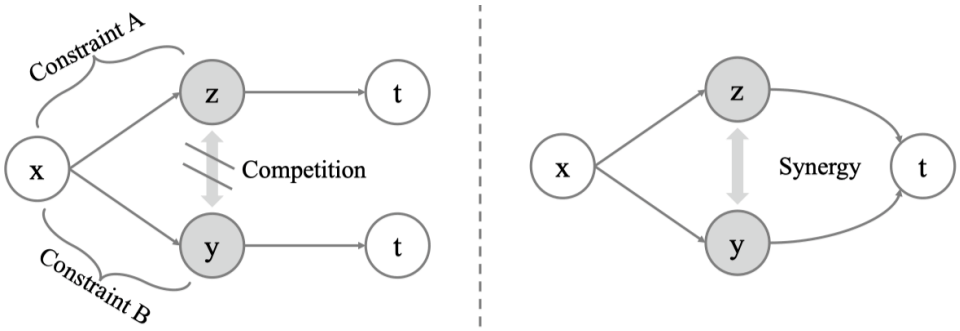
信息竞争式的多样化特征学习

1. 背景介绍

特征学习研究如何从输入提取出既有效又有意义的特征来完成下游任务。目前已经有大量的工作从信息约束的角度来学习更好的特征，但这些工作依然局限于对输入与特征之间加以单一约束。本文认为，要得到一种既有效（判别能力）又有意义（解耦能力）的特征，需要对特征的不同部分进行不同的信息约束。因此，本文提出了一种信息竞争式的多样化特征学习方法（Information Competing Process，ICP），在有监督的图像分类任务与自监督的图像解耦重构任务中验证了方法的有效性。

2. 目标函数推演

信息竞争式的多样化特征学习的主要思想如下图所示。在竞争阶段，输入为 x，对中间特征 z 和 y 加以不同的信息约束，并且令 z 与 y 在单独完成目标任务 t 的同时，约束 z 与 y 之间互相学习不到对方所携带的信息；在协同阶段，令 z 与 y 同时完成目标任务 t。



首先将有监督和自监督的目标任务统一表示为 t。在有监督的设定下，t 为输入 x 的标签；在自监督的设定下，t 为输入 x 自身。对于输入 x 的特征 r，原始的目标函数为：

$$\max\left[\mathcal{I}(r,t)\right], \tag{1}$$

其中，$\mathcal{I}(\cdot,\cdot)$ 表示的是变量之间的互信息。

为了增加特征的信息多样性，将特征 r 直接拆分成两部分 z 与 y，并对两部分加以完全不同方向的两个约束，可以得到如下目标函数：

$$\max\left[\mathcal{I}(r,t)+\alpha\mathcal{I}(y,x)-\beta\mathcal{I}(z,x)\right], \tag{2}$$

其中 $\alpha>0$，$\beta>0$。

让 z 与 y 互相直接进行竞争，一方面，约束 z 与 y 之间互相学习不到对方所携带的信息；另一方面，为了防止某部分的特征对下游任务占据主导，令 z 与 y 独立完成下游任务 t。因此，可以得到最终的目标函数：

$$\max\Big[\underbrace{\mathcal{I}(r,t)}_{\text{① Synergy}}+\underbrace{\alpha\mathcal{I}(y,x)}_{\text{② Maximization}}-\underbrace{\beta\mathcal{I}(z,x)}_{\text{③ Minimization}}+\underbrace{\mathcal{I}(z,t)+\mathcal{I}(y,t)-\gamma\mathcal{I}(z,y)}_{\text{④ Competition}}\Big], \tag{3}$$

其中，$\gamma>0$。

3. 目标函数优化

虽然目标函数中所有项都是在约束两个变量之间的互信息，但他们有不同的目标，所以需要不同的优化方法。我们将目标函数中不同的目标函数分为：互信息最小化项 $\mathcal{I}(z,x)$，互信息最大化项 $\mathcal{I}(y,x)$，目标任务推理项 $\mathcal{I}(z,t),\mathcal{I}(y,t),\mathcal{I}(r,t)$ 与预测最小化项 $\mathcal{I}(z,y)$，并给出了一种优化上述目标函数的例子。

3.1 互信息最小化项

对于互信息最小化项 $\mathcal{I}(z,x)$，我们需要找到一个将其表示为基于 x 的条件概率的可求上界进行优化，这个互信息项可以展开为：

$$\begin{aligned}\mathcal{I}(z,x) &= \int\int P(z,x)\log\frac{P(z,x)}{P(z)P(x)}dxdz = \int\int P(z,x)\log\frac{P(z|x)}{P(z)}dxdz \\ &= \int\int P(z,x)\log P(z|x)dxdz - \int\int P(x|z)P(z)\log P(z)dxdz \\ &= \int\int P(z,x)\log P(z|x)dxdz - \int P(z)\log P(z)dz.\end{aligned} \tag{4}$$

由于 $P(z)$ 不可求，我们利用先验 $Q(z)$ 来作为 $P(z)$ 的变分近似，所以有：

$$KL\big[P(z)||Q(z)\big] \geq 0 \Rightarrow \int P(z)\log P(z)dz \geq \int P(z)\log Q(z)dz. \tag{5}$$

结合式（4）和式（5），得到可求上界：

$$\mathcal{I}(z,x) \leq \int\int P(z|x)P(x)\log\frac{P(z|x)}{Q(x)}dxdz = \mathbb{E}_{x\sim P(x)}\Big[KL\big[P(z|x)||Q(z)\big]\Big], \tag{6}$$

3.2 互信息最大化项

对于互信息最大化项 $\mathcal{I}(y,x)$，寻找一个将其表示为基于 x 的条件概率的可求下界比较困难，我们转而寻找一个可求的替代项来优化这一项。对这一项进行展开，可以得到：

$$\mathcal{I}(y,x) = \int\int P(y,x)\log\frac{P(y,x)}{P(y)P(x)}dxdy = KL\big[P(y|x)P(x)||P(y)P(x)\big]. \tag{7}$$

式（7）认为最大化互信息 $\mathcal{I}(y,x)$ 等价于最大化 $P(y|x)P(x)$ 与 $P(y)P(x)$ 之间的 KL 散度。由于最大化 KL 散度不收敛，我们利用 JS 散度来替代，可以得到一个 JS 散度的可求变分估计：

$$\begin{aligned}JS\big[P(y|x)P(x)||P(y)P(x)\big] = \max\Big[&\mathbb{E}_{(y,x)\sim P(y|x)P(x)}\big[\log D(y,x)\big] \\ &+\mathbb{E}_{(\hat{y},x)\sim P(y)P(x)}\big[\log\big(1-D(\hat{y},x)\big)\big]\Big],\end{aligned} \tag{8}$$

其中 D 用于判别输入为正负样本的概率，(y,x) 为从 $P(y|x)P(x)$ 采样的正样本对，$(\hat{y},x)$ 为从 $P(y)P(x)$ 中采样的负样本对。由于 $P(y)$ 不可求，并且 $\hat{y}$ 应为基于输入的特征，我们通过随机重排 y 来得到负样本对。

3.3 目标任务推理项

对于目标任务推理项 $\mathcal{I}(z,t),\mathcal{I}(y,t),\mathcal{I}(r,t)$，我们以 $\mathcal{I}(r,t)$ 为例进行推导。目标是需要找到一个基于 r 的条件概率的可求下界进行优化，因此将其展开为：

$$\begin{aligned}\mathcal{I}(r,t) &= \int\int P(r,t)\log\frac{P(t|r)}{P(t)}drdt \\ &= \int\int P(r,t)\log P(t|r)dtdr - \int P(t)\log P(t)dt \\ &= \int\int P(r,t)\log P(t|r)dtdr + \mathcal{H}(t) \\ &\geq \int\int P(t|r)p(r)\log P(t|r)dtdr,\end{aligned} \tag{9}$$

其中 $\mathcal{H}(t)\geq 0$ 为信息熵。令先验 $Q(t|r)$ 为 $P(t|r)$ 的变分估计，可得：

$$KL\big[P(t|r)||Q(t|r)\big]\geq 0 \Rightarrow \int P(t|r)\log P(t|r)dt \geq \int P(t|r)\log Q(t|r)dt. \tag{10}$$

因此我们可以得到一个可求的下界：

$$\mathcal{I}(r,t)\geq \int\int P(r,t)\log Q(t|r)dtdr. \tag{11}$$

有监督设定下，t 为已知标签。假设特征 r 不取决于 t，即 $P(r|x,t)=P(r|x)$，可以得到：

$$P(x,r,t)=P(r|x,t)P(t|x)P(x)=P(r|x)P(t|x)P(x). \tag{12}$$

所以 r 和 t 的联合分布可以写为：

$$P(r,t)=\int P(x,r,t)dx=\int P(r|x)P(t|x)P(x)dx. \tag{13}$$

结合式（11）与（13），可以得到有监督设定下的下界为：

$$\begin{aligned}\mathcal{I}(r,t) &\geq \int\int\int P(x)P(r|x)P(t|x)\log Q(t|r)dtdrdx \\ &= \mathbb{E}_{x\sim P(x)}\Big[\mathbb{E}_{r\sim P(r|x)}\big[\int P(t|x)\log Q(t|r)dt\big]\Big].\end{aligned} \tag{14}$$

由于条件概率 $P(t|x)$ 在有监督设定下代表了已知标签的分布，因此式（14）实际上是一个分类的交叉熵损失函数。

自监督设定下，t 为输入 x 本身，因此式（11）可以直接写为：

$$\mathcal{I}(r,x)\geq \int\int P(r|x)P(x)\log Q(x|r)dxdt=\mathbb{E}_{x\sim P(x)}\Big[\mathbb{E}_{r\sim P(r|x)}\big[\log Q(x|r)\big]\Big]. \tag{15}$$

假设先验 $Q(t|r)$ 为标准高斯分布的情况下，式（15）实际上为输入 x 的 L2 重构损失。

3.4 预测最小化项

预测最小化项 $\mathcal{I}(z,y)$ 相当于令变量 z 与 y 互相独立，由（Predictability Minimization，PM）得到启发，我们可以引入一个预测器 H 来完成这个目标。具体而言，我们将 z 输入到 H 中预测 y，并利用其损失函数防止特征提取器提取能够预测 y 的 z。同理，对 y 也进行同样的操作。可以得到如下目标函数：

$$\min\max\Big[\mathbb{E}_{z\sim P(z|x)}\big[H(y|z)\big]+\mathbb{E}_{y\sim P(y|x)}\big[H(z|y)\big]\Big]. \tag{16}$$

至此，目标函数中所有项都得到了可求的替代。信息竞争式的多样化特征学习的优化过程可以总结为如下算法：

Algorithm 1: Optimization of Information Competing Process

Input: The source input x with the downstream task target t, the prior distribution $Q(z)$, $Q(t|r)$, $Q(t|z)$ and $Q(t|y)$ for variational approximation, and the hyperparameters α, β, γ.

Output: The learned representation extractor and downstream solver.

while *not Convergence* **do**
  Optimize Eq. 8 and Eq. 16 for discriminator D and predictor H;
  // Mutual Information Minimization Term:
  Replace $\mathcal{I}(z, x)$ in Eq. 3 with the tractable upper bound in Eq. 6;
  // Mutual Information Maximization Term:
  Replace $\mathcal{I}(y, x)$ in Eq. 3 with the tractable alternative in Eq. 8;
  // Inference Term:
  Replace $\mathcal{I}(z, t), \mathcal{I}(y, t), \mathcal{I}(r, t)$ in Eq. 3 with the tractable lower bound in Eq. 14;
  // Predictability Minimization Term:
  Replace $\mathcal{I}(z, y)$ in Eq. 3 with Eq. 16;
  Optimize Eq. 3 while fixing the parameters of D and H;
end

4. 实验

在实验中，所有条件概率都使用神经网络来拟合，我们假设 $Q(z), Q(t|r), Q(t|z), Q(t|y)$ 都为标准高斯分布。利用 VAE 中的重参数技巧，目标函数中每一项都可导。在有监督的图像分类任务中，我们使用一个单层的全连接层作为分类器；在自监督的图像重构解耦任务中，我们使用多层反卷积层来重构图像。

4.1 有监督的分类任务

4.1.1 分类结果

我们在 CIFAR-10 与 CIFAR-100 两个数据集上，利用 VGG16，GoogLeNet，ResNet20，DenseNet40 四种不同的网络结构，验证 ICP 的有效性。分类结果如下表所示：

Table 1: Classification error rates (%) on CIFAR-10 test set.

	VGG16 [29]	GoogLeNet [30]	ResNet20 [10]	DenseNet40 [13]
Baseline	6.67	4.92	7.63	5.83
VIB [1]	$6.81^{\uparrow 0.14}$	$5.09^{\uparrow 0.17}$	$6.95^{\downarrow 0.68}$	$5.72^{\downarrow 0.11}$
DIM* [12]	$6.54^{\downarrow 0.13}$	$4.65^{\downarrow 0.27}$	$7.61^{\downarrow 0.02}$	$6.15^{\uparrow 0.32}$
VIB$_{\times 2}$	$6.86^{\uparrow 0.19}$	$4.88^{\downarrow 0.04}$	$6.85^{\downarrow 0.78}$	$6.36^{\uparrow 0.53}$
DIM*$_{\times 2}$	$7.24^{\uparrow 0.57}$	$4.95^{\uparrow 0.03}$	$7.46^{\downarrow 0.17}$	$5.60^{\downarrow 0.23}$
ICP$_{\text{-ALL}}$	$6.97^{\uparrow 0.30}$	$4.76^{\downarrow 0.16}$	$6.47^{\downarrow 1.16}$	$6.13^{\uparrow 0.30}$
ICP$_{\text{-COM}}$	$6.59^{\downarrow 0.08}$	$4.67^{\downarrow 0.25}$	$7.33^{\downarrow 0.30}$	$5.63^{\downarrow 0.20}$
ICP	$\mathbf{6.10}^{\downarrow 0.57}$	$\mathbf{4.26}^{\downarrow 0.66}$	$\mathbf{6.01}^{\downarrow 1.62}$	$\mathbf{4.99}^{\downarrow 0.84}$

Table 2: Classification error rates (%) on CIFAR-100 test set.

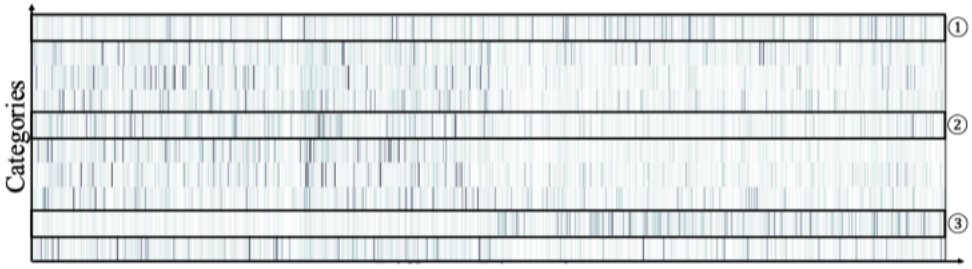
	VGG16 [29]	GoogLeNet [30]	ResNet20 [10]	DenseNet40 [13]
Baseline	26.41	20.68	31.91	27.55
VIB [1]	$26.56^{\uparrow 0.15}$	$20.93^{\uparrow 0.25}$	$30.84^{\downarrow 1.07}$	$26.37^{\downarrow 1.18}$
DIM* [12]	$26.74^{\uparrow 0.33}$	$20.94^{\uparrow 0.26}$	$32.62^{\uparrow 0.71}$	$27.51^{\downarrow 0.04}$
VIB$_{\times 2}$	$26.08^{\downarrow 0.33}$	$22.09^{\uparrow 1.41}$	$29.74^{\downarrow 2.17}$	$29.33^{\uparrow 1.78}$
DIM*$_{\times 2}$	$25.72^{\downarrow 0.69}$	$21.74^{\uparrow 1.06}$	$30.16^{\downarrow 1.75}$	$27.15^{\downarrow 0.40}$
ICP$_{\text{-ALL}}$	$26.73^{\uparrow 0.32}$	$20.90^{\uparrow 0.22}$	$28.35^{\downarrow 3.56}$	$27.51^{\downarrow 0.04}$
ICP$_{\text{-COM}}$	$26.37^{\downarrow 0.04}$	$20.81^{\uparrow 0.13}$	$32.76^{\uparrow 0.85}$	$26.85^{\downarrow 0.70}$
ICP	$\mathbf{24.54}^{\downarrow 1.87}$	$\mathbf{18.55}^{\downarrow 2.13}$	$\mathbf{28.13}^{\downarrow 3.78}$	$\mathbf{24.52}^{\downarrow 3.03}$

其中，Variational Information Bottleneck（VIB）只优化 $\mathcal{I}(z,t)-\beta\mathcal{I}(z,x)$；带有一个附加项的全局 Deep InfoMax（DIM*）只优化 $\mathcal{I}(y,t)+\alpha\mathcal{I}(y,x)$。角标 x2 的表示将上述方法特征扩展到和 ICP 相同的维度。消融实验中，ICP-ALL 表示去掉所有的约束条件的结果，即只优化式（1）；ICP-COM 表示去掉信息竞争的结果，即只优化式（2）。

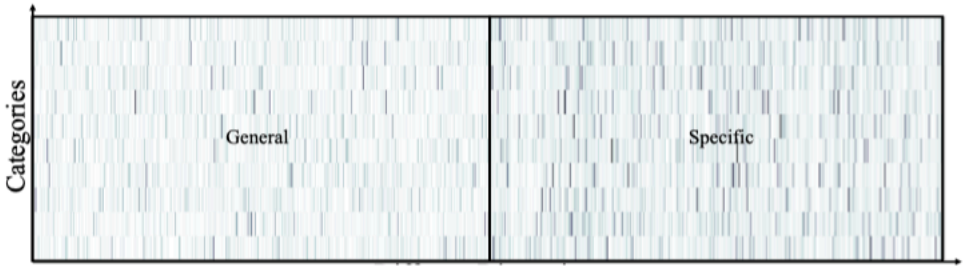
实验结果表明，由于缺少信息约束的多样性，VIB 与 DIM*只能达到一个次优的效果；ICP-ALL 由于大的模型容量与缺少约束，ICP-COM 由于单种特征占据主导，也都没有较好的效果。只有 ICP 在不同的结构中都取得了较好的效果。

4.1.2 多样化特征的可解释性初探

我们以在 CIFAR-10 上训练的 VGG 网络为例，初步探究了多样性特征的可解释性。下图展示了 VGG 网络正规化之后的分类器权重的绝对值：



(a) Correlation heatmap of ICP-ALL.



(b) Correlation heatmap of ICP.

其中，横轴为特征的每一个维度，纵轴为分类类别，颜色代表着相关程度。由上图（a）我们可以看到，ICP-ALL 的分类依赖于不同的部分，比如方框①均匀地取决于所有部分，方框②大部分取决于左边部分的特征，方框③大部分取决于右边部分的特征。相反的，上图（b）中可以明确的看到特征的取决关系分为两部分，分类基本上是由左边部分的少数维度搭配上右边部分的大部分维度完成。左半部分是最小化互信息得到的特征，右边部分是最大化互信息得到的特征。一般而言，大家认为最小化互信息得到的特征会携带着输入的泛化性的信息，因此只需要个别维度就可以完成分类；而最大化互信息得到的特征会携带着输入的特定性的信息，因此需要大量维度才能进行分类。

4.2 自监督的重构任务

4.2.1 数据集

我们在 2D 形状数据集 dSprites 和 3D 模拟人脸数据集 3D Faces 上定量和定性地测试了 ICP 的解耦重构能力。在更复杂的 CelebA 数据集上，我们重构了 128 x 128 大小的图片来定量测试 ICP 的重构能力和定性测试 ICP 的解耦能力。

4.2.2 定量实验结果

我们利用 Mutual Information Gap（MIG）这个指标，在 dSprites 和 3D Faces 两个数据集上定量测试了 ICP 的解耦效果。如下表所示：

Table 3: MIG score of disentanglement.

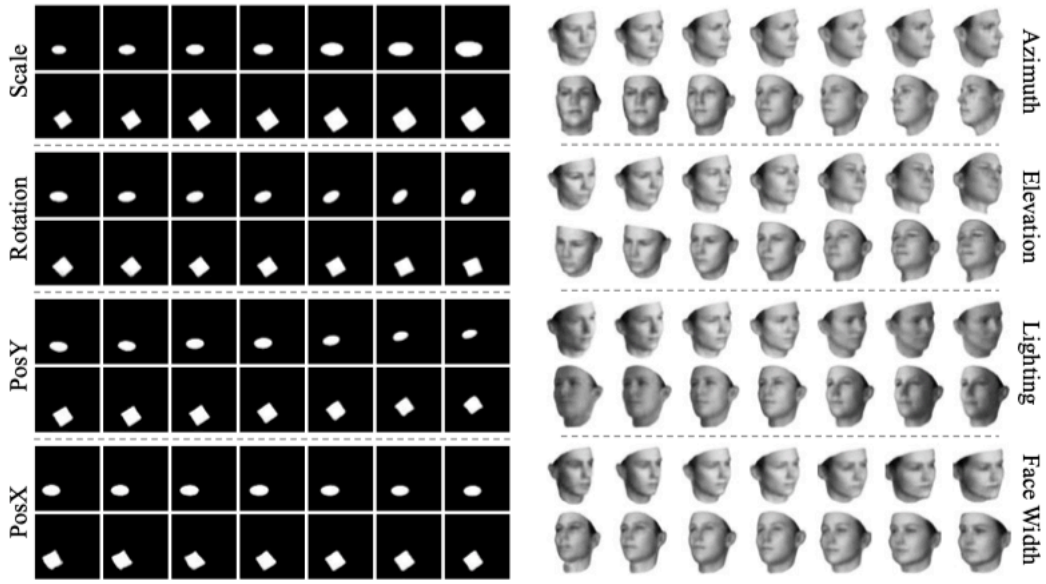
	dSprites [20]	3D Faces [23]
β-VAE [11]	0.22	0.54
β-TCVAE [6]	0.38	0.62
ICP-ALL	0.33	0.26
ICP-COM	0.20	0.57
ICP	**0.48**	**0.73**

其中，Beta-VAE 是我们的基准线，Beta-TCVAE 是目前 MIG 值最高的方法，ICP-ALL 和 ICP-COM 是消融实验的方法。可以看到 ICP 在 MIG 指标上超越了目前最好的方法。

此外，针对更加复杂的 CelebA 数据集，我们计算了 Mean Square Error（MSE）和 Structural Similarity Index（SSIM）来验证重构效果。ICP 的 MSE 为 0.0085，而基准线 Beta-VAE 的 MSE 为 0.0092；ICP 的 SSIM 为 0.62，而基准线 Beta-VAE 的 SSIM 为 0.60。结果证明，ICP 在重构上也优于基准线。

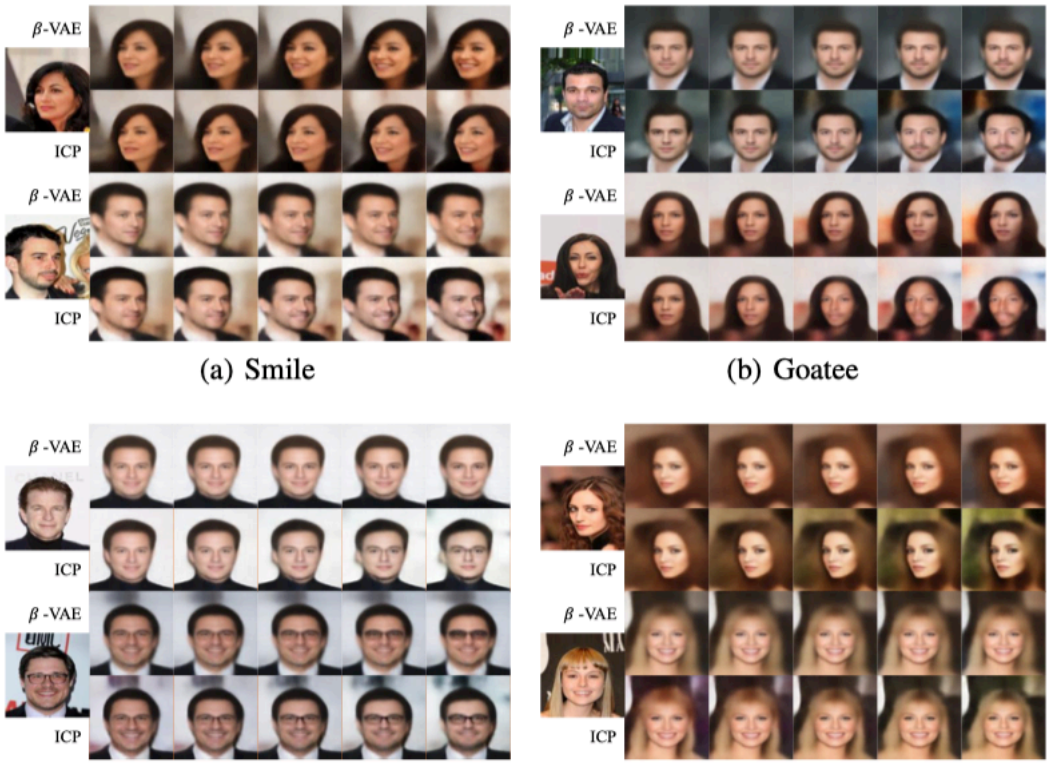
4.2.3 定性实验结果

对于定性实验，我们在 [-3, 3] 的范围内插值遍历了单独的某个特征维度，保持其他维度值不变，进行重构。我们手工挑选了与人类概念最相近的某些维度进行展示。其中，dSprites 和 3D Faces 数据集结果如下：



(a) dSprites (b) 3D Faces

我们可以看到，ICP 在 dSprites 数据集上解耦出了旋转，在 3D Faces 数据集上解耦出了人脸宽度这样细粒度的语义概念。在 CelebA 数据集上的结果如下：



(a) Smile (b) Goatee

(c) Eyeglasses (d) Hair Color

我们可以看到，ICP 在 CelebA 数据集上解耦出了类似胡子之类的细粒度语义概念。

5. 总结

信息竞争式的多样化特征学习的本质在于使得特征携带更多样有效的信息，为特征学习的研究提供了新思路。