NLP中的
自监督学习&对比学习

吴金龙@爱因互动

2020.09

吴金龙

一个AI 创建人，2016

- www.yige.ai
- 免费好用的中文Bot Maker

ChatbotsChina 发起人，2016

- Bot 相关的技术、产品、运营
- 微信公众号/交流群、微博

爱因互动 合伙人，2017

- www.einplus.cn
- 技术合伙人/算法负责人

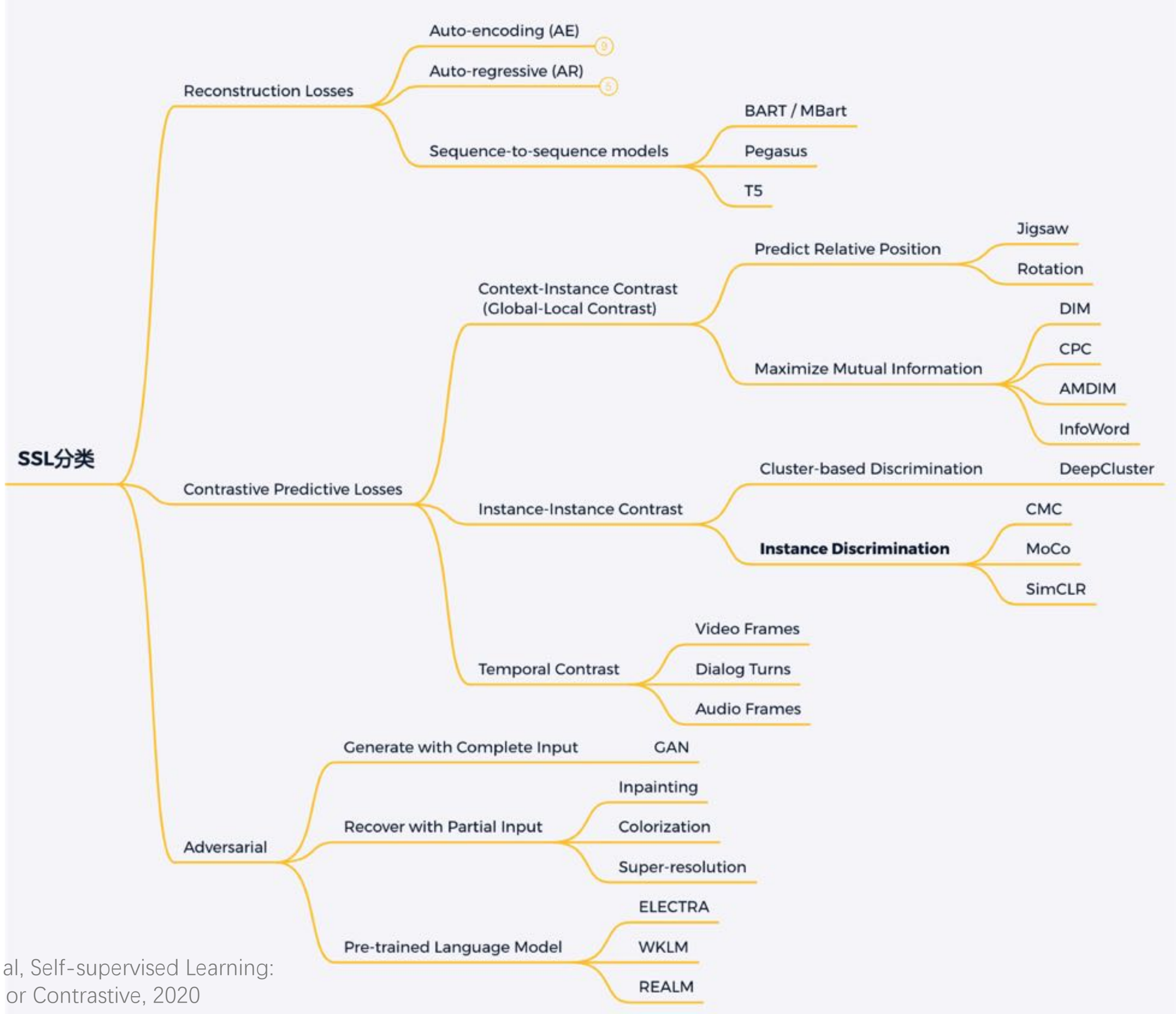
SSL

- 大致分成两类：生成式、判别式
- 生成式
  - 期望利用数据表示重构完整数据
- 判别式
  - 期望数据表示包含足够多信息即可

Contrastive Self-Supervised Learning | Ankesh Anand

SSL

- 分类



Xiao Liu et al, Self-supervised Learning: Generative or Contrastive, 2020

SSL

- 评价方法
  - Linear Evaluation Protocol
    - 下游任务训练时，模型参数固定，只用于获得样本表示向量，在有监督数据上训练线性模型看效果

  - Downstream Tasks
    - 看模型在单个或多个下游任务上的效果

对比学习

- **CL：让像的样本表示差异小，让不像的样本表示差异大**
- 期望学到更通用的知识，与辅助任务无关

- 学习方法

$$\underbrace{\text{score}(f(x), f(x^+))} >> \underbrace{\text{score}(f(x), f(x^-))}$$

- here x^+ is data point similar or congruent to x, referred to as a *positive* sample.
- x^- is a data point dissimilar to x, referred to as a *negative* sample.
- the score function is a metric that measures the similarity between two features.

  - 对比什么：什么和什么对比？
  - 对比评判：如何量化差异？
    - 损失函数

对比学习

- 常用损失函数
  - **Noise-contrastive Estimation (NCE):** negative sampling

$$\log \sigma(\boldsymbol{u}^T \boldsymbol{v}^+ / \tau) + \log \sigma(-\boldsymbol{u}^T \boldsymbol{v}^- / \tau)$$

  - **InfoNCE**

$$\mathcal{L}_q = -\log \frac{\exp(q \cdot k_+ / \tau)}{\sum_{i=0}^{K} \exp(q \cdot k_i / \tau)}$$

Temperature Hyperparameter

  - **Margin Triplet**

$$-\max(\boldsymbol{u}^T \boldsymbol{v}^- - \boldsymbol{u}^T \boldsymbol{v}^+ + m, 0)$$

对比学习

- **MI 和 InfoNCE**
  - **互信息 Mutual Information (MI)**

$$\boxed{I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \| P_X \otimes P_Y)}$$

$$\boxed{\mathrm{I}(X;Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} p_{(X,Y)}(x,y) \log\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)\, p_Y(y)}\right) dx\, dy, \qquad \text{(Eq.2)}}$$

    - **X与Y相关性越强，I(X; Y) 越大**

  - **InfoNCE是MI的下界**

$$I(A,B) \geq \mathbb{E}_{p(A,B)}\left[f_{\boldsymbol{\theta}}(a,b) - \mathbb{E}_{q(\tilde{\mathcal{B}})}\left[\log \sum_{\tilde{b} \in \tilde{\mathcal{B}}} \exp f_{\boldsymbol{\theta}}(a,\tilde{b})\right]\right] + \log |\tilde{\mathcal{B}}|$$

Linear Relational Embedding (LRE)

- LRE (Paccanaro & Hinton, 2000) introduced the idea of using a **contrastive loss** for modelling relational data
  - To model Mother-of (John) = Victoria, learn a vector j for John, a vector v for Victoria and a matrix M for mother-of
  - Want M j to be closer to v than to the vectors representing other objects
  - Maximize

$$-\|\mathbf{Mj}-\mathbf{v}\|^2 - \log\left(\sum_k e^{\left(-\|\mathbf{Mj}-\mathbf{k}\|^2\right)}\right)$$

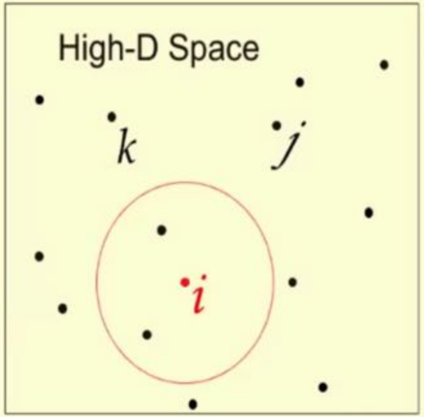
LRE → SNE (with one relation similar-to)

Applying the LRE objective function to dimensionality reduction

- Compute a big probability table that contains the probability that each high-dimensional data-point , i, would pick another data-point , j, as its "neighbor".
  - This probability is proportional to exp(- $||x_i – x_j||$^2)
- Learn to convert each high-dimensional data-point x_i to a 2-D map point, y_i
  - Compare resulting 2-D map point, y_i, with all the other 2-D map points, y_j, to get a probability that y_i would pick y_j as its neighbor.
  - Probability is proportional to exp(- $||y_i – y_j||$^2)
- Learn the 2-D locations of the map points so that the probabilities of picking a neighbor computed in the 2-D space match the probabilities computed in the original high-D space.

A picture of SNE

- Each point in high-D has a conditional probability of picking each other point as its neighbor.
- The distribution over neighbors is based on the high-D pairwise distances.



$$p_{j|i} = \frac{e^{-d_{ij}^2 / 2\sigma_i^2}}{\sum_k e^{-d_{ik}^2 / 2\sigma_i^2}}$$

probability of picking j given that you start at i

Stochastic Neighbor Embedding (SNE)

- 把N个高维的数据 x_1,…,x_N，对应映射成N个低维的数据 y_1,…,y_N

$$p_{j|i} = \frac{\exp\left(-\|x_i - x_j\|^2 / 2\sigma_i^2\right)}{\sum_{k \neq i} \exp\left(-\|x_i - x_k\|^2 / 2\sigma_i^2\right)}$$

$$q_{j|i} = \frac{\exp\left(-\|y_i - y_j\|^2\right)}{\sum_{k \neq i} \exp\left(-\|y_i - y_k\|^2\right)}$$

- Cost Function (opt: SGD with Momentum)

$$C = \sum_i KL(P_i \| Q_i) = \sum_i \sum_j p_{j|i} \log \frac{p_{j|i}}{q_{j|i}}$$

t-Distributed Stochastic Neighbor Embedding (t-SNE)

- Cost Function (conditional probability → joint probability)

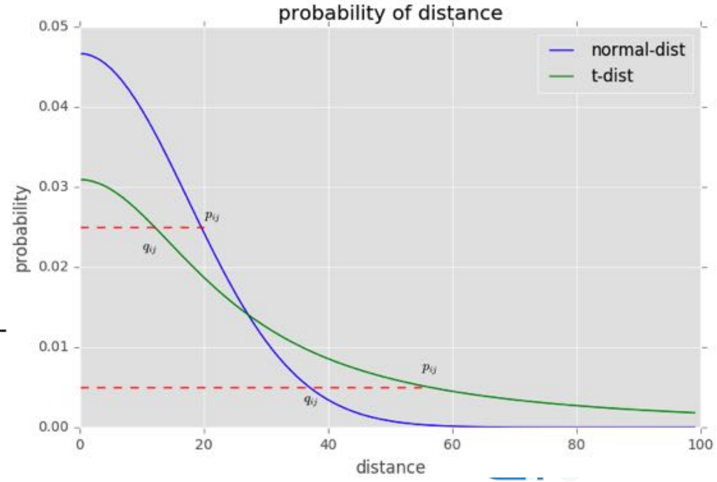
$$C = KL(P\|Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

- Symmetric SNE

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}$$

- Normal Dist → Student t-Dist

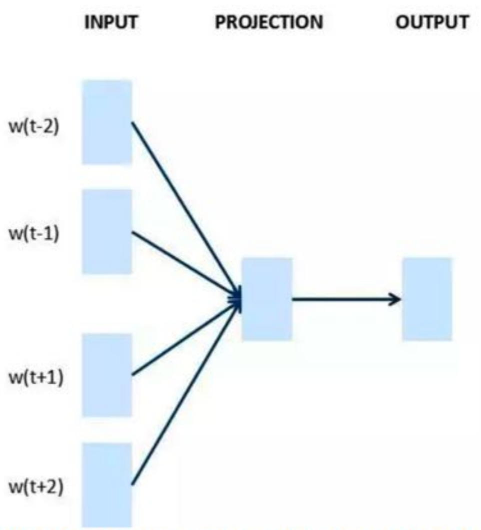
$$q_{ij} = \frac{\left(1 + \|y_i - y_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|y_k - y_l\|^2\right)^{-1}}$$



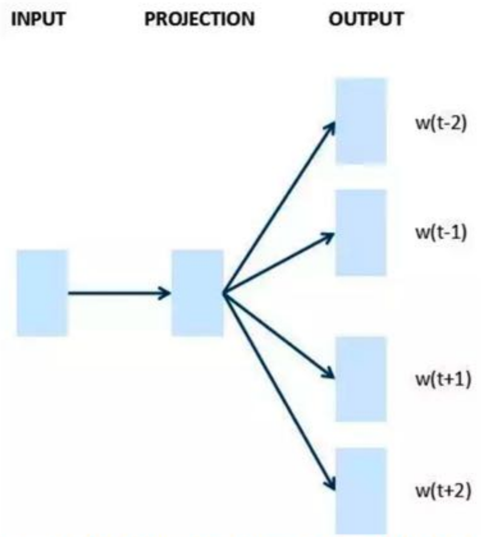
t-SNE完整笔记

Reconstruction / AE

- Word2Vec



CBOW (Continuous Bag-of-Words Model)

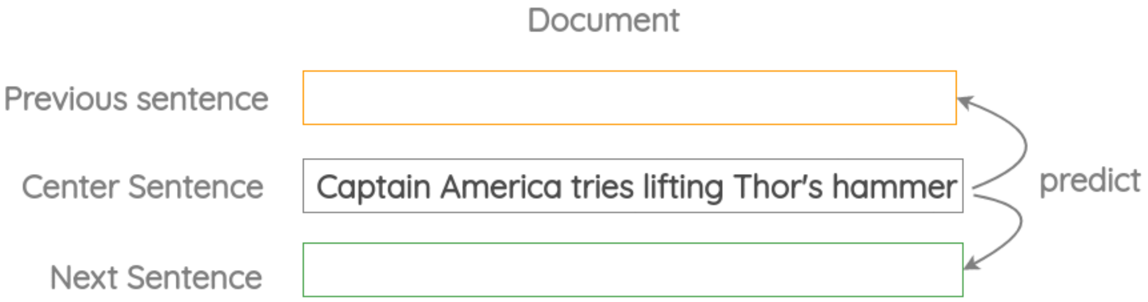


Skip-gram (Continuous Skip-gram Model)

Self–supervised Learning 再次入门 – 知乎

Reconstruction / AE

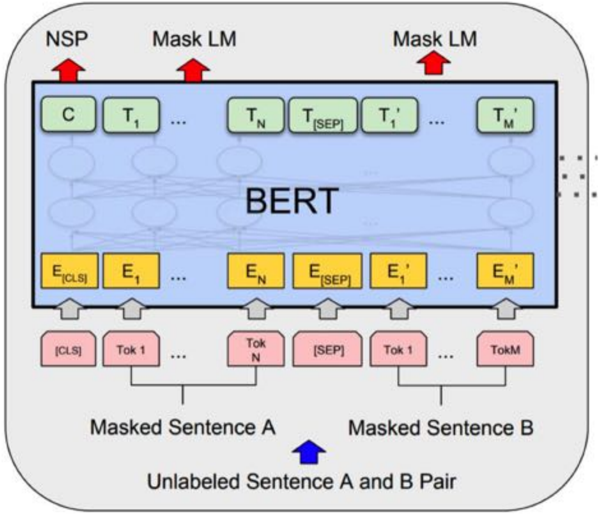
- Neighbor Sentence Prediction
  - **利用中间的句子，生成前一句与后一句**
  - **句级别的 Skip-gram**



used in the Skip-Thought Vectors paper

Reconstruction / AE

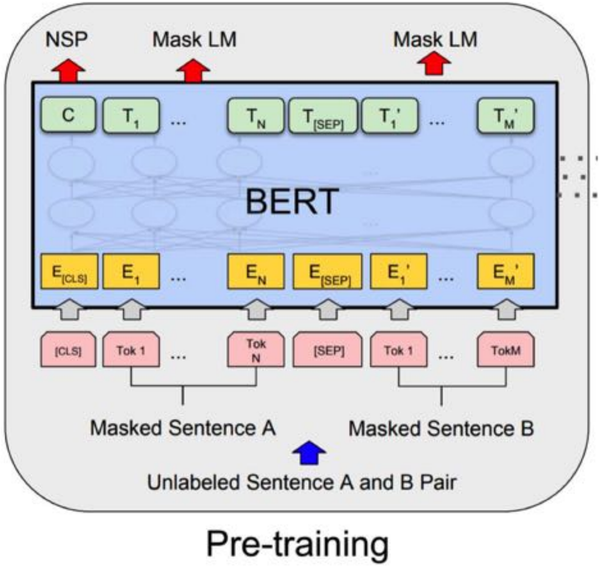
- Masked LM
  - BERT, RoBERTa, ALBERT



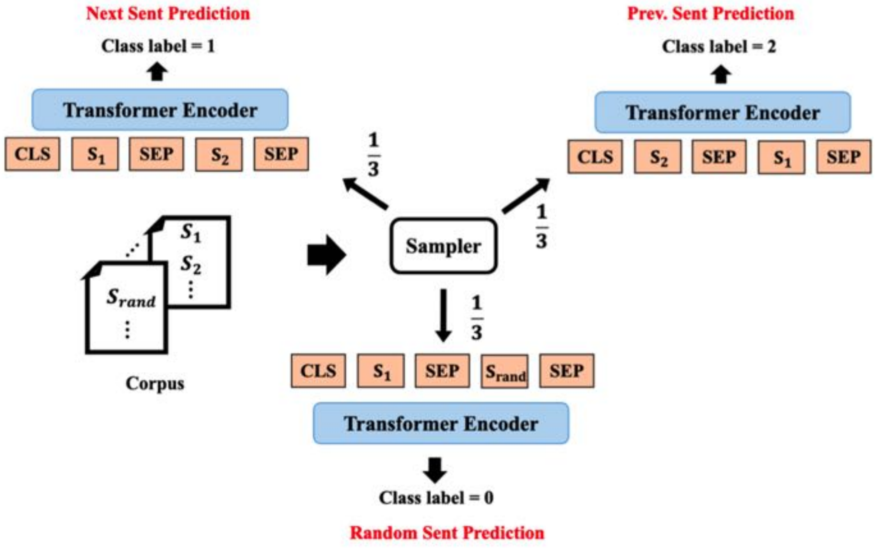
Pre-training

Reconstruction / AE

- **Next Sentence Prediction (NSP)**
  - BERT

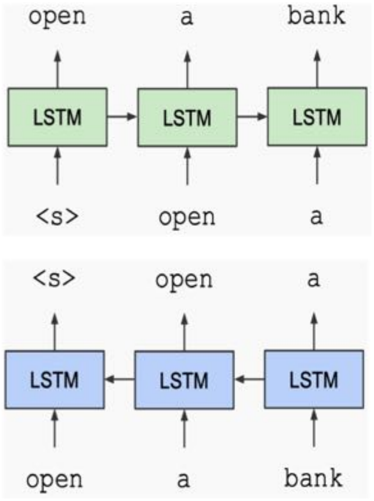


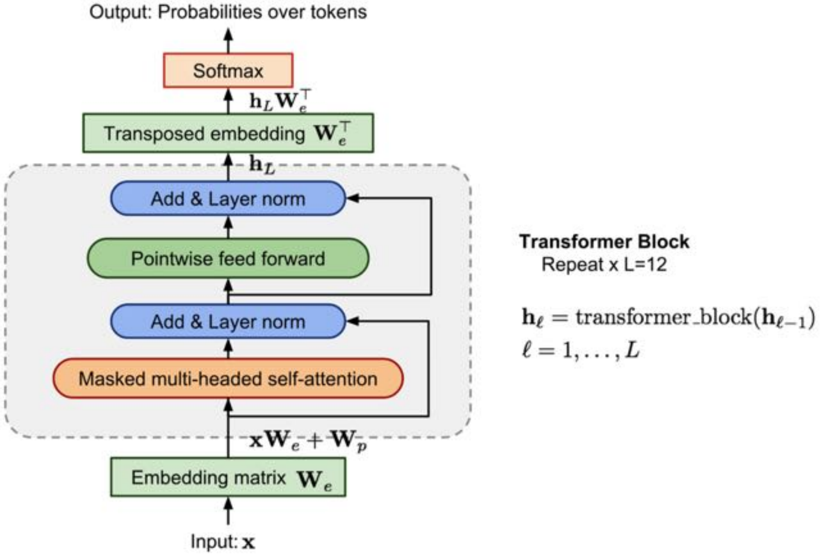
- **Sentence Order Prediction (SOP)**
  - ALBERT, StructBERT



Reconstruction / AR

- Auto-regressive LM
  - **LSTM, ELMo**
  - **GPT, GPT2**





Reconstruction / Sequence-to-sequence

- More Generative Tasks
  - **MASS**

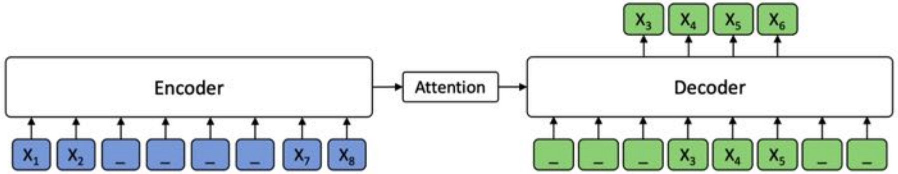
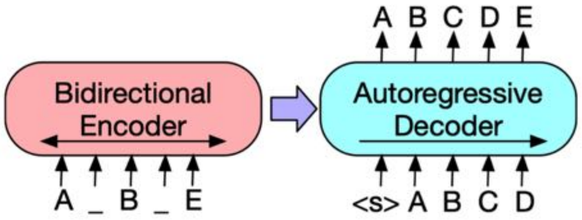
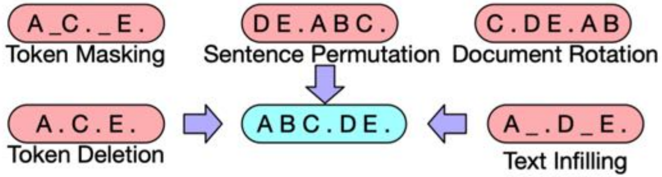


Figure 1. The encoder-decoder framework for our proposed MASS. The token "_" represents the mask symbol [M].

  - **T5**
  - **BART**
    - regenerate corrupted inputs
    - Transformations for noising the input





【FB-BART新的预训练模型】阅读笔记

Reconstruction / Sequence-to-sequence

- **Sentence Permutation**
  - 恢复被打乱顺序的句子（BART）

I did X. Then I did Y. Finally I did Z.

- **Document Rotation**
  - 随机选一个词作为旋转中心，旋转一段话（BART）

I am going outside. I will be back in the evening.

original text

Contrastive / Context-Instance

- **QuickThoughts**
  - Google 2018年的工作，把相邻的句子当做正样本，不相邻的句子作为负样本，做对比学习训练，提升相邻句子的概率值

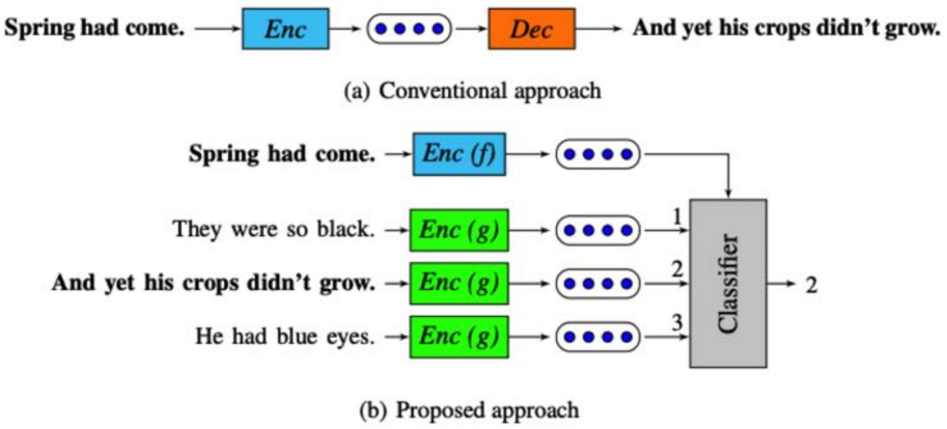


Figure 1: Overview. (a) The approach adopted by most prior work where given an input sentence the model attempts to generate a context sentence. (b) Our approach replaces the decoder with a classifier which chooses the target sentence from a set of candidate sentences.

Contrastive / Context-Instance

- **InfoWord**
  - anchor是一个句子，但把其中选中的n-grams mask掉，而正样本就是mask掉的n-grams，负样本就是其他anchor中对应的n-grams
  - DIM 在文本上的适配升级版

$$\mathcal{I}_{\text{INFOWORD}} = \lambda_{\text{MLM}}\mathcal{I}_{\text{MLM}} + \lambda_{\text{DIM}}\mathcal{I}_{\text{DIM}}$$

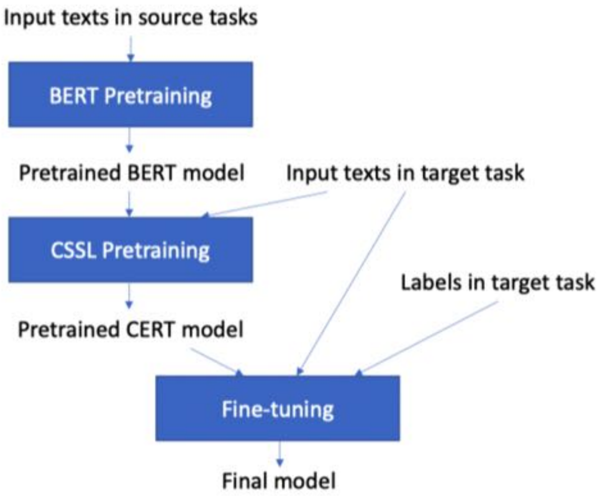
Table 1: Summary of methods as instances of contrastive learning. See text for details.

Objective	a	b	$p(a, b)$	g_ω	g_ψ
Skip-gram	word	word	word and its context	lookup	lookup
MLM	context	masked word	masked tokens probability	Transformer	lookup
NSP	sentence	sentence	(non-)consecutive sentences	Transformer	lookup
XLNet	context	masked word	factorization permutation	TXL++	lookup
DIM	context	masked n-grams	sentence and its n-grams	Transformer	not used

A Mutual Information Maximization Perspective of Language Representation Learning

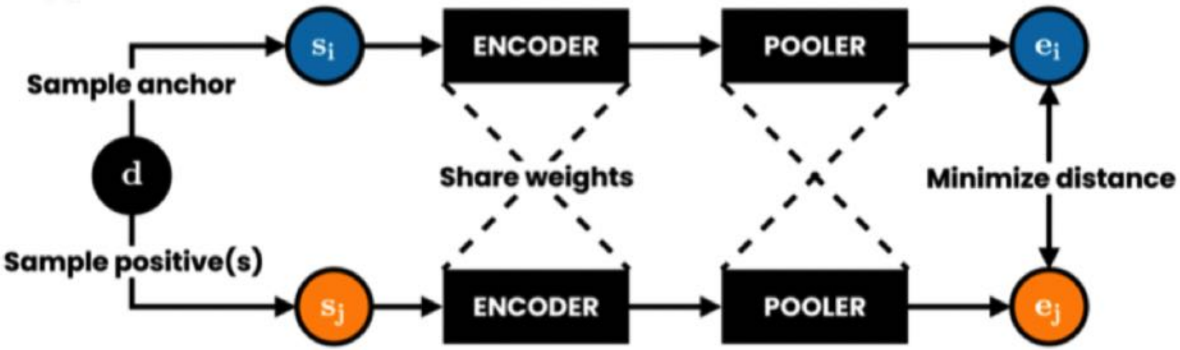
Contrastive / Instance-Instance

- **CERT: Contrastive Self-supervised Learning for Language Understanding**
  - 把MoCo直接搬到NLP，数据增强使用回译
  - 下游任务效果略优于BERT



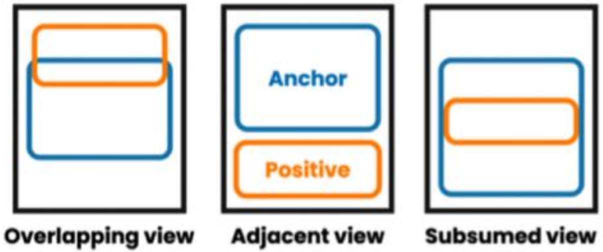
Contrastive / Instance-Instance

- **DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations**
  - 先从文档中随机选取anchor，然后在anchor附近随机选取正样本
  - 负样本就是其他anchor的正样本
  - 优化目标就是InfoNCE，加上MLM损失



Contrastive / Instance-Instance

- **DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations**
  - 正样本有以下三种情况（蓝色是anchor，黄色是正样本）



  - 负样本有两种：
    - 简单负样本：来自于其他文档
    - 困难负样本：来自同一个文档

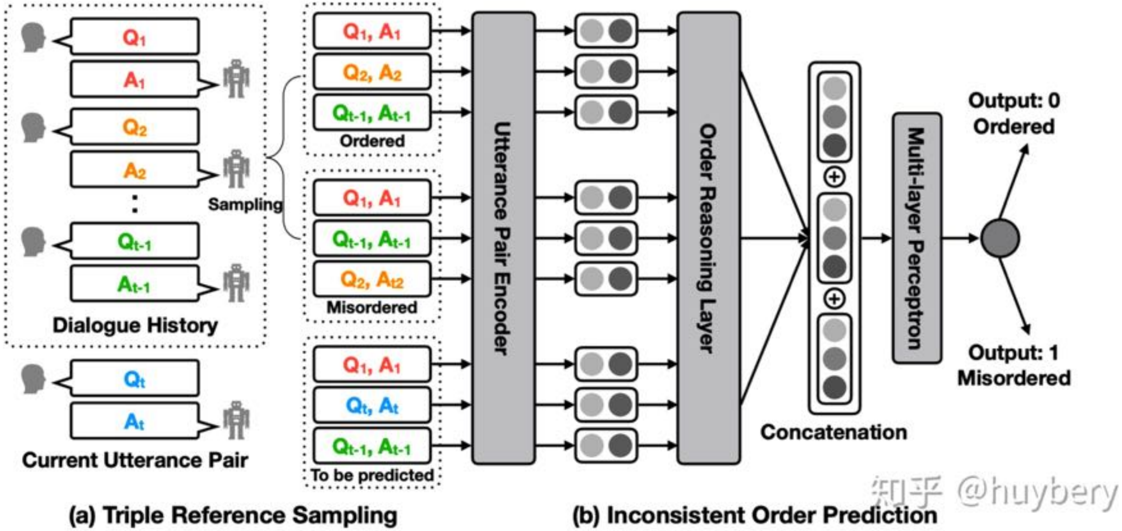
Contrastive / Instance-Instance

- **DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations**
  - SentEval 上的效果
    - 18 downstream tasks: representative NLP tasks such as sentiment analysis, natural language inference, paraphrase detection and image-caption retrieval
    - 10 probing tasks, which are designed to evaluate what linguistic properties are encoded in a sentence representation

			SentEval			
Model	Parameters	Embed. dim.	Downstream	Probing	Avg.	Δ
Unsupervised						
Transformer-small	82M	768	72.69	**74.27**	73.48	-2.50
Transformer-base	125M	768	72.22	73.38	72.80	-3.18
DeCLUTR-small (ours)	82M	768	76.80	73.84	75.32	-0.66
DeCLUTR-base (ours)	125M	768	78.16	73.80	**75.98**	–

Contrastive / Temporal

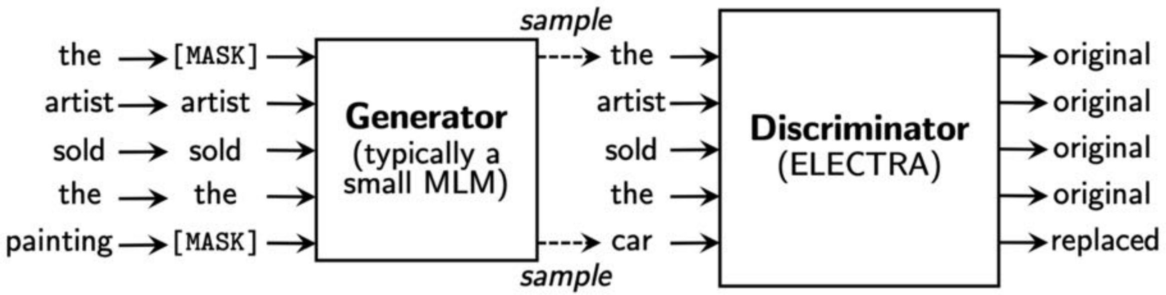
- 对话
  - 从大量的历史预料中挖掘出顺序的序列（positive）和乱序的序列（negative），通过模型来预测是否符合正确的顺序来进行训练



[Self-Supervised Dialogue Learning](#) [Self-supervised Learning 再次入门 - 知乎](#)

Adversarial

- **Replaced Token Detection (RTD)**
  - **ELECTRA**: instead of generating tokens for MLM, discriminate whether a token is **original**



  - More efficient (each token can be used for training)

Others / Supervised

- **Sentence-BERT**: Sentence Embeddings using Siamese BERT-Networks
  - 用训练好的 BERT 直接获得句子向量（`[CLR]`，或者mean/max pooling）效果一般
  - Sentence-BERT的思路是之后用带标注的数据集再做精调，这样可以获得更好的句子向量
  - 精调训练数据可以使用NLI数据（`contradiction`, `entailment`, and `neutral`），最上层的softmax classifier就是三分类模型

Others / Supervised

- **Sentence-BERT**: Sentence Embeddings using Siamese BERT-Networks
  - 精调方法如下，左边是精调训练结构，右边是推断结构

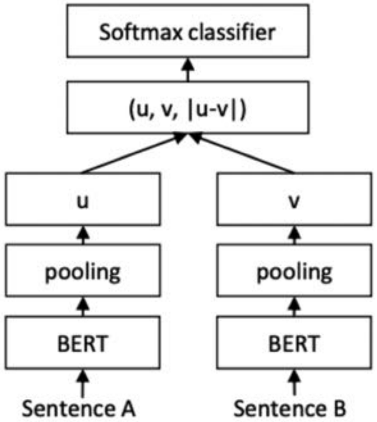


Figure 1: SBERT architecture with classification objective function, e.g., for fine-tuning on SNLI dataset. The two BERT networks have tied weights (siamese network structure).

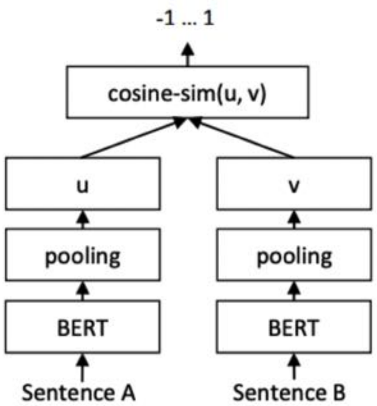
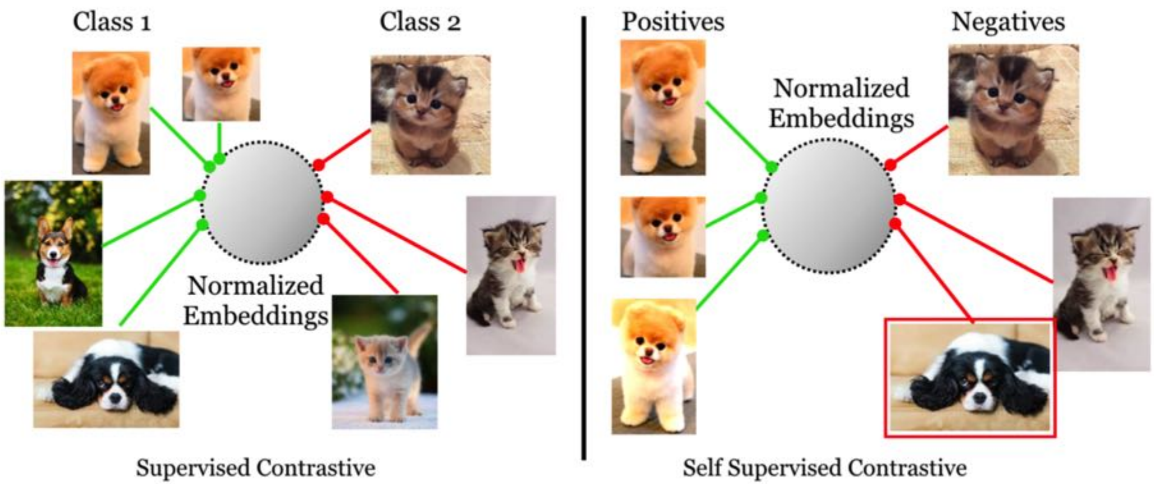


Figure 2: SBERT architecture at inference, for example, to compute similarity scores. This architecture is also used with the regression objective function.

Others / Supervised

- Supervised Contrastive Learning (**SupContrast**)
  - 同类别的数据作为正样本，不同类别的数据作为负样本



Others / Supervised

- Supervised Contrastive Learning (**SupContrast**)
  - 同类别的数据作为正样本，不同类别的数据作为负样本
  - Loss Function（原始的每个样本 x_i 扩充为两个新 x_i^1和x_i^2 ）

$$\mathcal{L}^{sup} = \sum_{i=1}^{2N} \mathcal{L}_i^{sup}$$

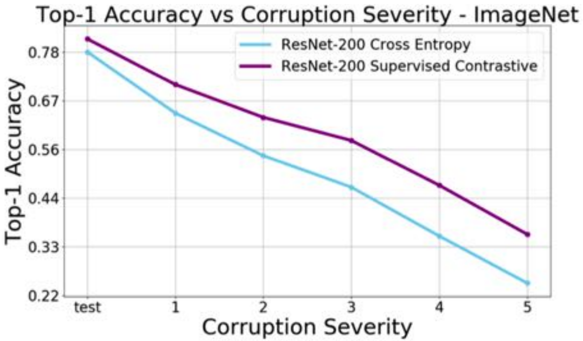
$$\mathcal{L}_i^{sup} = \frac{-1}{2N_{\tilde{\boldsymbol{y}}_i} - 1} \sum_{j=1}^{2N} \mathbb{1}_{i \neq j} \cdot \mathbb{1}_{\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{y}}_j} \cdot \log \frac{\exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_j / \tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{i \neq k} \cdot \exp\left(\boldsymbol{z}_i \cdot \boldsymbol{z}_k / \tau\right)}$$

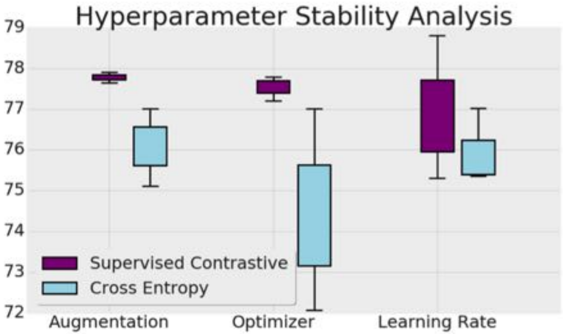
  - 两步
    1. 利用上面的损失函数做对比训练
    2. 去掉投影层，用标准分类损失精调上面的模型

Others / Supervised

- Supervised Contrastive Learning (**SupContrast**)
- 结论
  - better accuracy than cross entropy
  - robustness to Image Corruptions and Calibration
  - more stable to changes in hyperparameters

Loss	Architecture	Top-1	Top-5
Cross Entropy (baselines)	AlexNet [27]	56.5	84.6
	VGG-19+BN [42]	74.5	92.0
	ResNet-18 [20]	72.1	90.6
	MixUp ResNet-50 [56]	77.4	93.6
	CutMix ResNet-50 [55]	78.6	94.1
	Fast AA ResNet-50 [9]	77.6	95.3
	Fast AA ResNet-200 [9]	80.6	95.3
Cross Entropy (our implementation)	ResNet-50	77.0	92.9
	ResNet-200	78.0	93.3
Supervised Contrastive	ResNet-50	**78.8**	**93.9**
	ResNet-200	**80.8**	**95.6**





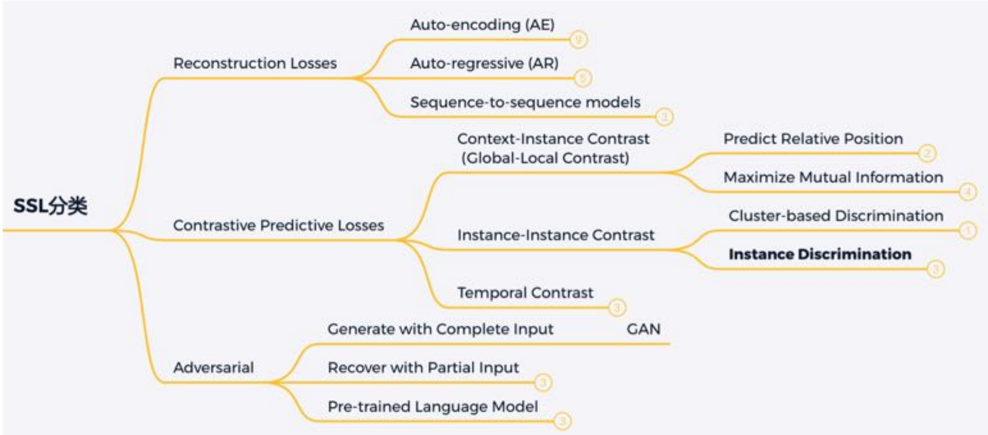
Others / Supervised

- **Emoji Prediction**
  - 把推文中的emoji作为类别，训练分类模型
  - fine-tuned it on emotion-related downstream tasks like sentiment analysis, hate speech detection and insult detection

used in the DeepMoji paper

总结

- 借助于CV中有效的数据增强方法，Instance Discrimination 类型的CL算法霸榜CV
- NLP很难有准确有效的数据增强方法，可能要走另一条路
  - 先利用任务同领域有监督数据（领域知识）精调模型，再利用任务数据二次精调模型
  - 探索背景和时序无监督数据
    - 改进 UDA？



总结

- NLP之路

Pretrained Model → 在任务同领域数据上精调模型 → 在任务数据上精调模型

Pretrained Model

- 大量无监督数据上训练好的模型，如BERT
- 可以在预训练之后，利用特定知识精调模型，如ERNIE

在任务同领域数据上精调模型

- Reconstruction Loss，如任务数据上继续训练LM模型
- CL Loss：基于背景和时序同领域无监督数据，使用CL精调模型
- Supervised：基于同领域有监督数据，精调模型

在任务数据上精调模型

NLP/Speech

Thanks !

Your business is in good hands .

References

- Self-supervised Learning 再次入门 - 知乎
- Self-Supervised Learning 入门介绍 - 知乎
- 无监督学习距离监督学习还有多远？Hinton组新作解读 - 知乎
- Contrastive Self-Supervised Learning | Ankesh Anand
- PW Live -对比学习及其在NLP中的应用
- github: awesome-self-supervised-learning
- Self-Supervised Learning. Andrew Zisserman (Oxford & Deepmind)
- The Illustrated Self-Supervised Learning
- Self-supervised learning and computer vision · fast.ai
- 【FB-BART新的预训练模型】阅读笔记
- The Illustrated PIRL: Pretext-Invariant Representation Learning
- The Illustrated SimCLR Framework

References

- Xiao Liu et al, Self-supervised Learning: Generative or Contrastive, 2020
- Self Supervised Representation Learning in NLP
- t-SNE完整笔记
- Summary of the models — transformers 3.1.0 documentation