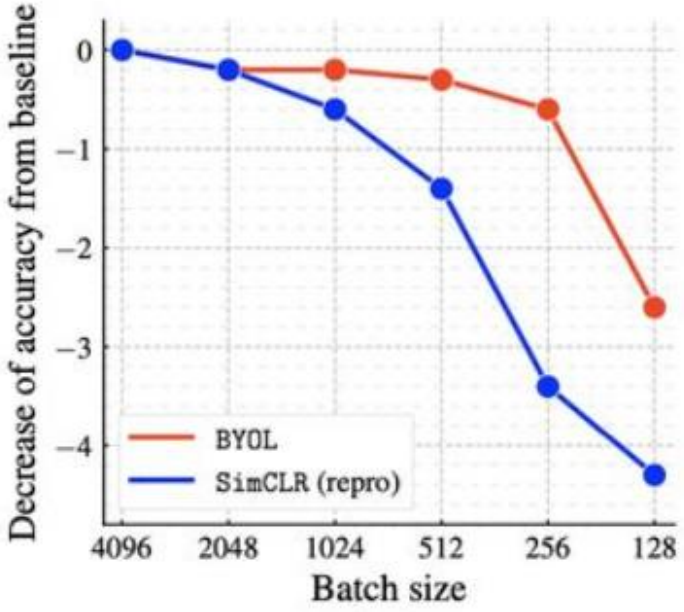
BYOL提出了利用两个Encoder（Target和Online）进行交互学习，分别输入同一图片不同变换后的版本，其中Target的权重随着Online滑动更新

BN?



结果

- Batch Size. SimCLR 中使用了很大的 batch size，由于本文不需要进行 negative 采样，因此性能对 batch size 的大小不是特别敏感。



(a) Impact of batch size

Thanks for listening