A Primer on Contrastive Pretraining in Language Processing: Methods, Lessons Learned, and Perspectives

NILS RETHMEIER, German Research Center for AI, Berlin, Germany, University of Copenhagen, Denmark
ISABELLE AUGENSTEIN, University of Copenhagen, Denmark, Denmark

Modern natural language processing (NLP) methods employ self-supervised pretraining objectives such as masked language modeling to boost the performance of various downstream tasks. These pretraining methods are frequently extended with recurrence, adversarial, or linguistic property masking. Recently, contrastive self-supervised training objectives have enabled successes in image representation pretraining by learning to contrast input-input pairs of augmented images as either similar or dissimilar. In NLP however, a single token augmentation can invert the meaning of a sentence during input-input contrastive learning, which led to input-output contrastive approaches that avoid the issue by instead contrasting over input-label pairs. In this primer, we summarize recent self-supervised and supervised contrastive NLP pretraining methods and describe where they are used to improve language modeling, zero to few-shot learning, pretraining data-efficiency, and specific NLP tasks. We overview key contrastive learning concepts with lessons learned from prior research and structure works by applications. Finally, we point to open challenges and future directions for contrastive NLP to encourage bringing contrastive NLP pretraining closer to recent successes in image representation pretraining.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Machine learning**; **Transfer learning**; **Neural networks;**

Additional Key Words and Phrases: Contrastive learning

ACM Reference format:
Nils Rethmeier and Isabelle Augenstein. 2023. A Primer on Contrastive Pretraining in Language Processing: Methods, Lessons Learned, and Perspectives. *ACM Comput. Surv.* 55, 10, Article 203 (February 2023), 17 pages.
https://doi.org/10.1145/3561970

1 INTRODUCTION

Current downstream machine learning applications heavily rely on the effective pretraining of representation learning models. Contrastive learning is one such technique that enables pretraining of general or task-specific encoder models in a supervised or self-supervised fashion. While contrastive pretraining in computer vision has enabled the recent successes in self-supervised image representation pretraining, the benefits and best practices of contrastive pretraining in **natural**

Authors' addresses: N. Rethmeier, Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Alt-Moabit 91c, 10559 Berlin, Germany; email: nils.rethmeier@dfki.de; I. Augenstein (corresponding author), Department of Computer Science, University of Copenhagen, Universitetsparken 1, 2100 Copenhagen, Denmark; email: augenstein@di.ku.dk.

0360-0300/2023/02-ART203 $15.00
https://doi.org/10.1145/3561970

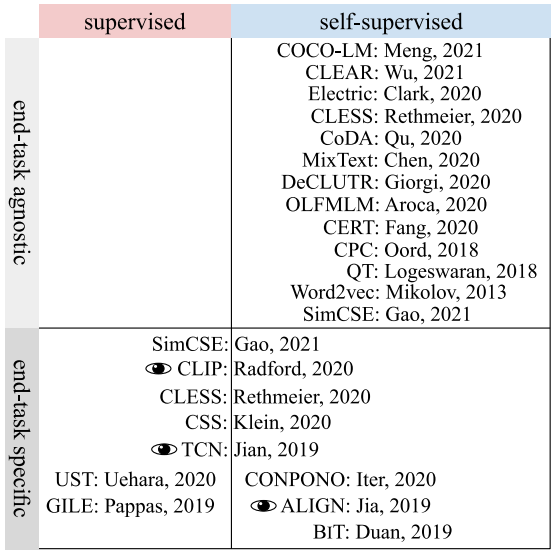

Fig. 1. Types of contrastive pretraining and works that fall within these categories. 👁 marks text-image contrastive works.

language processing (NLP) are less established [24]. However, a first wave of works on contrastive NLP methods, seen in Figure 1, shows strong performance and data-efficiency benefits of (self-)supervised contrastive NLP pretraining. For example, even supervised contrastive pretraining enables zero-shot prediction of unseen text classes and improves few-shot performance [47]. Moreover, task-agnostic self-supervised contrastive pretraining systems have been shown to improve overall language modeling performance [9, 15, 38, 64], data and label-efficiency [49, 53], or semantic similarity tasks [14]. Besides that, there are many task-specific uses of contrastive self-supervision, e.g., for pronoun disambiguation [31], discourse representation learning [23], or text summarization [12], to name a few—see Section 3.

Contributions. In this primer to contrastive pretraining, we therefore summarize recent supervised and self-supervised contrastive NLP pretraining methods. We then describe how they enable zero-shot learning and improve language modeling, few-shot learning, pretraining data-efficiency or rare event prediction. We cover basic concepts and crucial design lessons of contrastive NLP, while detailing the resulting benefits such as zero-shot prediction and efficient training. Then, we structure existing research as supervised or self-supervised contrastive pretraining and explain connections to **Energy-based models (EBMs)**, since many works refer to EBMs. Finally, we point out open challenges and outline future and underrepresented research directions in contrastive NLP pretraining.

2 CONTRASTIVE LEARNING AND ITS BENEFITS

At their core, contrastive methods learn to distinguish between pairs of similar or dissimilar data points. A pair of similar data points is called a positive sample, which in self-supervised contrastive learning is generated by augmenting an original data point. For example, SimCSE [14] applies two dropout masks to an input sentence to create two slightly different sentence embeddings that are then used as a pair of positive (matching) sentence embeddings for self-supervised pretraining. Negative samples are pairs where the two data points are of different data instances, e.g., in

SimCSE the authors simply use the embeddings of other sentences in a training batch as negatives. Contrastive objectives have been demonstrated to have certain desirable properties over other common losses. Reference [16] has shown that a contrastive loss is more resistant to label noise than the commonly used softmax objective. Additionally, Zimmermann et al. [66] demonstrate that contrastive self-supervision effectively "inverts a data generating process." This results in very data-efficient pretraining as both they and Rethmeier and Augenstein [53] demonstrate. Other potential benefits include modelling redundancy reduction [65] and disentangling representations [52]. Below, in Section 2.1, we overview popular contrastive losses for NLP and summarize how to avoid pitfalls. Then, we overview connections to other machine learning fields and specifically outline Energy-based models in Section 2.2, since they are used in much of the cited research. Finally, we organize methods into input-input contrastive and the NLP specific input-output contrastive methods to highlight their respective benefits.

2.1 Noise Contrastive Estimation (NCE)

Noise contrastive estimation (NCE) is the objective used by most contrastive learning approaches within NLP. Thus, we briefly outline its main variants, Binary and Ranking NCE, and the core ideas behind them, while pointing to Ma and Collins [40][1] for detailed, yet readily understandable explanations of the two main NCE variants. Both variants can intuitively be understood as classification with undersampling of negative (non-active) classes. Either method is used to predict a similarity score between two text embeddings, where the prediction score 1 means similar, and a score of 0 or -1 means dissimilar, i.e., a direct analog to the standard (self-)supervised classification objectives. Training is done with positive and negative samples, where x_i is an original text input embedding, while a_i^- is an augmented text embedding that is dissimilar to x_i, and a_i^+ is a text embedding that is considered similar to x_i. A positive sample is then describes as a pair of similar text embeddings $<x_i, a_i^+>$ that is annotated, manually or automatically, with a similarity (class) of 1 via an indicator variable. A negative sample is a pair $<x_i, a_i^->$ of dissimilar text embeddings that is annotated with a similarity of 0, or -1, depending of the similarity function (e.g., cosine) to indicated dissimilarity or a mismatch. Either method uses one positive sample $<x_i, a_i^+>$, and sub-samples K negative samples $<x_i, a_i^->$ for contrast. Below, we describe both variants and will point out an easy way to remember both variants at the end.

Binary NCE:. The first variant expresses NCE as a binary objective (loss) in the form of maximum log likelihood, where only K negatives are considered:

$$L_B(\theta, \gamma) = log\, \sigma(s(x_i, a_{i,0}^+; \theta), \gamma) + \sum_{k=1}^{K} log(1 - \sigma(s(x_i, a_{i,k}^-; \theta), \gamma). \quad (1)$$

Here, $s(x_i, a_{i,\circ}; \theta)$ is a similarity or scoring function that measures the compatibility between a single text input x_i and a contrast sample $a_{i,\circ}$. This sample is another input text or an output label (text) to model NLP tasks as "text-to-text" prediction similar to language models. The similarity or scoring function is typically a cosine similarity, a dot product, or a small neural network that computes a similarity or matching score between a pair of text embeddings [47, 53]. The $\sigma(z, \gamma)$ is a scaling function, which for use in Equation (1) is typically the sigmoid $\sigma(z) = \exp(z - \gamma)/(1 + \exp(z - \gamma))$ with a hyperparameter $\gamma \geq 0$ (temperature), that is tuned or omitted depending on how negative samples a_i^- are attained [40].

[1] https://vimeo.com/306156327 talk by Ma and Collins [40].

Ranking NCE, InfoNCE, NT-Xent: learns to rank a single positive pair $(x_i, a^+_{i,0})$ above K negative pairs $(x_i, a^-_{i,k})$:

$$L_R(\theta) = log \frac{e^{\bar{s}(x_i, a^+_{i,0}; \theta)}}{e^{\bar{s}(x_i, a^+_{i,0}; \theta)} + \sum_{k=1}^{K} e^{\bar{s}(x_i, a^-_{i,k}; \theta)}}. \tag{2}$$

In Jaiswal et al. [25], Section 5, it can be seen that the Ranking NCE objective has the same form as the InfoNCE objective in CPC [62] or the NT-Xent objective in SimCLR [6], except that the SimCLR version uses a similarity scaling factor (temperature) τ. The names InfoNCE and NT-Xent are commonly used in computer vision, while all names are used in NLP. Interestingly, van den Oord et al. [62] also proved that "minimizing this loss maximizes a lower bound on the mutual information" of a positive sample (between a pair) $<x_i, a^+_{i,0}>$—i.e., in their notation between $x_{t+k}, c_t := <a^+_{i,0}, x_i>$. This means that the Ranking NCE can also be understood as approximate mutual information maximization objective.

Additionally, as discussed in Ma and Collins [40] Section 3.2, some older works define a modified similarity (scoring) function $\bar{s}(x_i, a_{i,\circ}) = s(x_i, a_{i,\circ}) - log\, p_{\mathcal{N}}(a_{i,\circ})$ to subtract the probability of the current sample $a_{i,\circ}$ under a chosen noise distribution $p_{\mathcal{N}}(a_{i,\circ})$. For example, Mikolov et al. [43] set $p_{\mathcal{N}}(a_{i,\circ})$ as corpus word-unigram probabilities p_{word} [43] to make the learning of word embeddings more robust. Works like Deng et al. [11] set the noise distribution to the probability p_{LM} of a sequence under a language model LM, to learn contrastive sequence prediction. As Ma and Collins [40] state, when desirable, adding this noise term can make the Binary NCE objective self-normalizing allowing it to converge faster toward the MLE solution. Ranking NCE is already self-normalizing, but Ma and Collins [40] showed that adding the noise term can still improve RankingNCE results. While some older works like Reference [44] set the parameters of the noise distribution term to zero, for computational reasons, recent models do not use a noise distribution term [53, 64].

Generalization to an Arbitrary Number of Positives. As Khosla et al. [30] discuss original contrastive losses use only one positive sample per text instance (see, e.g., Logeswaran and Lee [38], Mikolov et al. [43]), while recent methods mine multiple positives per sample [48, 53]. This means that the positive term in Equation (1) extends to become a sum over P positive samples:

$$\sum_{p=1}^{P} log\, \sigma(s(x_i, a^+_{i,p}; \theta, \gamma)). \tag{3}$$

A way to easily remember Binary and Ranking NCE. Binary NCE can be remembered as an undersampled version of Binary Cross Entropy over similarity scores. When used with multiple positives, but *without* a noise distribution term for self-normalization, it can learn multi-label problems, where an instance can have multiple active labels (classes) that are independent of each other. When using a single positive class, Binary NCE learns a contrastive version of multi-class classification, especially when adding the optional noise distribution term for self-normalization, which makes it act even closer to an undersampled softmax objective.

Ranking NCE can be remembered as an undersampled softmax over similarity scores, as it uses an undersampled normalization, which suits multi-class learning, where classes are mutually exclusive and normalization can be thought of as 'inducing a ranking and mutual exclusivity' between classes. While this objective if often appropriate when learning representations, other ranking losses can be used to induce ranking oriented semantics, e.g., SPECTER and SciNCL [10, 46] use the triplet (ranking) loss for contrastive pretraining of citation representations. However, for practical applications the impact of ranking loss variant may be minor, as pointed out by Musgrave

et al. [45], who give a concise overview of relevant losses and a critical analysis of realistic benefits and drawbacks.

Lessons on Effective Negative and Positive Sampling. A key component (and pitfall) of effective contrastive learning is how positive and negative samples are generated. Saunshi et al. [55] prove and empirically validate that "sampling more negatives improves performance, *but only if they do not collide with positive samples*," which otherwise deteriorates performance. Instead, in Section 6.3 of their work they propose to "sample negatives from blocks of similar data points," i.e., from similar contexts such as the same paragraph or sentence. Instances of such contextual contrast sampling can be found in References [23, 53, 55]. For example, Rethmeier and Augenstein [53] sample words from a current text instance to construct positives for self-supervised pretraining of a contrastive text autoencoder model. Recent works use multiple positive samples to boost supervised contrast [30]. Additionally, during self-supervision, multiple positives should be sampled from similar contexts [63] or "diversely from *common and rare* positives when pretraining long-tail recognition language models" [53].

2.2 Contrastive Learning in Machine Learning

Contrastive learning methods are related to other machine learning concepts, all of which describe the same underlying intuitions of how to learn representations in either a supervised or self-supervised fashion. All these methods are related in that they are learning from the similarity (contrastive, metric learning), shared information (mutual information), or compatibility (Energy-based Models) between views or augmentations of the same input or output data.

Mutual information. For one, as explained in Section 2.1, InfoNCE (or Ranking NCE) has been shown to maximize the lower bound of **mutual information (MI)** between similar augmented and non-augmented inputs [21, 62], while works like References [3, 32] show this MI perspective between inputs and labels or inputs and input sub-sequences. Importantly, Tschannen et al. [60] demonstrate that maximizing mutual information with contrastive losses can deteriorate end-task performance, and does not necessarily lead to learning useful representations. They also state that the "mutual information gets biased by the end-task objective, such that end task performance is maximized." In contrast, when thinking of the recent successes with contrastive pretraining in computer vision [6, 7], it becomes apparent that the bias introduced via the sampling design, i.e., the input augmentations used to produce positive and negative samples, is the tool for introducing task biases that lead to learning relevant representations. For this reason, starting from Section 3, we will discuss contrastive works in the context of NLP subfields to provide pointers to what kinds of sampling and augmentations induce desirable biases that help a model learn NLP task-relevant representations.

Metric learning. Because contrastive methods learn classification over similarity scores between instances, it is a part of the more general field of metric learning. Metric learning uses losses like triplet loss, focal loss, Neighborhood Component Analysis, and many others to learn (dis-)similarities between inputs [45]. Interestingly, Musgrave et al. [45] find that even basic contrastive losses perform surprisingly well when fairly compared to advanced metric losses, while Zimmermann et al. [66] point out that contrastive objectives are "theoretically more deeply understood than most metric losses."

Energy-based Models. Many recent works describe contrastive learning as EBMs. LeCun et al. [33], LeCun and Huang [34] define an Energy-based Model $E(W, X, Y)$ as one that "instead of trying to classify inputs X to labels Y, we would like to predict if a certain pair of $<x, y>$ fit together or not under the model parameters W—i.e., find whether a y is compatible with x according to W.

Especially LeCun and Huang [34] describes how an EBM E(X,Y,W) can be expressed in probabilistic model notation $P(Y|X, W)$ or as a non-normalized model, as we have already seen in Section 2.1 with Ranking NCE and Binary NCE. The modernized graphical and mathematical notion, as used in the excellent EBM lecture by LeCun and Canziani,[2] still heavily relates to the ones used in References [33, 34].

Therefore, we overview the two most NLP-relevant EBM formulations and reuse not only their mathematical notation but also adopt the graphical notation from LeCun et al. [33] for figures. In keeping with this notion, we categorize methods as either input-input (x_i, x_j) or input-output (x_i, y_j), which also allows us to better discuss their respective benefits. Contrastive computer vision methods learn from input-input (image-image) pairs (x_i, x_j) [7, 24]. As a result, using 10 samples incurs 10 times the computational load, which spawned efforts to reduce this load by reusing computation [20]. In NLP, some recent methods reduce this burden by instead using input-output (text, label) pairs (x_i, y_c), where the labels are produced by a separate, very lightweight, encoder—see Section 2.2. Here x_i, x_j are input text embeddings, while y_c are embeddings of "a short text span that describes a real or self-supervision label," i.e., an extreme summarization. This works as follows.

Input-output Contrastive EBM. The binary NCE variant from Equation (1) is a special case of a "Contrastive Free Energy" loss as described in Figure 6(b) of LeCun et al. [33], while Figure 2 and Section 3.3 of LeCun and Huang [35] describe it as the negative log-likelihood loss with negative undersampling. LeCun et al. [33] devise an input-output EBM E variation that learns the compatibility between input-output pairs (x_i, y_c) with $x_i \in X$ and $y_c \in Y$,

$$E(X, Y) \text{ or } E(W, X, Y). \tag{4}$$

Here, W (θ in Equation (1)) are model parameters that encode inputs X and labels Y, while these X and Y are views of either the same data point (positives) or different data points (negatives). The energy function E measures the compatibility between views (X, Y), where $E(\circ) = 0$ indicates optimal compatibility—e.g., $E(X = Tiger, Y = felidae)$=0 means X and Y match. Note that in the probabilistic framework $P(Y = felidae|X = Tiger, W) = 1$. Figure 2 shows two recent works [47, 53] that use an input-output contrastive learning approach. These methods encode an input text x_i using a text-encoder T and a label description text y_c using a very small, computationally cheap, label-encoder L. The input text and label text encoding is then concatenated into a single text input-output encoding pair $(T(x_i), L(y_c))$, which feeds a classifier that trains a binary NCE objective L_B, as in Equation (1). The left method in Figure 2 by Pappas and Henderson [47] uses *supervised text-to-label pair pretraining* to allow zero-shot prediction of unseen test time classes. The right-hand side method [53] instead samples input words $x_i \in X$ to use them as "pseudo label" encodings $y'_c = L(x_i)$ for *contrastive self-supervised pretraining*. Once this method pretrains on sampled input words (pseudo labels), the prediction head is directly reused during supervision via textual labels. This enables zero-shot prediction, *without using supervision labels*, by unifying supervision and self-supervision as a single task of learning to contrast (mis)matching (real) label encodings $L(y_c)$ or pseudo label encodings $y'_c = L(x_i)$.

Input-input Contrastive EBM. Expressing input-input contrastive learning as an EBM is straight forward [33]. Input-input methods in Figure 3 contrast input texts X from augmented input texts X' rather than from labels Y as in Figure 2. For example, Clark et al. [9] replace a subset of input text words $x_{i,w}$ with other words $x_{i,w'}$ sampled from the vocabulary for self-supervised contrastive pretraining. The original text x_i is augmented into a text a_i to provide a positive sample augment

[2] https://atcold.github.io/pytorch-Deep-Learning/en/week07/07-1/—EBM definition by Yann LeCun and Alfredo Canziani.

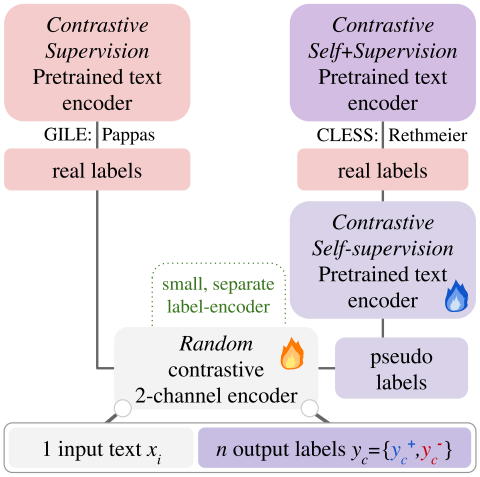

Fig. 2. Contrastive input-output (X, Y) pretraining: Texts and labels are encoded independently via a medium sized text encoder and a very small label-encoder. This encodes 1 text for n labels with minimal computation to enable large-scale K negative sampling.

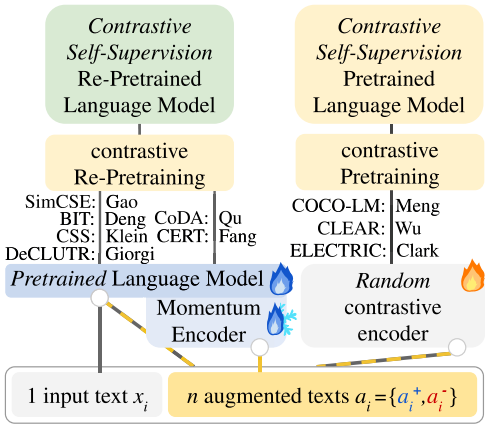

Fig. 3. Contrastive input-input (X, X') Pretraining: Input-input methods contrast an original text with augmented positive a_i^+ and negative a_i^- texts $a_i \in X'$, which requires more computation than input-output methods.

a_i^+ or a negative sample augment a_i^-. Self-supervised pretraining then contrasts pairs (x_i, a_i) of original texts against augmented ones via the binary NCE as in Equation (1). As a direct analog to the EBM in Equation (4), this can be written as

$$E(X, X') \text{ or } E(W, X, X'). \tag{5}$$

Methods on the left in Figure 3 re-pretrain an already pretrained language model such as BERT, using a contrastive objective. Methods on the right implement what amounts to contrastive (self-supervised) language model pretraining. Though the difference between input-input and input-output contrast seems semantical at first, each approach implies specific practical benefits and drawbacks as follows.

Input-input contrastive methods. Input-input contrastive methods such as the ones listed in Figure 3 recently explored improving large-scale NLP pretraining. Re-pretraining methods like Deng et al. [11], Fang et al. [13], Gao et al. [14], Giorgi et al. [15], Qu et al. [48] (see Figure 3 left tower) apply "a second stage of contrastive pretraining" to an otherwise pretrained Transformer model, to save computation by contrastively training with less input augmentations. This can be seen as either making use of (an advantage) or requiring (a limitation) otherwise pretrained models. Other input-input contrastive methods such as Clark et al. [9], Meng et al. [41], Wu et al. [64] *do not rely* on otherwise pretrained encoders—see Figure 3 right tower. Additionally, works such as Radford et al. [49] provide evidence that the same efficiency benefit of contrastive learners apply to large-scale models across modalities. For instance, Radford et al. [49] replace a Transformer by a CNN to speed up self-supervised zero-shot prediction learning by a factor of 3, and add text contrastive pretraining to speed up learning by another factor of 4.

The Benefits and Weaknesses of Input-output or Input-input Contrastive Methods. Input-output contrastive methods in Figure 2 are capable of zero-shot prediction, since they learn a pretraining NCE classifier, which can be reused or re-tuned to suit any downstream task labels, without having to initialize a new classifier per task—i.e., multi-task learning becomes single-task learning. As a contrastive analog to text-to-text Transformers like T5 [50], they unify self and supervised prediction with zero-shot transfer as a "text-to-text embedding similarity prediction" objective, whereas most, but not all, input-input methods, still have to initialize a new classifier for supervised downstream tasks. Input-output contrastive methods encode labels by using a small, compute efficient label-encoder, while encoding the more compute intensive text input X encoding only once. Input-input methods, however, construct "augmented views" X' by running a large (Transformer) text encoder over an augmented version X' of the original input X. This multiplies their training time by the number $K+P$ of negative and positive samples and presents the most important challenge to the more wide-spread adaptation of contrastive pretraining. For this reason, input-input research often argues that fewer negative samples should be used. Instead, input-output contrastive self-supervision [53] and contrastive supervision [19, 27, 47] enable very data-efficient pretraining and improved zero to few-shot, as well as long-tail learning. They can be pretrained on very small text collections with commodity hardware, which is very important for many applications in industry, medicine, and places where large amounts of GPU compute are less easy to attain. Thus, combining *highly expressive input-input with compute efficient input-output methods* provides a logical progression for future research.

Both input-input and input-output methods are well suited for self-supervised pretraining of language representations. These contrastively pretrained representations can subsequently be used for transfer learning of later supervised end-tasks during fine-tuning just as masked or autoregressive language model pretraining improves transfer. To make it easy to understand which contrastive methods have been explored within the different subfields of NLP, we overview popular self-supervised and supervised contrastive NLP methods Section 3 and later point to open questions and opportunities Section 4.

3 SELF- OR SUPERVISED CONTRASTIVE PRETRAINING

The goal of contrastive pretraining is to initialize model weights for efficient zero-shot transfer or fine-tuning to downstream tasks. Pretraining is either supervised or self-supervised. Supervised contrastive pretraining methods use corpora of hand-annotated data such as paraphrased parallel sentences, textual labels, or text summarizations to define text data augmentations for contrastive pretraining. Self-supervised contrastive methods aim to scale pretraining by contrasting automatically augmented input texts X' or textual output pseudo-labels $Y' \sim P(X)$—see Section 2.2

for input-input versus input-output contrastive methods. Both self-supervised and supervised contrastive methods are used to train language models from scratch, or can "re-pretrain" or fine-tune a model that was already pretrained using another pretraining method, e.g., a masked language model such as RoBERTa [37]. Below, we structure self- and supervised contrastive pretraining by technique and application.

3.1 Self-supervised Contrastive Pretraining

Input-input Contrastive Text Representation Pretraining via Automated Text Augmentation. Figure 3 compares methods that use input-input contrastive (EBM) learning as overviewed in Section 2.2. Qu et al. [48] use combinations of recently proposed text data augmentations like "cutoff, back translation, adversarial augmentation and mixup." They find that mixing augmentations is most useful when the augmentations provide sufficiently different views of the data. Further, since constructing text augmentations that do not alter the meaning (semantics) of a sentence is very difficult, they introduce two losses to ensure both sufficient difference and semantic consistency of sentence augmentations. They define a consistency loss to guarantee that augmentations lead to similar predictions y_c and a contrastive loss that makes augmented text representations a_i similar to the original text x_i. To ensure that a sufficiently large amount of negative text augmentations are sampled, they use an augmentation-embedding memory bank with a momentum encoder. Fang et al. [13] only use back-translation, Meng et al. [41], Wu et al. [64] investigate sentence augmentation methods, Giorgi et al. [15] contrast text spans, Clark et al. [9], Meng et al. [41] replace input words by re-sampling a language model, Gao et al. [14] sample positive text embeddings via dropout, and Simoulin and Crabbé [57] investigate contrastive sentence structure pretraining. Finally, Meng et al. [41] also contrasts cropped sentences after augmentation via word re-sampling.

Contrasting Next or Surrounding Sentence (or Word) Prediction (NSP, SSP). Sentence prediction is a popular input-input contrastive method as in Section 2.2. Next sentence prediction, NSP, and surrounding sentence prediction, SSP, take inspiration from the skip-gram model [43], where surrounding and non-surrounding words are contrastively predicted given a central word to learn word embeddings using an NCE Section 2.1 variant [43]. Methods mostly differ in how they sample positive and negative sentences, where negative sampling strategies such as undersampling frequent words, in Mikolov et al. [42], are crucial. Logeswaran and Lee [38] propose contrastive NSP, to predict the next sentence as a positive sample against n random negative sample sentences. Instead of generating the next sentence, they learn to discriminate which sentence encoding follows a given sentence. This allows them to train a better text encoder model with less computation but sacrifices the ability to generate text. Liu et al. [37] investigate variations of the contrastive NSP objective used in the BERT model. The method contrasts a consecutive sentence as a positive text sample against multiple non-consecutive sentences from other documents as negative text samples. They find that sampling negatives from the same document during self-supervised BERT pretraining is critical to downstream performance, but that removing the original BERT NSP task improves downstream performance. Iter et al. [23] find that predicting surrounding sentences in a k-sized window around a given central anchor sentence "improves discourse performance of language models." They sample surrounding sentences: (a) randomly from the corpus to construct easy negatives, and (b) from the same paragraph, but outside the context window as hard (to contrast) negative samples. Contextual negative sampling is theoretically and empirically proven by Saunshi et al. [55], who demonstrate that: "increased negative sampling only helps if negatives are taken from the original texts' context or block of information," i.e., the same document, paragraph, or sentence. Aroca-Ouellette and Rudzicz [2] study how to combine different variants of the NSP pretraining tasks with non-contrastive, auxiliary self-supervision signals, while Simoulin and Crabbé [57] explore contrastive sentence structure learning.

Input-output Contrastive Text Representation Pretraining. Rethmeier and Augenstein [53] use output label embeddings as an alternative view Y (labels) of text input embeddings X for contrastive learning of (dis-)similar text-label embedding pairs (X, Y) via binary NCE from Section 2.1. Using a separate label and text encoder enables the model to efficiently compute many negative label samples, while encoding the text X only once, unlike input-input view methods in Figure 3. They pretrain with random input words as pseudo-labels for self-supervised pretraining on a very small corpus, which despite the limited pretraining data enables unsupervised zero-shot prediction, largely improved few-shot and markedly better rare concept (long-tail) learning.

Contrastive Distillation. Sun et al. [59] propose CoDIR, a contrastive language model distillation method to pretrain a smaller student model from an already pretrained larger teacher such as a Masked Transformer Language Model. Compressing a pretrained language model is challenging, because nuances such as interactions between the original layer representation are easily lost—without noticing. For distillation, they extract layer representations from both the large teacher and the small student network over the same or two different input texts, to create a student and teacher view of said texts. Using the constrastive InfoNCE loss [62], they then learn to make the student representation similar to teacher representations for the same input texts, and dissimilar if they receive different texts. The score or similarity function in InfoNCE is measured as the cosine distance between mean pooled student and teacher Transformer layer representations. For negative sampling in pretraining, they use text inputs from the same topic, e.g., a Wikipedia article, to mine hard negative samples—i.e., they sample views from similar contexts as recommended for contrastive methods in Reference [55].

Text Generation as a Discriminative EBM. Deng et al. [11] combine an auto-regressive language model, with a contrastive text continuation EBM model for improved text generation. During pretraining, they learn to contrast real data text continuations and language model generated text continuations via conditional NCE from Section 2.1. For generation, they sample the top-k text completions from the auto-regressive language model and then score the best continuation via the trained EBM, to markedly improve model perplexity. However, the current approach is computationally expensive.

Cross-modal Contrastive Representation Pretraining. Representations for zero-shot image classification can be pretrained using image caption text for contrastive self-supervised pretraining. Jia et al. [26] automatically mine a large amount of noisy text captions for images in ALIGN, to then noise-filter and use them to construct matching and mismatching pairs of image and augmented text captions for contrastive training. Radford et al. [49] use the same idea in CLIP, but pretrain on a large collection of well annotated image caption datasets. Both methods allow for zero-shot image classification and image-to-text or text-to-image generation and are inherently zero-shot capable. Radford et al. [49] also run a zero-shot learning efficiency analysis for CLIP and find two things. First, they find that using a data-efficient CNN text encoder increases zero-shot image prediction convergence threefold compared to a Transformer text encoder, which they state to be computationally prohibitive. Second, they find that adding contrastive self-supervised text pretraining increases zero-shot image classification performance fourfold. Thus, CLIP [49] shows that contrastive self-supervised CNN text encoder pretraining can substantially outperform current Transformer pretraining methods, while ALIGN [26] also automates the image and caption data collection process to increase data scalability.

Vision-language grounding. Shi et al. [56] show that contrastive RoI-Feature prediction pretraining can increase performance on vision language grounding tasks compared to self-supervised masked token prediction or image-sentence matching predictions, especially when there is a gap

between the pretraining domain and end-task domain. Akula et al. [1] use a crowdsourced dataset to create a harder vision-language grounding task for robustness tests against common models and find that a contrastive method they design does increase performance, but is outperformed by a multi-task learning approach. Since contrastive losses have been shown to not overfit random labels (robustness) by Graf et al. [16], it is not clear whether there may be a label noise problem due to the crowdsourced nature of the data.

3.2 Supervised Contrastive Pretraining

Input-output Contrastive Supervised Text Representation Pretraining. Seen in Figure 2, Pappas and Henderson [47] train a two-input-lane Siamese CNN network, which encodes text as the input view x_i in one lane, and labels via a label encoder in a second data view y_c, to learn to contrast pairs of (x_i, y_x) as similar (1) or not (0). Rather than encoding labels as multi-hot vectors such as $[0, 1, 0, 0, 1]$, they express each label by a textual description of said label. These textual label descriptions can then be encoded by a label encoder subnetwork, which in the simplest case constructs a label embedding by averaging over the word embeddings of the words that describe a label. However, this requires manually describing each label. Using embeddings of supervised labels, they pretrain a contrastive text classification network on known positive and negative labels, and later apply the pretrained network to unseen classes for zero-shot prediction. Their method thus provides supervised, but zero-shot capable pretraining. While Rethmeier and Augenstein [53] also support supervised contrastive input-output pretraining, they automate label descriptions construction, and conjecture that in real-world scenarios, most labels, e.g., the word 'elephant', are already part of the input vocabulary and can thus be pretrained as word embeddings via methods such as Word2Vec [42]. They also note that: "once input words are labels, one can sample input words as pseudo label embeddings for contrastive self-supervised pretraining," as described in Section Section 3.1. Either method is contrastively pretrained via binary NCE as described in Section 2.1. Furthermore, both methods markedly boost few-shot learning and enable zero-shot predictions, while Rethmeier and Augenstein [53] enables unsupervised zero-shot learning via self-supervised contrastive pretraining. The added contrastive self-supervision further boosts few-shot and long-tailed learning performance, while also increasing convergence speed over supervised-only contrastive learning in Pappas and Henderson [47].

Contrastive Commonsense Pretraining. Klein and Nabi [31] use contrastive self-supervised pretraining to refine a pretrained BERT language model to drastically increase performance on pronoun disambiguation and the Winograd Schema Commonsense Reasoning task. Their method contrasts over candidate trigger words that affect which word a pronoun refers to. They first mine trigger word candidates from text differences in paraphrased sentences and then maximize the contrastive margin between candidate pair likelihoods. While general pretraining provides little pronoun disambiguation learning signal, their method demonstrates the design of task-specific contrastive learning to produce strong performance increases in *un- and supervised commonsense reasoning*.

Contrastive Text Summarization. Duan et al. [12] use a Transformer attention mechanism during abstractive sentence summarization learning to optimize two contrasting loss objectives. One loss maximizes the contributions of tokens with the most attention when predicting the summarized sentence. The other loss is connected to a second decoder head, which learns to minimize the contribution of the attention to other, non-summarization-relevant, tokens. This method can perhaps best be understood as contrastive, layer attention noise reduction. The main drawback of this method is the current dual network head prediction, which introduces a larger complexity compared to other contrastive methods.

Cross and Multi-modal Supervised Contrastive Text Pretraining for Representation Learning. Recent work from computer vision and time series prediction train with contrastive supervised losses to enable zero-shot learning or improve data-to-text generation. Jiang et al. [27] fuse image and text description information into the same representation space for generalized zero-shot learning—i.e., where at test time some classes are unseen, zero-shot, while other classes were seen during training. To do so, they first pretrain a supervised text-image encoder network to contrast (*image*, *text*, *label*) triplets of human annotated image classes. At test time, this contrastive network decides which text description best matches a given image. This works for seen and unseen classes, because classes are represented as text descriptions. Li et al. [36], Radford et al. [49] perform pretraining on manually annotated textual image descriptions to enable better generalization to unseen image classes. Uehara et al. [61] turn stock price value time series into textual stock change descriptions where the contrastive objectives markedly increase the fluency and non-repetitiveness of generated texts, especially when trained with little data.

Datasets Construction for Contrastive pretraining. Raganato et al. [51] automatically create a corpus of contrastive sentences for word sense disambiguation in machine translation by first identifying sense ambiguous source sentence words, and then creating replacement word candidates to mine sentences for contrastive evaluation.

4 CHALLENGES AND FUTURE OPPORTUNITIES

Opportunities: Data Efficiency, Fairness, and Small-scale Pretraining. Zimmermann et al. [66] proved that contrastive methods effectively recover data properties even from very limited data, which explains their few-shot label efficiency in both supervised contrastive fine-tuning [18] and contrastive re-pretraining of pretrained language model [13, 23, 58]. Additionally, contrastive language models like Clark et al. [9], Meng et al. [41], Rethmeier and Augenstein [53], Wu et al. [64] do not require other pretrained models, while largely improving pretraining data efficiency (zero-shot learning) or label efficiency (few-shot learning). For example, Rethmeier and Augenstein [53] propose a small contrastive language model to markedly improve *long-tail learning* over large pretrained language models. Contrastive modeling thus provides a promising direction to reducing algorithmic fairness issues that have been linked to a loss of minority (tail) information by Hooker et al. [22]. These aspects indicate that contrastive self-supervised models require far less pretraining data than other objectives, which opens their applications to data sparse domains, languages, productivity gains, and scalable or budget friendly language model pretraining.

Challenges: Better Negative and Positive Generation. Current methods require sampling many negative instances for contrastive learning to work well. Sampling hard [4] or context-relevant negatives [55] are known to boost sample efficiency during contrastive self-supervised learning, while sampling diverse negatives [43, 45, 53] have been demonstrated to improve generalization in open class set (or open tasks) applications such as pretraining. Sampling multiple positives for supervised contrast is common in multi-label metric learning and therefore somewhat studied. However, explicit experiments on the benefits of generating multiple positive samples for *contrastive self-supervision* are poorly understood, especially in NLP. To date, Wang and Isola [63] showed that positive self-supervision samples should be sampled close to one another (in computer vision), while sampling more contextual positives has been linked to largely improved language model pretraining sample efficiency gains in Rethmeier and Augenstein [53]. Works like BYOL [17] or Barlow Twins [65] do not require negative sampling. Their momentum contrast or redundancy reduction-based learning, may be adapted for contrastive language modeling to overcome current compute challenges of input-input contrastive NLP.

Challenge and Opportunity: Text Augmentation Quality. Self-supervised text augmentation research in NLP (Section 3.1) is gaining momentum and Chen et al. [5], Qu et al. [48] and many others analyze using mixes of recent text data augmentations. However, these input-input contrastive methods often use computationally expensive or non-robust mechanisms like: back translation, initializing a new prediction head per downstream task, or rely on already otherwise pretrained models like RoBERTa. Fortunately, more scalable and robust input augmentations have already been proposed by Iter et al. [23], Wu et al. [64], which is a promising step to cost effective future extensions.

Opportunities: for Data Limited NLP Sub-fields. Chi et al. [8] use contrastive pretraining to reduce data limitations in multi-lingual models, while Jiang et al. [28] use adversarial sample generation to make contrastive pretraining more sample efficient and robust. The contrastive **word sense disambiguation (WSD)** dataset construction method by Raganato et al. [51] is potentially adaptable to automatically mine inputs for contrastive pronoun learning in Klein and Nabi [31]. Such automation would help to scale contrastive common sense learning.

Opportunities: Underresearched NLP applications. An underresearched direction for contrastive NLP are data-to-text tasks that turn non-text inputs into a textual description. Uehara et al. [61], for instance, contrastively learn to generate stock change text descriptions from stock price time series using limited data, while works such as Jia et al. [26], Radford et al. [49] show that contrastive text supervision and self-supervision can multiply the zero-shot learning efficiency in cross-modal representation learning. Deng et al. [11] improve text generation with contrastive importance resampling of language model generated text continuations, while Duan et al. [12] propose contrastive abstractive sentence summarization, which using Momentum Contrast can potentially improve. Sun et al. [59] compress a large language model. Future work could adapt their method to fuse multiple language models or mutually transfer knowledge between models. Jiang et al. [29] present long-tail preserving vision model compression by contrasting easily pruned (forgotten) model information with large model information. Together with Rethmeier and Augenstein [53], Wu et al. [64] this may be used to learn and compress large long-tail capable language models that retain more tail class information to reduce compression-induced minority fairness losses as first identified by Hooker et al. [22].

Opportunity: Fringe Science. Works like Luss et al. [39] and Ross et al. [54] generate contrastive (counterfactual) explanations in vision or NLP, because humans give contrastive explanations. This could be used to ease the creation of semantically sensible input augmentations, which are used as negative or positive samples for contrastive learning. This would result in an optimization loop between explanation and contrast pair generation that amounts to energy minimization toward an equilibrium. Additionally, human annotations could be incorporated for data-efficient, explanation guided human-in-the-loop learning.

5 CONCLUSION

This primer on contrastive pretraining surveys contrastive learning concepts and their relations to other sub-fields like EBMs to ease advanced reading into the connected literature. It highlights recent methodological and theoretical insights that are important to designing effective contrastive learners for NLP. Finally, the primer structures contrastive pretraining as self-learning versus supervised learning summarises challenges and provides pointers to future research directions.

REFERENCES

[1] Arjun Akula, Spandana Gella, Yaser Al-Onaizan, Song-Chun Zhu, and Siva Reddy. 2020. Words Aren't enough, their order matters: On the robustness of grounding visual referring expressions. In *Proceedings of the 58th Annual Meeting*

of the Association for Computational Linguistics. Association for Computational Linguistics, 6555–6565. https://doi.org/10.18653/v1/2020.acl-main.586

[2] Stéphane Aroca-Ouellette and Frank Rudzicz. 2020. On losses for modern language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.403

[3] Malik Boudiaf, Jérôme Rony, Imtiaz Masud Ziko, Eric Granger, Marco Pedersoli, Pablo Piantanida, and Ismail Ben Ayed. 2020. A unifying mutual information view of metric learning: Cross-entropy vs. pairwise losses. In *Proceedings of the European Conference on Computer Vision (ECCV'20) (Lecture Notes in Computer Science, Vol. 12351)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer. https://doi.org/10.1007/978-3-030-58539-6_33

[4] Tiffany Tianhui Cai, Jonathan Frankle, David J. Schwab, and Ari S. Morcos. 2020. Are All Negatives Created Equal in Contrastive Instance Discrimination? Retrieved from https://arXiv:2010.06682.

[5] Jiaao Chen, Zichao Yang, and Diyi Yang. 2020. MixText: Linguistically-informed interpolation of hidden space for semi-supervised text classification. In *Proceedings of the Association for Computational Linguistics (ACL'20)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.194

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. 1597–1607. Retrieved from http://proceedings.mlr.press/v119/chen20j.html.

[7] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E. Hinton. 2020. Big self-supervised models are strong semi-supervised learners. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS'20)*. Retrieved from https://proceedings.neurips.cc/paper/2020/hash/fcbc95ccdd551da181207c0c1400c655-Abstract.html.

[8] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'21)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.280

[9] Kevin Clark, Minh-Thang Luong, Quoc Le, and Christopher D. Manning. 2020. Pre-training transformers as energy-based cloze models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.20

[10] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2270–2282. https://doi.org/10.18653/v1/2020.acl-main.207

[11] Yuntian Deng, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2020. Residual energy-based models for text generation. In *Proceedings of the International Conference on Learning Representations (ICLR'20)*. Retrieved from https://openreview.net/forum?id=B1l4SgHKDH.

[12] Xiangyu Duan, Hongfei Yu, Mingming Yin, Min Zhang, Weihua Luo, and Yue Zhang. 2019. Contrastive attention mechanism for abstractive sentence summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1301

[13] Hongchao Fang, Sicheng Wang, Meng Zhou, Jiayuan Ding, and Pengtao Xie. 2020. CERT: Contrastive Self-supervised Learning for Language Understanding. Retrieved from https://arXiv:2005.12766.

[14] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. Retrieved from https://arXiv:2104.08821.

[15] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. DeCLUTR: Deep contrastive learning for unsupervised textual representations. In *Proceedings of the Association for Computational Linguistics (ACL'21)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.72

[16] Florian Graf, Christoph Hofer, Marc Niethammer, and Roland Kwitt. 2021. Dissecting supervised constrastive learning. In *Proceedings of the International Conference on Machine Learning (ICML'21) (PMLR, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR. Retrieved from https://proceedings.mlr.press/v139/graf21a.html.

[17] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. 2020. Bootstrap your own latent—A new approach to self-supervised learning. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS'20)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2020/file/f3ada80d5c4ee70142b17b8192b2958e-Paper.pdf.

[18] Beliz Gunel, Jingfei Du, Alexis Conneau, and Veselin Stoyanov. 2021. Supervised contrastive learning for pre-trained language model fine-tuning. In *Proceedings of the 9th International Conference on Learning Representations (ICLR'21)*. OpenReview.net. Retrieved from https://openreview.net/forum?id=cu7IUiOhujH.

[19] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2021. Few-Shot Cross-Lingual Stance Detection with Sentiment-based Pre-Training. Retrieved from https://arXiv:2109.06050.

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR'20)*. 9726–9735. https://doi.org/10.1109/CVPR42600.2020.00975

[21] R. Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Philip Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *Proceedings of the 7th International Conference on Learning Representations (ICLR'19)*. Retrieved from https://openreview.net/forum?id=Bklr3j0cKX.

[22] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2020. What Do Compressed Deep Neural Networks Forget? Retrieved from https://arxiv.org/pdf/1911.05248.pdf.

[23] Dan Iter, Kelvin Guu, Larry Lansing, and Dan Jurafsky. 2020. Pretraining with contrastive sentence objectives improves discourse performance of language models. In *Proceedings of the Association for Computational Linguistics (ACL'20)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.439

[24] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2021). https://doi.org/10.3390/technologies9010002

[25] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2021. A survey on contrastive self-supervised learning. *Technologies* 9, 1 (2021). Retrieved from https://www.mdpi.com/2227-7080/9/1/2.

[26] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the International Conference on Machine Learning (ICML'21) (PMLR, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR. Retrieved from http://proceedings.mlr.press/v139/jia21b.html.

[27] Huajie Jiang, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2019. Transferable contrastive network for generalized zero-shot learning. In *Proceedings of the (IEEE/CVF ICCV'19)*. https://doi.org/10.1109/ICCV.2019.00986

[28] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. 2020. Robust pre-training by adversarial contrastive learning. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS'20)*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper/2020/file/ba7e36c43aff315c00ec2b8625e3b719-Paper.pdf.

[29] Ziyu Jiang, Tianlong Chen, Bobak J. Mortazavi, and Zhangyang Wang. 2021. Self-damaging contrastive learning. In *Proceedings of the International Conference on Machine Learning (ICML'21) (PMLR, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR. Retrieved from https://proceedings.mlr.press/v139/jiang21a.html.

[30] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Proceedings of the Conference and Workshop on Neural Information Processing Systems (NeurIPS'20)*. Retrieved from https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

[31] Tassilo Klein and Moin Nabi. 2020. Contrastive self-supervised learning for commonsense reasoning. In *Proceedings of the Association for Computational Linguistics (ACL'20)*. Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.acl-main.671.

[32] Lingpeng Kong, Cyprien de Masson d'Autume, Lei Yu, Wang Ling, Zihang Dai, and Dani Yogatama. 2020. A mutual information maximization perspective of language representation learning. In *Proceedings of the 8th International Conference on Learning Representations (ICLR'20)*. Retrieved from https://openreview.net/forum?id=Syx79eBKwr.

[33] Yann LeCun, Sumit Chopra, Raia Hadsell, Marc Aurelio Ranzato, and Fu Jie Huang. 2006. *A Tutorial on Energy-based Learning*. MIT Press.

[34] Yann LeCun and Fu Jie Huang. 2005. Loss functions for discriminative training of energy-based models. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AIStats'05)*. Retrieved from http://yann.lecun.com/exdb/publis/pdf/lecun-huang-05.pdf.

[35] Yann LeCun and Fu Jie Huang. 2005. Loss functions for discriminative training of energy-based models. In *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AIStats'05)*. Retrieved from http://www.gatsby.ucl.ac.uk/aistats/fullpapers/207.pdf.

[36] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq R. Joty, Caiming Xiong, and Steven C. H. Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. Retrieved from https://arxiv.org/abs/2107.07651.

[37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. Retrieved from http://arxiv.org/abs/1907.11692.

[38] Lajanugen Logeswaran and Honglak Lee. 2018. An efficient framework for learning sentence representations. In *Proceedings of the (ICLR'18)*. Retrieved from https://openreview.net/forum?id=rJvJXZb0W.

[39] Ronny Luss, Pin-Yu Chen, Amit Dhurandhar, Prasanna Sattigeri, Yunfeng Zhang, Karthikeyan Shanmugam, and Chun-Chen Tu. 2021. Leveraging latent features for local explanations. In *Proceedings of the (KDD SIGKDD'21)*, Feida Zhu, Beng Chin Ooi, and Chunyan Miao (Eds.). ACM. https://doi.org/10.1145/3447548.3467265

[40] Zhuang Ma and Michael Collins. 2018. Noise contrastive estimation and negative sampling for conditional models: Consistency and statistical efficiency. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'18)*. https://doi.org/10.18653/v1/d18-1405

[41] Yu Meng, Chenyan Xiong, Payal Bajaj, Saurabh Tiwary, Paul Bennett, Jiawei Han, and Xia Song. 2021. COCO-LM: Correcting and contrasting text sequences for language model pretraining. Retrieved from https://arxiv.org/abs/2102.08473.

[42] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR Workshop Track Proceedings*. Retrieved from http://arxiv.org/abs/1301.3781.

[43] Tomás Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *NeurIPS*. Retrieved from https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

[44] Andriy Mnih and Yee Whye Teh. 2012. A fast and simple algorithm for training neural probabilistic language models. In *Proceedings of the International Conference on Machine Learning (ICML'12)*. Omnipress, Madison, WI, 8 pages.

[45] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check. In *Proceedings of the European Conference on Computer Vision (ECCV'20)*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (Eds.). Springer International Publishing, Cham.

[46] Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. Retrieved from https://arxiv.org/abs/2202.06671.

[47] Nikolaos Pappas and James Henderson. 2019. GILE: A generalized input-label embedding for text classification. *Trans. Assoc. Comput. Linguistics* 7 (2019). Retrieved from https://transacl.org/ojs/index.php/tacl/article/view/1550.

[48] Yanru Qu, Dinghan Shen, Yelong Shen, Sandra Sajeev, Weizhu Chen, and Jiawei Han. 2021. CoDA: Contrast-enhanced and diversity-promoting data augmentation for natural language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR'21)*. Retrieved from https://openreview.net/forum?id=Ozk9MrX1hvA.

[49] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. Retrieved from https://cdn.openai.com/papers/Learning_Transferable_Visual_Models_From_Natural_Language.pdf.

[50] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67. Retrieved from http://jmlr.org/papers/v21/20-074.html.

[51] Alessandro Raganato, Yves Scherrer, and Jörg Tiedemann. 2019. The MuCoW test suite at WMT 2019: Automatically harvested multilingual contrastive word sense disambiguation test sets for machine translation. In *Proceedings of the Workshop on Statistical Machine Translation (WMT'19)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-5354

[52] Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. 2021. Do generative models know disentanglement? Contrastive learning is all you need. Retrieved from https://arxiv.org/abs/2102.10543.

[53] Nils Rethmeier and Isabelle Augenstein. 2020. Long-Tail Zero and Few-Shot Learning via Contrastive Pretraining on and for Small Data. Retrieved from https://arXiv:2010.01061.

[54] Alexis Ross, Ana Marasović, and Matthew Peters. 2021. Explaining NLP models via minimal contrastive editing (MiCE). In *Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP'21)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.336

[55] Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. 2019. A theoretical analysis of contrastive unsupervised representation learning. In *Proceedings of the International Conference on machine Learning (ICML'19) (PMLR, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR. Retrieved from http://proceedings.mlr.press/v97/saunshi19a.html.

[56] Lei Shi, Kai Shuang, Shijie Geng, Peng Gao, Zuohui Fu, Gerard de Melo, Yunpeng Chen, and Sen Su. 2021. Dense contrastive visual-linguistic pretraining. In *Proceedings of the ACM Multimedia Conference (MM'21)*. 5203–5212. https://doi.org/10.1145/3474085.3475637

[57] Antoine Simoulin and Benoit Crabbé. 2021. Contrasting distinct structured views to learn sentence embeddings. In *Proceedings of the EACL Student Research Workshop*. Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.eacl-srw.11.

[58] YuSheng Su, Xu Han, Yankai Lin, Zhengyan Zhang, Zhiyuan Liu, Peng Li, and Maosong Sun. 2021. CSS-LM: A contrastive framework for semi-supervised fine-tuning of pre-trained language models. Retrieved from https://arxiv.org/abs/2102.03752.

[59] Siqi Sun, Zhe Gan, Yuwei Fang, Yu Cheng, Shuohang Wang, and Jingjing Liu. 2020. Contrastive distillation on intermediate representations for language model compression. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.36

[60] Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. 2020. On mutual information maximization for representation learning. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=rkxoh24FPH.

[61] Yui Uehara, Tatsuya Ishigaki, Kasumi Aoki, Hiroshi Noji, Keiichi Goshima, Ichiro Kobayashi, Hiroya Takamura, and Yusuke Miyao. 2020. Learning with contrastive examples for data-to-text generation. In *Proceedings of the International Conference on Computational Linguistics (COLING'20)*. International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.213

[62] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. Retrieved from http://arxiv.org/abs/1807.03748.

[63] Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the Proceedings of the International Conference on machine Learning (ICML'20) (PMLR, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR. Retrieved from https://proceedings.mlr.press/v119/wang20k.html.

[64] Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. CLEAR: Contrastive Learning for Sentence Representation. Retrieved from https://arXiv:2012.15466.

[65] Jure Zbontar, Li Jing, Ishan Misra, Yann Lecun, and Stephane Deny. 2021. Barlow twins: Self-supervised learning via redundancy reduction. In *Proceedings of the Proceedings of the International Conference on machine Learning (ICML'21) (PMLR, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR. Retrieved from https://proceedings.mlr.press/v139/zbontar21a.html.

[66] Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. 2021. Contrastive Learning Inverts the Data Generating Process. Retrieved from https://arXiv:2102.08850.

Received 11 November 2021; revised 30 June 2022; accepted 23 August 2022