# 对比学习的神经主题模型

Thong Nguyen, Luu Anh Tuan

**摘要**

最近的实证研究表明，对抗性主题模型 (ATM) 可以通过将文档与另一个不同的样本区分开来，成功地捕获文档的语义模式。然而，利用这种判别生成架构有两个重要的缺点：（1）该架构不关联那些具有相同显着词的文档词分布的相似文档；(2) 它限制了整合外部信息的能力，例如已被证明有利于神经主题模型训练的文档的情感。为了解决这些问题，我们从数学分析的角度重新审视了对抗性主题架构，提出了一种将判别目标重新表述为优化问题的新方法，并设计了一种有助于整合外部变量的新型采样方法。重新表述的这一新方法提出在模型中结合相似样本之间的关系，并对不同样本之间的相似性施加约束；而基于内部输入和重构输出的采样方法有助于为模型提供对主要主题有贡献的显着词。实验结果表明，在主题连贯性方面，我们的框架在属于不同领域、词汇量和文档长度的三个常见基准数据集中优于其他最先进的神经主题模型。

**关键词：**对比学习；神经主题模型

## 1 引言

主题模型已成功应用于自然语言处理领域，并具有各种应用，例如信息提取，文本聚类，摘要和情感分析[1-4]。最流行的常规主题模型，潜在狄利克雷分布（Latent Dirichlet Allocation，LDA）[5]，通过 Gibbs 采样和平均场近似来学习文档主题分布和主题词分布。为了将深度神经网络应用于主题模型，Miao 等人[6]提出使用神经变分推理作为训练方法，而 Srivastava 和 Sutton[7]采用逻辑正态先验分布。然而，最近的研究[8-9]表明，高斯和逻辑正态先验都无法捕获文档的多模态方面和语义模式，但文档的多模态方面和语义模式对于保持主题模型的质量至关重要。

为了解决这个问题，具有结合生成器和鉴别器的对抗机制的对抗主题模型 (ATM) 被提出了[8-9]。通过寻求生成器和鉴别器之间的平衡，生成器能够学习到文档的有意义的语义模式。尽管如此，这个框架有两个主要的局限性。首先，对抗主题模型 (ATM) 依赖于关键要素: 利用真实分配与假 (负) 分配的区别来指导培训。由于假分布的采样不以真实分布为条件，因此它几乎不产生正样本，而正样本在很大程度上保留了真实样本的语义内容。这限制了正样本与真实样本中的共有信息的表现，这一表现已被证明是在无监督学习中学习有用表示的关键驱动因素。其次，对抗主题模型 (ATM) 从先前的分配中获取随机样本注入到生成器。先前的工作[10]表明，合并其他变量 (例如元数据或情感) 来估计主题分布有助于学习连贯的主题。对抗主题模型 (ATM) 依靠预先定义的先验分布，因此它并不能很好地结合这些变量。

为了解决上述缺点，在本文中，我们提出了一种新颖的方法来建模样本之间的关系，而不依赖于生成判别体系结构。特别是，我们将目标表述为一个优化问题，旨在将输入 (或原型) 的表示形式更接近共享语义内容的表示形式，即正样本。我们也考虑了原型和负数样本的关系，通过形成辅助约束来使模型将负样本的表示与原型表示最大可能的分开。我们的数学框架以对比目标结尾，该目标将与神经主题模型的证据下限共同优化。

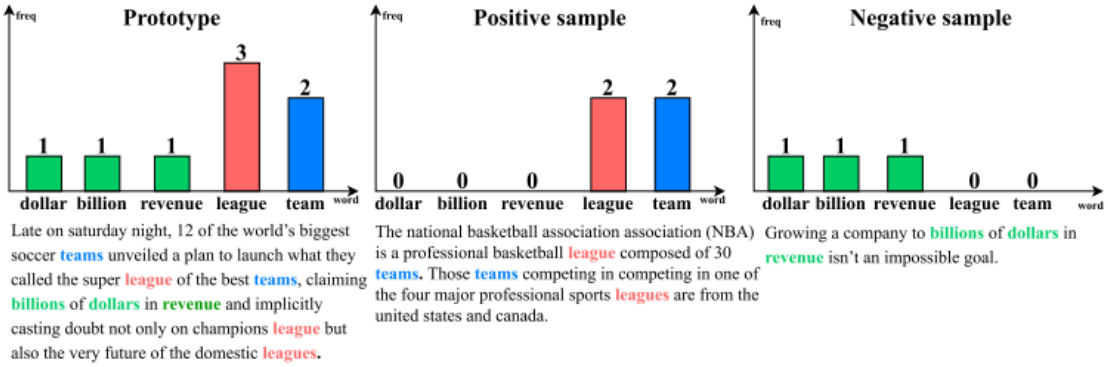


图 1: Illustration of a document with one positive and negative pair.

尽管如此，另一个挑战还是出现了: 如何在神经主题模型下有效地生成正负样本？最近的努力已经解决了正向采样策略和方法，以生成图像的硬负样本[11]。但是，文献中忽略了使这些技术适应于神经主题模型的相关研究。在这项工作中，我们介绍了一种新颖的采样方法，该方法模仿了人类抓住一对文档相似性的方式，该方法基于以下假设:

假设 1. 根据其显着单词的相对频率，可以表现出原型和正样本的共同主题。

我们使用图1中的示例来解释我们方法的思想。由于“league（联盟）”和“teams（团队）”等突出词的频率与正样本中的对应词成比例，因此人类能够分辨出输入与正样本的相似性。另一方面，输入和负样本之间的分离是可以诱导的，因为输入中的那些词不会出现在负样本中，尽管它们都包含“billions（十亿）”和“dollars（美元）”，这些并不是在输入的上下文中的突出词。基于这种直觉，我们的方法通过保持原型中显著条目的权重和改变不重要条目的权重来生成主题模型的正样本，同时对负样本执行相反的过程。本质上，由于我们的方法不依赖于固定的先验分布来绘制我们的样本，因此我们不限制将外部变量合并为更好地学习主题提供额外的知识。

简而言之，我们从新的数学角度出发，通过对样本之间的关系进行建模来捕获有意义的表示，并提出了一种新的对比目标，该目标与神经主题模型的证据下限共同优化。我们发现，捕获原型与其正样本之间的共有信息为构建连贯的主题提供了坚实的基础，相比之下区分原型和负样本的作用就不那么重要了。

我们提出了一种新颖的采样策略，该策略在比较不同的文档时受到人类行为的激励。通过依靠重构的输出，我们使采样适应模型的学习过程，与其他抽样策略相比，产生了信息最多的样本。

我们在三个常见的主题建模数据集中进行了广泛的实验，并通过在全局和逐个主题的基础上在主题一致性方面优于其他最新方法来证明我们的方法的有效性。

## 2 相关工作

### 2.1 Neural Topic Model (NTM)

神经主题模型 (NTM) 已被研究以使用潜在向量对大量文档进行编码。受变分自动编码器的启发，NTM 继承了 VAE 特定早期作品的大部分技术，例如重新参数化技巧[12]和神经变分推理[13]。尝试应用主题模型的后续工作[6-7,14]专注于研究各种先验分布，例如高斯或逻辑正态分布。最近，研究直接针对通过将主题连贯性制定为优化目标[15]、结合上下文语言知识[16] 或传递外部信息（例如情感，文档组，作为输入[10]。生成人类可解释的主题已成为各种最新努力的目标。

### 2.2 Adversarial Topic Model（ATM）

对抗主题模型（ATM）[8]是一种主题建模方法，它使用基于 GAN 的架构对主题进行建模。该架构中的关键组件包括一个生成器，该生成器投影随机采样的文档主题分布以尽可能获得最真实的文档词分布，以及一个试图区分生成样本和真实样本的鉴别器[8-9]。

### 2.3 Contrastive Framework and Sampling Techniques

有各种努力研究对比方法来学习有意义的表征。对于视觉信息，对比框架应用于图像分类[17]、目标检测[18]、图像分割[19]等任务。其他不同于图像的应用包括对抗训练[20]、图[21]和序列建模[22]。已经提出了特定的正采样策略来提高对比学习的性能，例如应用基于视图的转换来保留图像中的语义内容。另一方面，最近人们对研究负采样方法的兴趣激增。Chuang 等人[23]提出了一种去偏方法，用于纠正假负样本中的事实。尽管得到广泛研究，但很少有人努力将对比技术应用于神经主题模型。

## 3 本文方法

### 3.1 本文方法概述

在本文中，我们将对比框架应用于神经主题模型，将神经主题模型中学习文档表示的目标重新表述为对比目标。目标的形式主要与 Robinson[11]等人有关。然而，有两个关键的区别：（1）由于他们使用与负样本影响相关的权重因子作为搜索硬负样本分布的工具，我们将其视为自适应参数来控制学习的正负样本的影响。(2) 我们将正样本的影响视为实现有意义表示的主要驱动力，而它们利用负样本的影响。我们的方法更适用于主题建模，正如对区分文档的人类行为的调查所证明的那样。

### 3.2 神经主题模型部分

在本文中，我们专注于提高神经主题模型 (NTM) 的性能，通过主题连贯性来衡量。NTM 继承了 Variational Auto-encoder（VAE）的架构，其中将潜在向量作为主题分布。假设词汇表有 $V$ 个唯一单词，每个文档表示为单词计数向量 $x \in R^V$ 和 $T$ 个主题上的潜在分布：$z \in R^T$。神经主题模型假设 $z$ 是从先验分布 $p(z)$ 中生成的，$x$ 是解码器 $\phi$ 从主题 $p_\phi(x|z)$ 的条件分布生成的。模型的目的是在给定字数的情况下推断文档主题分布。换句话说，它必须估计后验分布 $p(z|x)$，它由编码器 $\theta$ 建模的变分分布 $q_\theta(z|x)$ 近似。NTM 通过最小化以下目标进行训练

$$\mathrm{L_{VAE}}(x) = -E_{q_\theta(z|x)}[\log p_\phi(x|z)] + KL[q_\theta(z|x)||p(z)] \tag{1}$$

### 3.3 对比学习部分

令 $\chi = \{x\}$ 表示文档词袋的集合。每个向量 $x$ 都与一个负样本 $x^-$ 和一个正样本 $x^+$ 相关联。我们假设一组离散的潜在类别 $C$，因此 $(x, x^+)$ 具有相同的潜在类别，而 $(x, x^-)$ 则没有。在这项工作中，我们选择使用语义点积来衡量原型 x 与绘制样本之间的相似性。我们的目标是学习映射函数 $f_\theta(R^V \rightarrow R^T)$：可以将 $x$ 变换为潜在分布 $z$（$x^-$ 和 $x^+$ 分别变换为 $z^-$ 和 $z^+$）的编码器 $\theta$。一个合理的映射函数必须满足两个特性：(1) $x$ 和 $x^+$ 被映射到附近的位置；(2) $x$ 和 $x^-$ 被映射到较远的位置。将目标（1）作为主要目标，将目标（2）作为强制模型学习不同样本之间关系的约束，我们指定约束优化问题，其中 $\varepsilon$ 表示约束的强度

$$\max_\theta E_{x\sim\chi}(z \cdot z^+) \quad subject \quad to \quad E_{x\sim\chi}(z \cdot z^-) < \varepsilon \tag{2}$$

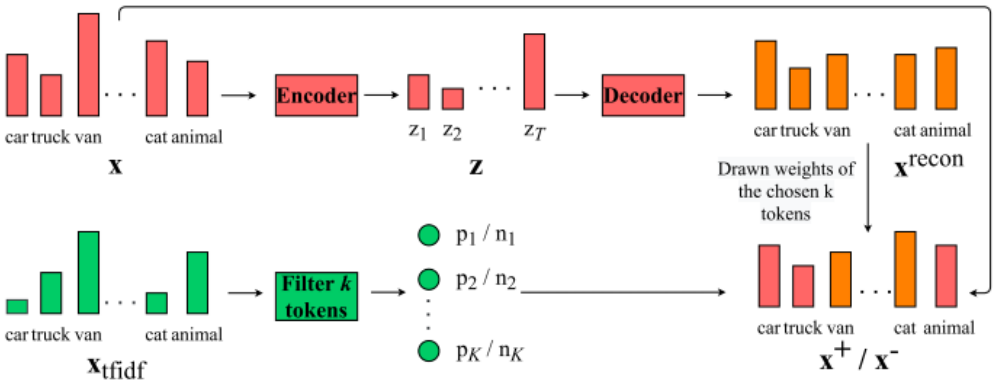


图 2: Word-Based Sampling Strategy

根据 KKT 条件下的 Lagragian[24-25]重写公式2，我们获得：

$$F(\theta, x, x^+, x^-) = E_{x\sim\chi}(z \cdot z^+) - \alpha \cdot [E_{x\sim\chi}(z \cdot z^-) < \varepsilon] \tag{3}$$

其中 KKT 乘数 $\alpha(\alpha > 0)$ 是控制负样本对训练效果的正则化系数。公式3可以推导得出加权对比损失：

$$F(\theta, x, x^+, x^-) \geqslant L_{cont}(\theta, x, x^+, x^-) = E_{x\sim\chi}[\log \frac{\exp(z \cdot z^+)}{\exp(z \cdot z^+) + \beta \cdot \exp(z \cdot z^-)}] \tag{4}$$

其中 $\alpha = exp(\beta)$。之前的工作[17,21,23] 认为正样本和负样本的可能性相同，所以设置 $\beta = 1$。在本文中，我们利用不同的 $\beta$ 值来指导模型集中在不同的样本上的输入。因此，合理的 $\beta$ 值将在数据集中的主题之间提供清晰的分离。

### 3.4 基于词的采样策略

在这里，我们提供了采样方法的技术动机和细节。为了选择与输入具有相同基础主题的样本，过滤掉在文档-主题分布中具有较大值的 M 个主题是合理的，因为它们被神经主题模型认为是重要的。随后，该程序将在每个主题中绘制显着词，这些词将为绘制的样本贡献权重。我们称这种策略为基于主题的采样策略。

然而，如[6]所示，主题选择过程对训练性能很敏感，并且确定每个输入代表的最佳主题数具有挑战性。Miao 等人[6]实施了一个 stick breaking 程序来专门预测每个文档的主题数量。他们的策略要求近似于增加文档表示的主题数量。由于他们的过程占用了大量的计算，我们提出了一种更简单的方法，即基于词来绘制正样本和负样本。

对于每个文档及其关联的词数向量 $x \sim \chi$，我们形成 $tf - idf$ 表示 $x^{tf-idf}$。然后，我们将 $x$ 馈送到神经主题模型以获得潜在向量 $z$ 和重构文档 $x^r econ$。我们基于词的采样策略如图2所示。

负采样我们选择 $k$ 个标记 $N = n_1, n_2, ..., n_k$ 具有最高的 $tf - idf$ 分数。我们假设这些词主要有助于表达文档的主题。通过将原始输入 $x$ 中所选标记的权重替换为重构表示 $x^r econ$ 的权重：$x^-_{n_j} = x^{recon}_{n_j}, j \in \{1, .., k\}$，我们强制负样本 $x^-$ 的主要内容偏离原始输入 $x$。请注意，由于模型随着训练的进行而提高了其重建能力，因此重建输出中显着词的权重接近原始输入中的显着词权重（但不相等）。该模型应该采取更仔细的学习步骤来适应这种情况。由于负样本控制因子 $\beta$ 在收敛到最后的训练步骤时会衰减其值，由于我们在 3.3 节中提到的自适应调度方法，它能够适应这种现象。

正采样与负采样相反，我们选择 $k$ 个具有最低 $tf - idf$ 分数的标记 $P = p_1, p_2, ..., p_k$。我们通过 $x^-_{p_j} = x^{recon}_{p_j}, j \in \{1, .., k\}$ 将 $x^{recon}$ 中所选标记的权重分配给 $x^+$ 中的对应标记，从而获得与原始输入具有相似主题的正样本。这形成了一个有效的正采样程序，因为修改了不重要标记的权重保留了源文

档中的显着主题。

### 3.5 训练目标

我们将重建原始输入的目标与实际后分布相匹配，与第3.3节中指定的对比目标相匹配。

$$L(x,\theta,\phi) = -E_{z\sim q(z|x)}[\log(p_\theta(x|z)) + KL(q_\theta(z|x)||p(z))] - E_{z\sim q(z|x)}[\log \frac{\exp(z\cdot z^+)}{\exp(z\cdot z^+) + \beta\cdot\exp(z\cdot z^-)}] \quad (5)$$

## 4 复现细节

### 4.1 与已有开源代码对比

复现过程中引用了论文的官方代码，代码地址为https://github.com/nguyentthong/CLNTM。在分析论文代码时，我发现作者为了计算的速度而使用了简化方法，直接令 $x^{recon} = x^{tf-idf}$，通过计算 $x^{recon}$ 的 $top-k$ 来生成正负样本，且使用的仅仅是生成了负样本的模型。为了与论文方法保持一致，我编写了计算每篇文本 $tf-idf$ 相关代码，并修改整体代码框架，使得模型能够使用对应文本的 $tf-idf$ 的 $top-k$ 来生成正负样本，并且使用能正常生成正负样本的完整的模型。

同时由于论文的代码是在 Scholar 的代码架构基础上进行的升级，因此我对 Scholar 模型也进行了复现，使得复现出来的模型可以与其他模型进行对比。Scholar 模型的官方代码地址为https://github.com/dallascard/scholar。

伪代码如下所示：

**Procedure 1** Contrastive Neural Topic Model

**Input:** Dataset $D = \{x^i_{tfidf}, x^i_{BOW}\}_{i=1}^N$, model parameter $\theta$, model $f$, push-pull balancing factor $\alpha$, contrastive controlling weight $\gamma$
**Output:** Topic-word distribution
Compute $x_{tfidf}$ from $x_{BOW}$
**for** *k=1 to maxepochs* **do**
  **for** $i = 1$ *to* $N$ **do**
    Compute $z^i$, $x^i_{recon}$ from $x^i_{BOW}$;
    Obtain $top-k$ indices of words with smallest $tf-idf$ weights $K_{pos} = p_1, p_2, ..., p_k$;
    Sample $x^i_{pos}$ from $K^i_{pos}$ and $x^i_{recon}$
    Obtain $top-k$ indices of words with largest $tf-idf$ weights $K_{neg} = n_1, n_2, ..., n_k$;
    Sample $x^i_{neg}$ from $K^i_{neg}$ and $x^i_{recon}$
  **end**
  Compute the loss function $L$ defined in Eq.5;
  Update $\theta$ by gradients to minimize the loss;
**end**

### 4.2 实验环境搭建

python3

pandas

gensim

numpy

torchvision

pytorch 1.7.0

scipy

表 1: 用 NPMI 衡量的神经主题模型的结果

| | **20NG** | | **IMDb** | | **Wiki** | |
|---|---|---|---|---|---|---|
| | T=50 | T=200 | T=50 | T=200 | T=50 | T=200 |
| Scholar | 0.319±0.005 | 0.274±0.003 | 0.160±0.002 | 0.172±0.001 | 0.274±0.015 | 0.362±0.008 |
| Original model k=1 | 0.324±0.005 | 0.274±0.002 | 0.158±0.005 | 0.174±0.002 | 0.487±0.012 | 0.452±0.004 |
| Original model k=5 | 0.322±0.002 | 0.277±0.002 | 0.166±0.007 | 0.175±0.003 | 0.473±0.010 | 0.450±0.005 |
| Original model k=15 | **0.326±0.004** | **0.277±0.001** | **0.167±0.004** | **0.176±0.004** | **0.454±0.023** | **0.435±0.005** |
| Updated model k=1 | **0.324±0.002** | **0.273±0.002** | **0.167±0.005** | **0.170±0.002** | **0.476±0.012** | **0.450±0.003** |
| Updated model k=5 | 0.323±0.005 | 0.268±0.001 | 0.162±0.004 | 0.160±0.001 | 0.442±0.015 | 0.453±0.006 |
| Updated model k=15 | 0.315±0.008 | 0.265±0.003 | 0.162±0.004 | 0.153±0.002 | 0.407±0.011 | 0.425±0.008 |

### 4.3 界面分析与使用说明

训练模型：runModel.sh

执行评估程序：evaluateModel.sh

### 4.4 创新点

由于论文的官方代码中使用的是便于计算的简化版本的模型，但简便计算的方法与原论文方法终究有所不同，使用便于计算的方法所获得的性能并不能说明方法的有效性。因此我完全按照原论文中的方法复现出整个模型，通过实验去验证论文提出方法的有效性。

## 5 实验结果分析

通过对三个不同数据集（20ng,IMDb,wiki）进行操作，然后用主题连贯性（topic coherence）这一评价指标的高低来验证方法的有效性。通过计算每个数据集所求得的主题-词分布的 NPMI 值来评价主题连贯性，若 NPMI 值较高则说明主题连贯性越好。

Scholar、原模型和更新后的模型性能对比如表1所示。原模型与 Scholar 模型相比，在主题连贯性方面确实有一定的提升，也确实如论文中的结论所说，当 $top-k$（即表中的 k 值）增大时，模型的主题连贯性会有稳步的提升。但是通过数据而言，更新后的模型结果（即论文中提出的方法）相比 Scholar 模型也有一定的提升，但是当 top-k 不断增大时，所得到的主题连贯性却有所下降。

## 6 总结与展望

### 6.1 总结

所得实验结果与论文中所得出的结论并不相同，正在探究产生当前情况的原因，其中可能的原因如下：1、计算 $tf-idf$ 的部分代码出错，导致模型使用到错误的数据；2、$tf-idf$ 这一指标并不能突出表明对主题贡献较大的词；3、论文中的结论不够严谨，便于计算的方法并不会与提出的方法有相似的结果。进一步的探究还在进行中。

### 6.2 展望

在这篇论文中提出了一种对比学习框架、一种自适应的调度策略调整正负样本的比重和一种新的基于词的采样策略，以提高神经主题模型在主题连贯性方面的性能。方法十分新颖，并且可以解决对抗主题模型的主要缺点。但还有几方面可以有更深入的探索，例如 1、使用一种比 $tf-idf$ 更好的评价指标去生成正负样本；2、在对比学习的应用中，出现过只生成正样本，不生成负样本的模型同样提高了整体的评价指标，不知在这个模型中是否适用；3、在基础的 VAE 模型中加入对比学习框架会

对其有多大的提升。这几点我会在之后进一步探索。

## 参考文献


[1] NGUYEN T LUU A T. Contrastive learning for neural topic model. Advances in Neural Information Processing Systems, 2021, 34: 11974-11986.

[2] SUBRAMANI S, SRIDHAR V, SHETTY K. A novel approach of neural topic modelling for document clustering//2018 IEEE Symposium Series on Computational Intelligence (SSCI). 2018: 2169-2173.

[3] WANG R, ZHOU D, HE Y. Open event extraction from online text using a generative adversarial network. arXiv preprint arXiv:1908.09246, 2019.

[4] WANG M MENGONI P. How pandemic spread in news: text analysis using topic model//2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT). 2020: 764-770.

[5] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation. Journal of machine Learning research, 2003, 3(Jan): 993-1022.

[6] MIAO Y, GREFENSTETTE E, BLUNSOM P. Discovering discrete latent topics with neural variational inference//International Conference on Machine Learning. 2017: 2410-2419.

[7] SRIVASTAVA A SUTTON C. Autoencoding variational inference for topic models. arXiv preprint arXiv:1703.01488, 2017.

[8] WANG R, ZHOU D, HE Y. Atm: Adversarial-neural topic model. Information Processing & Management, 2019, 56(6): 102098.

[9] WANG R, HU X, ZHOU D, Neural topic modeling with bidirectional adversarial training. arXiv preprint arXiv:2004.12331, 2020.

[10] CARD D, TAN C, SMITH N A. Neural models for documents with metadata. arXiv preprint arXiv:1705.09296, 2017.

[11] ROBINSON J, CHUANG C Y, SRA S, Contrastive learning with hard negative samples. arXiv preprint arXiv:2010.04592, 2020.

[12] KINGMA D P WELLING M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[13] REZENDE D J, MOHAMED S, WIERSTRA D. Stochastic backpropagation and approximate inference in deep generative models//International conference on machine learning. 2014: 1278-1286.

[14] MIAO Y, YU L, BLUNSOM P. Neural variational inference for text processing//International conference on machine learning. 2016: 1727-1736.

[15] DING R, NALLAPATI R, XIANG B. Coherence-aware neural topic modeling. arXiv preprint arXiv:1809.02687, 2018.

[16] HOYLE A, GOEL P, RESNIK P. Improving neural topic models using knowledge distillation. arXiv preprint arXiv:2010.02377, 2020.

[17] KHOSLA P, TETERWAK P, WANG C, Supervised contrastive learning. Advances in Neural Information Processing Systems, 2020, 33: 18661-18673.

[18] XIE E, DING J, WANG W, Detco: Unsupervised contrastive learning for object detection/ /Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 8392-8401.

[19] ZHAO X, VEMULAPALLI R, MANSFIELD P A, Contrastive learning for label efficient semantic segmentation/ /Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 10623-10633.

[20] HO C H NVASCONCELOS N. Contrastive learning with adversarial examples. Advances in Neural Information Processing Systems, 2020, 33: 17081-17093.

[21] YOU Y, CHEN T, SUI Y, Graph contrastive learning with augmentations. Advances in Neural Information Processing Systems, 2020, 33: 5812-5823.

[22] LOGESWARAN L LEE H. An efficient framework for learning sentence representations. arXiv preprint arXiv:1803.02893, 2018.

[23] CHUANG C Y, ROBINSON J, LIN Y C, Debiased contrastive learning. Advances in neural information processing systems, 2020, 33: 8765-8775.

[24] BERTSEKAS D P. Nonlinear programming. Journal of the Operational Research Society, 1997, 48(3): 334-334.

[25] KARUSH W. Minima of functions of several variables with inequalities as side constraints. M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago, 1939.