CPC

无监督对比预测学习

自动化钱61　戚伟健

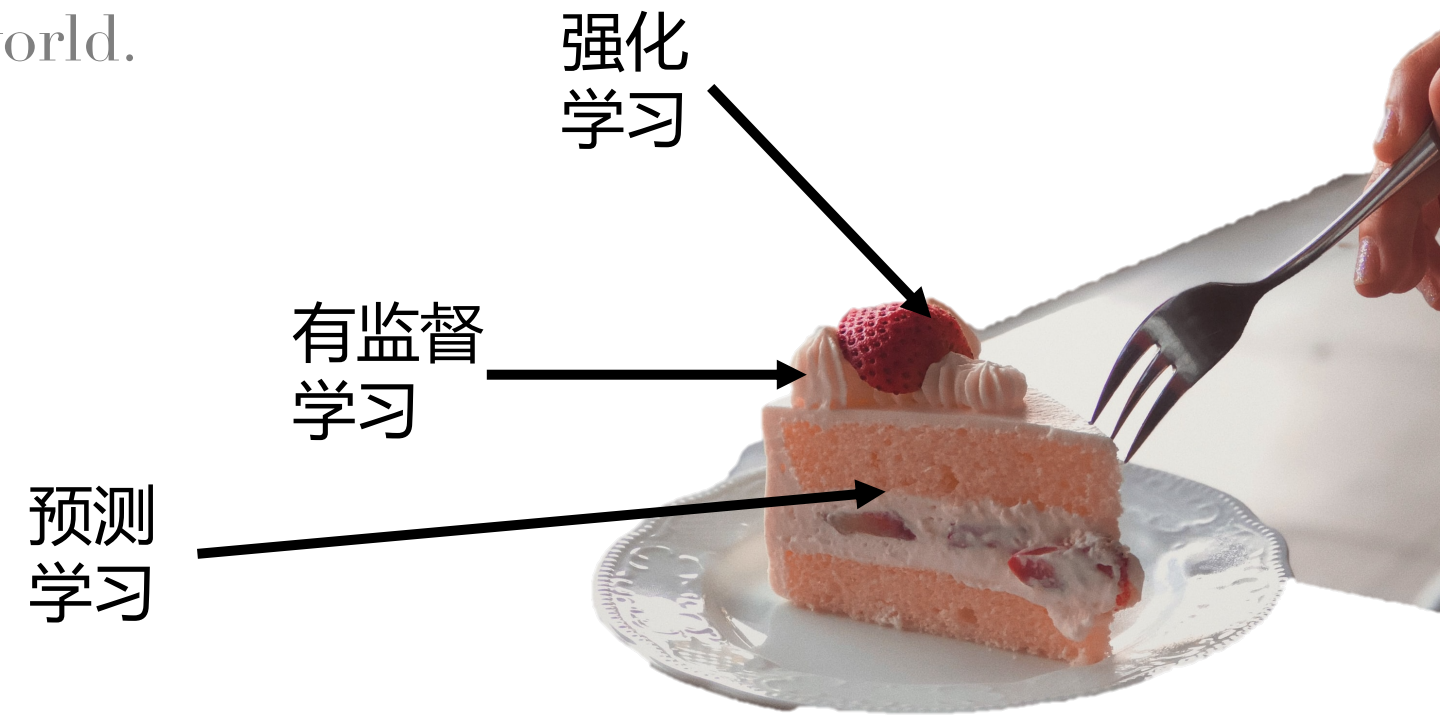
应数61　秦天毓

工业工程71　周梦豪

预测学习

“ Current AI systems are limited in that they can’t understand the real world. These systems **lack common sense** and **the ability to predict** based on an understanding of the natural world.

Yann Le Cunn @2016NIPS



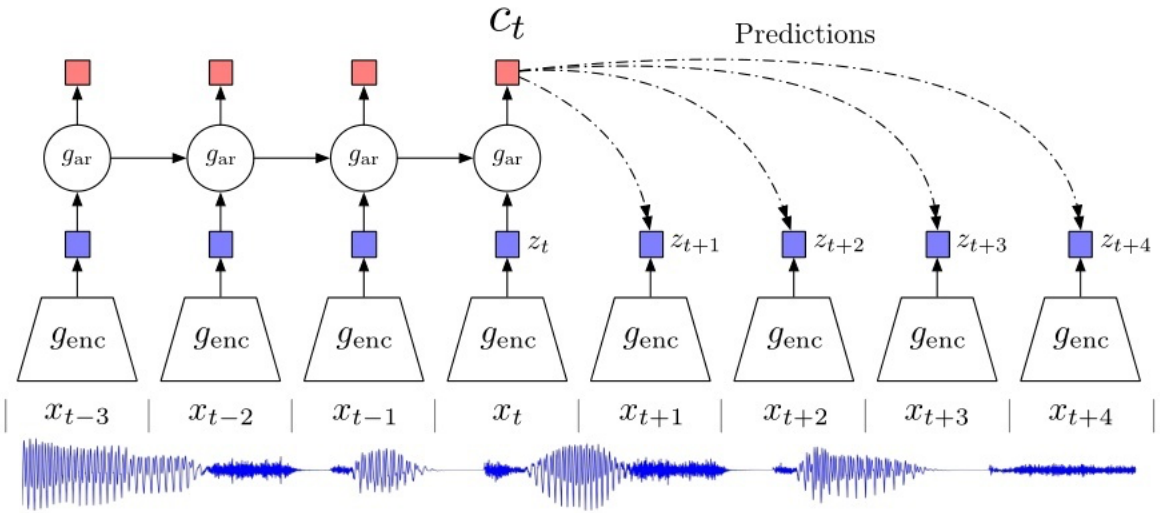


Figure 1: Overview of Contrastive Predictive Coding, the proposed representation learning approach. Although this figure shows audio as input, we use the same setup for images, text and reinforcement learning.

Contrastive Predictive Code

Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding[J]. arXiv preprint arXiv:1807.03748, 2018.

数据压缩/Encoder

Mutual information

$$\mathrm{I}(x;c) = \sum_{x,c} p(x,c) \log \frac{p(x|c)}{p(x)}$$

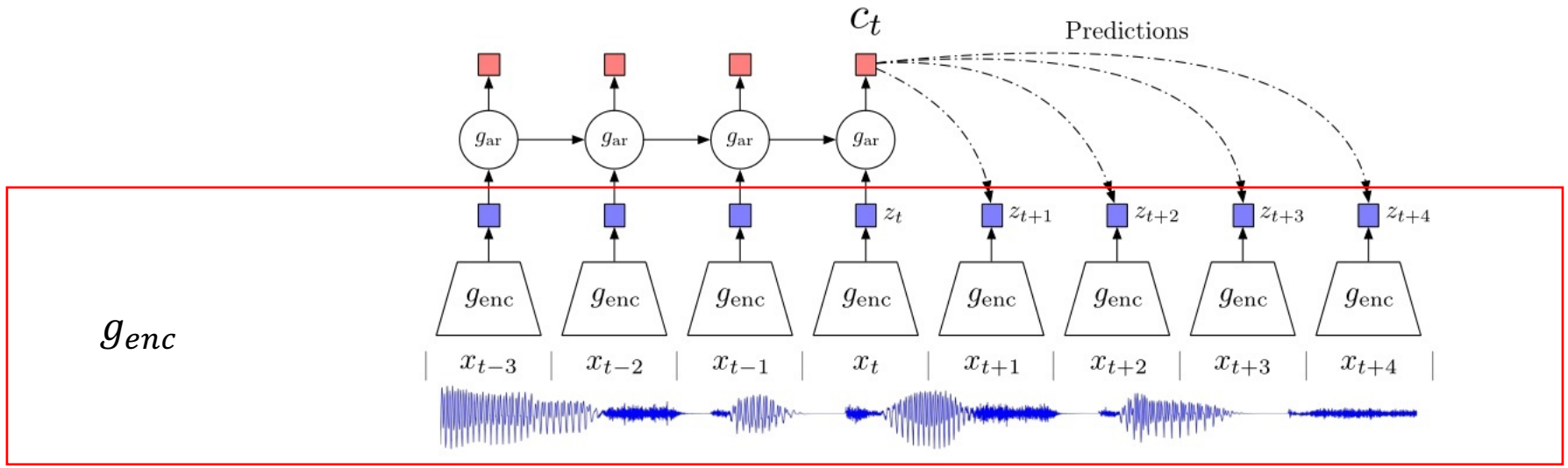


Figure 1: Overview of Contrastive Predictive Coding, the proposed representation learning approach. Although this figure shows audio as input, we use the same setup for images, text and reinforcement learning.

Contrastive Predictive Coding

$$f_k(x_{t+k}, c_t) \propto \frac{p(x_{t+k}|c_t)}{p(x_{t+k})}$$

$$f_k\left(x_{t+k}, c_t\right) = \exp\left(z_{t+k}^T W_k c_t\right)$$

自回归模型g_{ar}

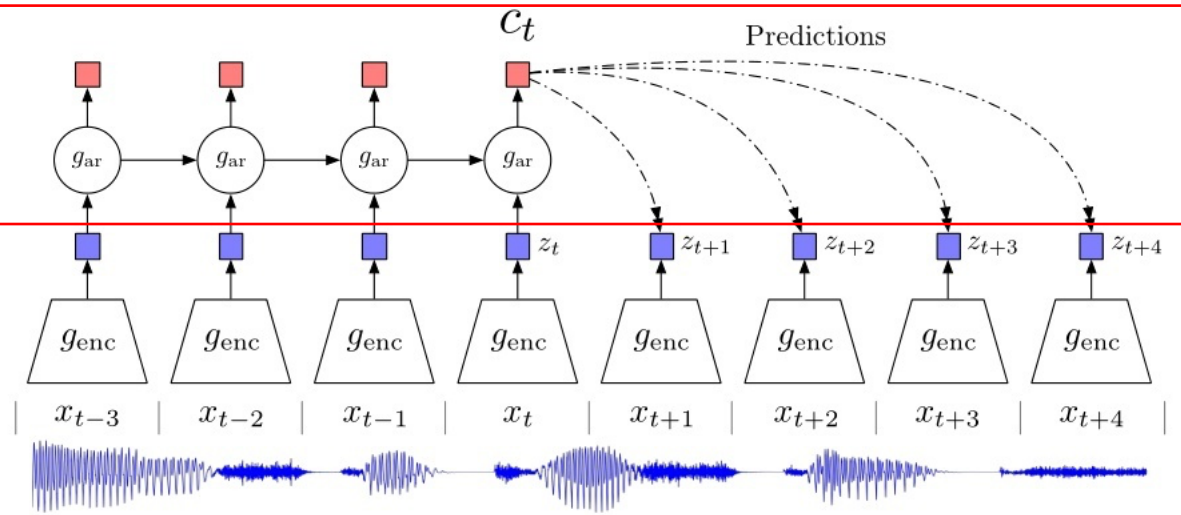


Figure 1: Overview of Contrastive Predictive Coding, the proposed representation learning approach. Although this figure shows audio as input, we use the same setup for images, text and reinforcement learning.

InfoNCE Loss

同时优化encoder和自回归模型

$$\mathcal{L}_{\mathrm{N}} = -\mathop{\mathbb{E}}_{X}\left[\log \frac{f_k(x_{t+k}, c_t)}{\sum_{x_j \in X} f_k(x_j, c_t)}\right]$$

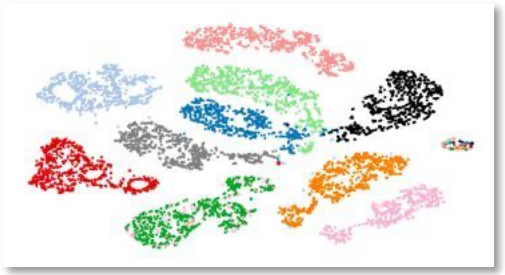
样本来自于假设条件分布的概率

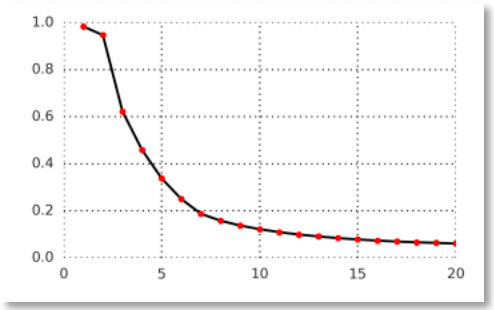
$$p(d=i|X,c_t) = \frac{p(x_i|c_t)\prod_{l\neq i} p(x_l)}{\sum_{j=1}^{N} p(x_j|c_t)\prod_{l\neq j} p(x_l)} = \frac{\frac{p(x_i|c_t)}{p(x_i)}}{\sum_{j=1}^{N}\frac{p(x_j|c_t)}{p(x_j)}}.$$

交互信息下界

$$I\left(x_{t+k}, c_t\right) \geq \log(N) - \mathcal{L}_{\mathrm{N}}$$

实验（来自论文）





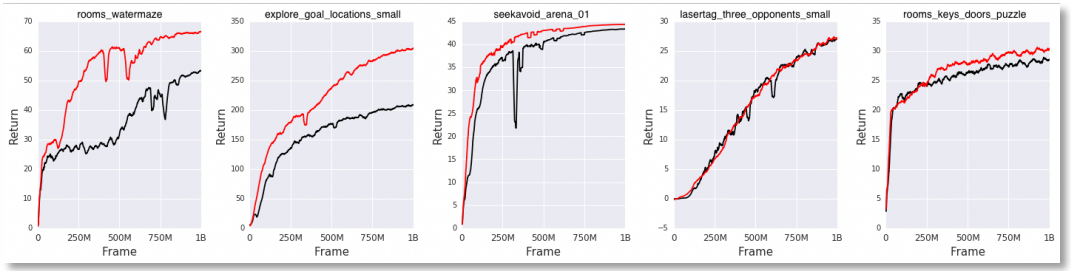
Method	ACC
Phone classification	
Random initialization	27.6
MFCC features	39.7
CPC	64.6
Supervised	74.6
Speaker classification	
Random initialization	1.87
MFCC features	17.6
CPC	97.4
Supervised	98.5

Method	ACC
#steps predicted	
2 steps	28.5
4 steps	57.6
8 steps	63.6
12 steps	64.6
16 steps	63.8
Negative samples from	
Mixed speaker	64.6
Same speaker	65.5
Mixed speaker (excl.)	57.3
Same speaker (excl.)	64.6
Current sequence only	65.2

Speech recognition

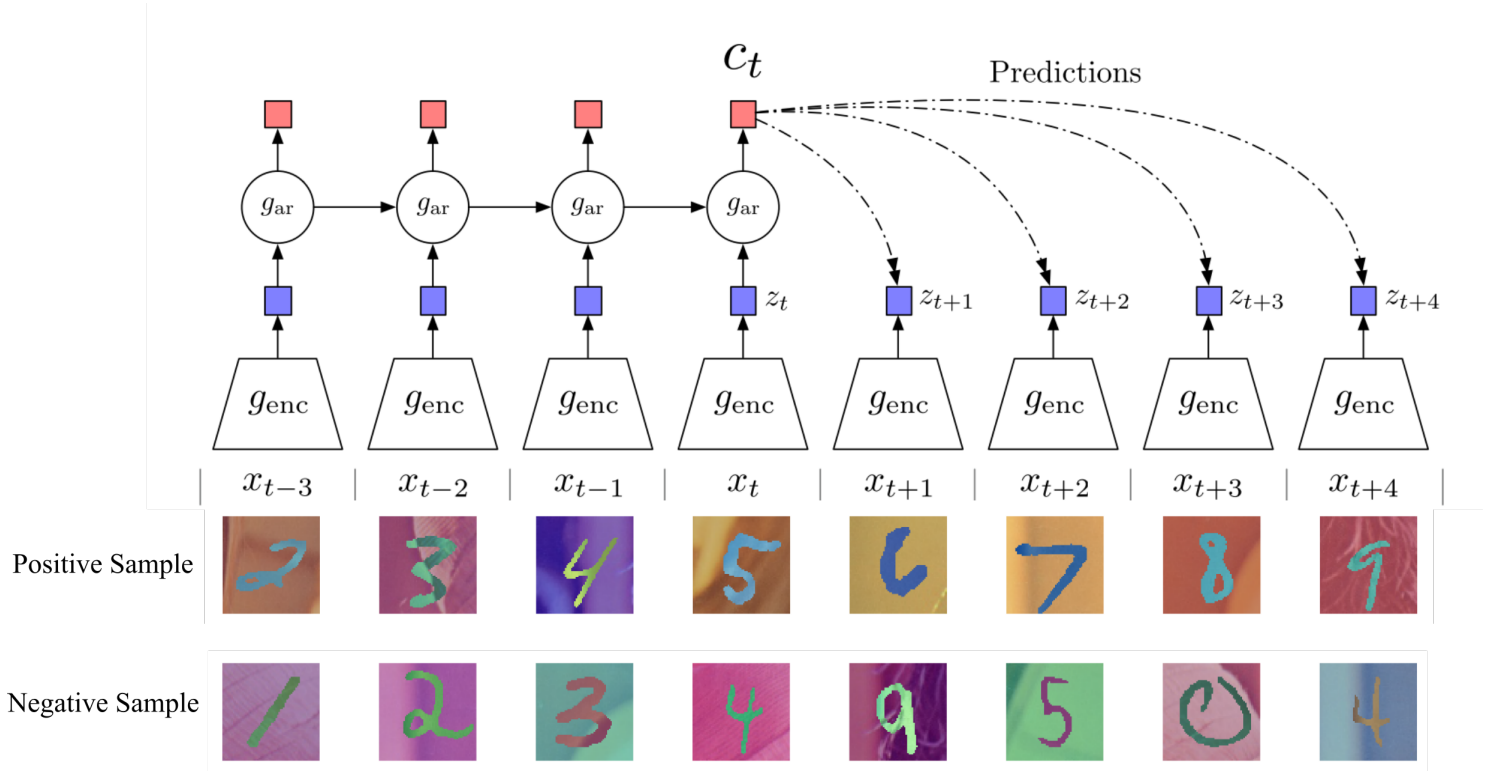
Method	MR	CR	Subj	MPQA	TREC
Paragraph-vector [40]	74.8	78.1	90.5	74.2	91.8
Skip-thought vector [26]	75.5	79.3	92.1	86.9	91.4
Skip-thought + LN [41]	79.5	82.6	93.4	89.0	-
CPC	76.9	80.1	91.2	87.7	96.8

NLP



Reinforcement learning

动手实践



99%
accuracy

https://github.com/davidtellez/contrastive-predictive-coding

半监督的CPC

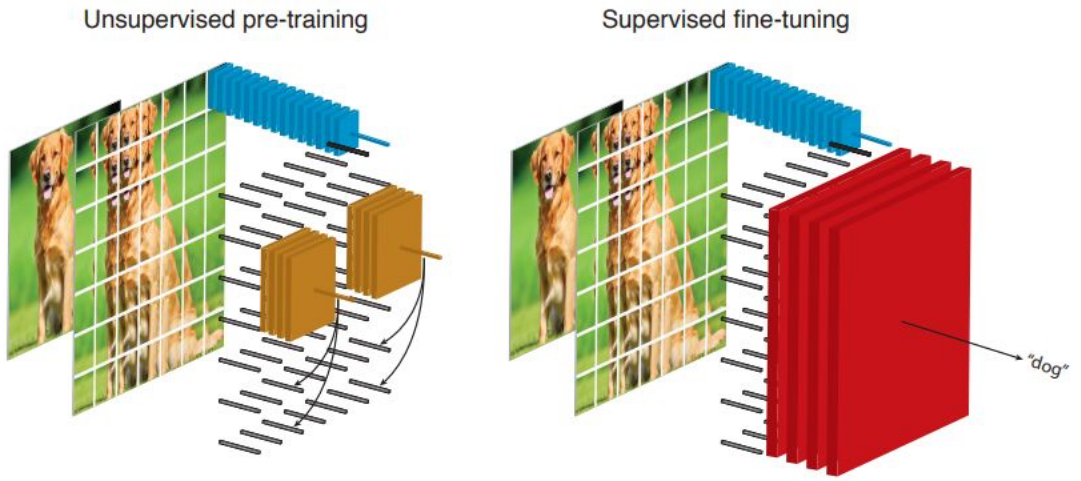
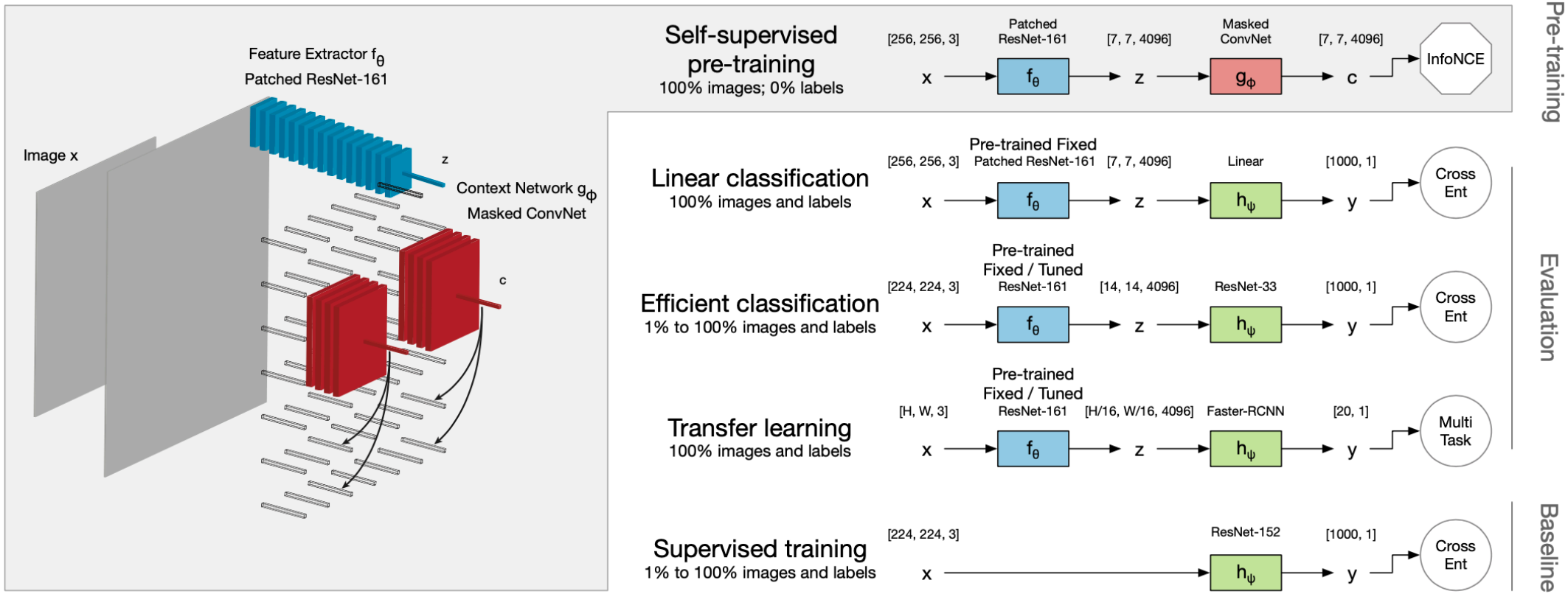


Figure 2. Overview of the framework for semi-supervised learning with Contrastive Predictive Coding. Left: unsupervised pre-training with a spatial prediction task. First, an image is divided into a grid of overlapping patches. Each patch is encoded independently from the rest with a feature extractor (blue) which terminates with a mean-pooling operation, yielding a single feature vector for that patch. Doing so for all patches yields a field of such feature vectors (wireframe vectors). Feature vectors above a certain level (in this case, the center of the image) are then aggregated with a context network (brown), yielding a row of context vectors which are used to linearly predict (unseen) features vectors below. Right: using the CPC representation for a classification task. Having trained the encoder network, the context network is discarded and replaced by a classifier network (red) which can be trained in a supervised manner. For some experiments, we also fine-tune the encoder network (blue) for the classification task.

Hénaff O J, Razavi A, Doersch C, et al. Data-efficient image recognition with contrastive predictive coding[J]. arXiv preprint arXiv:1905.09272, 2019.

半监督CPC框架



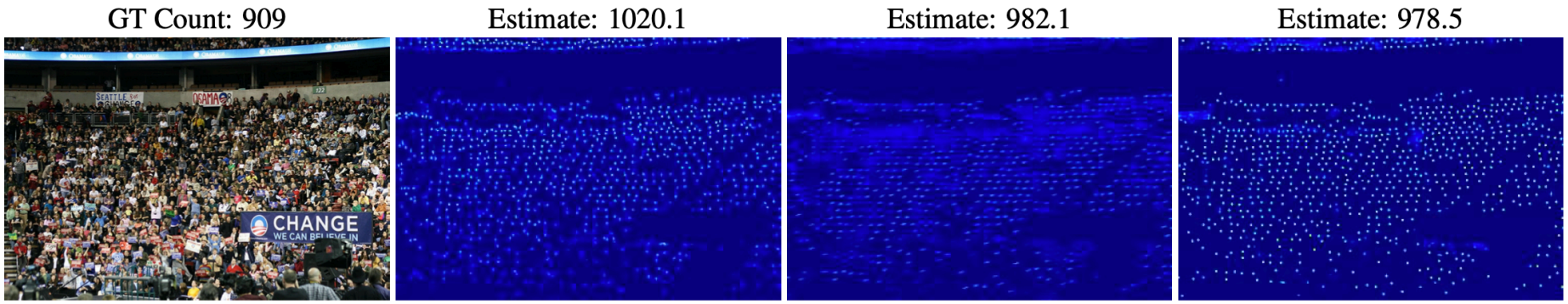
$$\theta^* = \arg\min_\theta \frac{1}{N}\sum_{n=1}^{N} \mathcal{L}_{\text{CPC}}[f_\theta(x_n)]$$

$$\psi^* = \arg\min_\psi \frac{1}{M}\sum_{m=1}^{M} \mathcal{L}_{\text{Sup}}[h_\psi \circ f_{\theta^*}(x_m), y_m]$$

性能

Method	Architecture	Top-5 accuracy				
Labeled data		1%	5%	10%	50%	100%
†Supervised baseline	ResNet-200	44.1	75.2⋆	83.9	93.1	95.2#
Methods using label-propagation:						
Pseudolabeling [63]	ResNet-50	51.6	-	82.4	-	-
VAT + Entropy Minimization [63]	ResNet-50	47.0	-	83.4	-	-
Unsup. Data Augmentation [61]	ResNet-50	-	-	88.5	-	-
Rotation + VAT + Ent. Min. [63]	ResNet-50 ×4	-	-	**91.2**	-	95.0
Methods using representation learning only:						
Instance Discrimination [60]	ResNet-50	39.2	-	77.4	-	-
Rotation [63]	ResNet-152 ×2	57.5	-	86.4	-	-
ResNet on BigBiGAN (fixed)	RevNet-50 ×4	55.2	73.7	78.8	85.5	87.0
ResNet on AMDIM (fixed)	Custom-103	67.4	81.8	85.8	91.0	92.2
ResNet on CPC v2 (fixed)	ResNet-161	77.1	87.5	90.5	95.0	96.2
ResNet on CPC v2 (fine-tuned)	ResNet-161	**77.9**⋆	**88.6**	**91.2**	**95.6**#	**96.5**

移植CPC



Ma Z, Wei X, Hong X, et al. Bayesian loss for crowd count estimation with point supervision[C]//Proceedings of the IEEE International Conference on Computer Vision. 2019: 6142-6151.