Unsupervised Natural Language Inference via Decoupled Multimodal Contrastive Learning

Wanyun Cui[§] **Guangyu Zheng**[‡] **Wei Wang**[‡]
cui.wanyun@sufe.edu.cn, simonzheng96@gmail.com,weiwang1@fudan.edu.cn
[§]Shanghai University of Finance and Economics
[‡]Shanghai Key Laboratory of Data Science, Fudan University

Abstract

We propose to solve the natural language inference problem without any supervision from the inference labels via task-agnostic multimodal pretraining. Although recent studies of multimodal self-supervised learning also represent the linguistic and visual context, their encoders for different modalities are coupled. Thus they cannot incorporate visual information when encoding plain text alone. In this paper, we propose Multimodal Aligned Contrastive Decoupled learning (MACD) network. MACD forces the decoupled text encoder to represent the visual information via contrastive learning. Therefore, it embeds visual knowledge even for plain text inference. We conducted comprehensive experiments over plain text inference datasets (i.e. SNLI and STS-B). The unsupervised MACD even outperforms the fully-supervised BiLSTM and BiLSTM+ELMO on STS-B.

1 Introduction

Humans are not supervised by the natural language inference (NLI). Supervision is necessary for applications in human-defined domains. For example, humans need the supervision of what is a noun before they do POS tagging, or what is a tiger in Wordnet before they classify an image of tiger in ImageNet. However, for NLI, people are able to entail that ⓐ `A man plays a piano` contradicts ⓑ `A man plays the clarinet for his family` without any supervision from the NLI labels. In this paper, we define such inference as a more general process of establishing associations and inferences between texts, rather than strictly classifying whether two sentences entail or contradict each other. Inspired by this, we raise the core problem in this paper: *Given a pair of natural language sentences, can machines entail their relationship without any supervision from inference labels?*

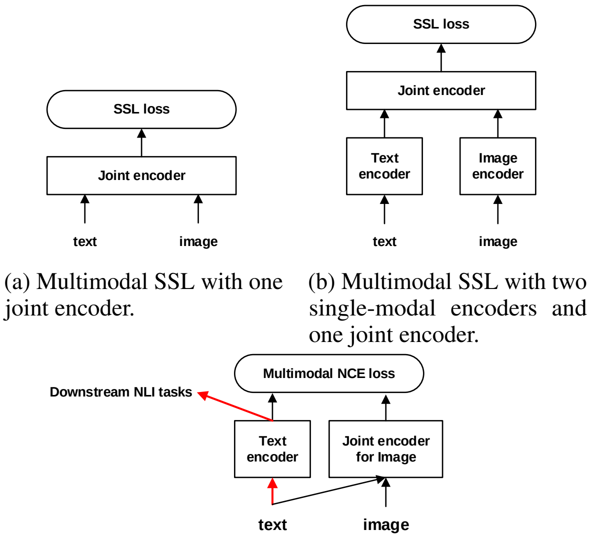
In his highly acclaimed paper, neuroscientist Moshe Bar claims that *"predictions rely on the existing scripts in memory, which are the result of real as well as of previously imagined experiences"* (Bar, 2009). The exemplar theory argues that humans use **similarity** to recognize different objects and make decisions (Tversky and Kahneman, 1973; Homa et al., 1981).

Analogy helps humans understand a novel object by linking it to a similar representation existing in memory (Bar, 2007). Such linking is facilitated by the object itself and its *context* (Bar, 2004). Context information has been widely applied in self-supervision learning (SSL) (Devlin et al., 2018; de Sa, 1994; He et al., 2020). Adapting context to NLI is even more straightforward. A simple idea of **constant conjunction** is that *A* causes *B* if they are constantly conjoined. Although constant conjunction contradicts "correlation is not causation", modern neuroscience has confirmed that humans use it for reasoning in their mental world (Levy and Steward, 1983). For example, they found an increase in synaptic efficacy arises from a presynaptic cell's repeated and persistent stimulation of a postsynaptic cell in Hebbian theory (Hebb, 2005). As to the natural language, the object and its context can be naturally used to determine the inference. For example, ⓐ contradicts ⓑ because they cannot happen simultaneously in the same **context**.

The context representation learned by SSL (e.g. BERT (Devlin et al., 2018)) has already achieved big success in NLP. From the perspective of context, these models (Devlin et al., 2018; Liu et al., 2019) learn the sentence level contextual information (i.e. by next sentence prediction task) and the word level contextual information (i.e. by masked language model task).

Besides linguistic contexts, humans also link other modalities (e.g. visions, voices) to novel inputs (Bar, 2009). Even if the goal is to reason about plain texts, other modalities still help (although they are not provided as inputs) (Kiela et al., 2018). For example, if only textual information is used, it is difficult to entail the contradiction between ⓐ and ⓑ. We need the commonsense that a man only has two arms, which cannot play the piano and clarinet simultaneously. This commonsense is hard to obtain from the text. However, if we link the sentences to their *visual scenes*, the contradiction is much clearer because the two scenes cannot happen in the same visual context. We think it is necessary to incorporate other modalities for the unsupervised natural language inference.

The idea of adapting multimodal in SSL is not new. According to (Su et al., 2020), we briefly divide previous multimodal SSL approaches into two categories based on their encoder infrastructures. As shown in Fig. 1a, the first category uses one joint encoder to represent the multimodal inputs (Sun et al., 2019; Alberti et al., 2019; Li et al., 2019, 2020; Su et al., 2020). Obviously, if the downstream task is only for plain text, we cannot extract the representation of text separately from the joint encoder. So the first category is infeasible for the natural language inference. The second category (Lu et al., 2019; Tan and Bansal, 2019; Sun et al., 2019) first encodes the text and the image separately by two encoders. Then it represents the multimodal information via a joint encoder over the lower layer encoders. This is shown in Fig. 1b. Although the textual representation can be extracted from the text encoder in the lower layer, such representation does not go through the joint learning module and contains little visual knowledge. In summary, the encoders in previous multimodal SSL approaches are coupled. If only textual inputs are given, they cannot effectively incorporate visual knowledge in their representations. Thus their help for entailing the contradiction between ⓐ and ⓑ is limited.



(a) Multimodal SSL with one joint encoder.

(b) Multimodal SSL with two single-modal encoders and one joint encoder.

(c) Our proposed multimodal aligned contrastive decoupled network. When adapting to downstream NLI tasks, we directly leverage the representation by the text encoder through the red lines, which only requires text as input.

Figure 1: Comparison of different multimodal SSL approaches.

In order to benefit from multimodal data in plain text inference, we propose the Multimodal Aligned Contrastive Decoupled learning (MACD) network. This is shown in Fig. 1c. Its text encoder is decoupled, which only takes the plain text as inputs. Thus it can be directly adapted to downstream NLI tasks. Besides, we use multimodal contrastive loss between the text encoder and the image encoder, thereby forcing the text representation to align with the corresponding image. Therefore even if the text encoder in MACD only takes the plain text as input, it still represents visual knowledge. In the downstream plain text inference tasks, without taking images as input, the text encoder of MACD still implicitly incorporating the visual knowledge learned by the multimodal contrastive loss. Note that we do not need a decoupled image encoder in the SSL. So the image encoder in Fig. 1c in MACD takes texts as inputs to provides a more precise image encoder. We will elaborate this in section 2.1.

2 Problem Formulation

We outline the general decoupled SSL process of MACD in section 2.1, and the downstream unsupervised NLI task in section 2.2.

2.1 Decoupled Multimodal SSL

For pretraining MACD, we use the multimodal training data $\mathcal{D}_{t2i} = \{x_i, y_i\}_{i=1}^{N}$ with N samples. Each sample $\{x_i, y_i\}$ consists of a pair of text x_i and image y_i, which describe the same context. It is straightforward to extend our method to modalities other than texts and images.

MACD learns from $\mathcal{D}_{t2i}$. Since text2image is many-to-many, we use energy-based models to represent their correlations. We first encode x_i and y_j into one pretext-invariant representation space (Misra and van der Maaten, 2020). The encoders are denoted by $f(x_i; \theta_f)$ and $g(x_i, y_i; \theta_g)$, respectively. We define the energy function

$\sigma(x_i, y_i) : X \times Y \to \mathbb{R}$ as

$$\sigma(x_i, y_i) = d(f(x_i; \theta_f), g(x_i, y_i; \theta_g)) \quad (1)$$

where $f(x_i; \theta_f)$ denotes the text encoder and $g(x_i, y_i; \theta_g)$ denotes the image encoder. d is a non-parametric distance metric (e.g. cosine). In the rest of this paper, we will use $f(x)$ and $g(x, y)$ instead of $f(x; \theta_f)$ and $g(x, y; \theta_g)$ for convenience.

Note that the text encoder $f(x)$ only takes the text as input, while the image encoder $g(x, y)$ takes both the image and the text as input. The higher the value of the energy function $\sigma()$, the higher the probability that x and y are in the same context, and vice versa. The forms of the encoders have the following advantages:

- The text encoder $f(x)$ and the image input y are *decoupled*. Therefore we represent x separately without knowing y. This allows us to use $f(x)$ in the downstream plain text inference.

- $g(x, y)$ represents the *one-to-many* relationship via implicitly introducing the "predictive sparse coding" (Gregor and LeCun, 2010). One image has multiple corresponding texts. To use energy-based models to represent the one-to-many relationship, one common approach is to introduce a noise vector z to allow multiple predictions through one image (Bojanowski et al., 2018). Note that such z can be quickly estimated by the given text x and image y (Gregor and LeCun, 2010). In our proposed image encoder $g(x, y)$, although z is not explicitly introduced, the encoder allows multiple predictions for one image via taking different images as input. Besides, it allows the image to interact with the text in the inner computation, which is an implicit alternative for the predictive z.

2.2 Downstream Unsupervised NLI

We use the representation from the pre-trained multimodal SSL to predict the relations of natural language sentence pairs under the unsupervised learning scenario. The testing data can be formulated as $\mathcal{D}_{test} = \{x_i^T, z_i\}_{i=1}^M$, each $x_i^T = (x_i^1, x_i^2)$ is composed of a sentence pair x_i^1 and x_i^2. z_i indicates the relation between x_i^1 and x_i^2. Under the unsupervised setting, we predict z_i for given x_i^T by the similarity of $f(x_i^1)$ and $f(x_i^2)$ (e.g. cosine similarity).

3 Methods

This section elaborates our major methodology. In section 3.1, we show how we maximize the cross-modal mutual information (MI) for the decoupled representation learning. In section 3.2, we show how we incorporate the mutual information (MI) of local structures. We elaborate the encoders in section 3.3. In Section 3.4, in order to solve the catastrophic forgetting problem, we use lifelong learning regularization to anchor the text.

3.1 Decoupled Representation Learning by Cross-Modal Mutual Information Maximization

As discussed in section 1, the query object and its context determine the inference. NLI depends on whether the two sentences are in the same context. In this paper, we consider context from different modalities (e.g. text or images).

Mutual information maximization has become a trend for SSL (Tian et al., 2019; Hjelm et al., 2019). For cross-modal SSL, we also leverage mutual information $\mathcal{I}(X, Y)$ to represent the correspondence between the text and the image. Intuitively, high mutual information means that the text and the image are well-matched. More formally, the goal of multimodal representation learning is to maximize their mutual information:

$$\mathcal{I}(X, Y) = \sum_{x,y} P(x, y) \log \frac{P(x|y)}{P(x)} \quad (2)$$

Eqn. (2) is intractable and thereby hard to compute. To approximate and maximize $\mathcal{I}(X, Y)$, we use Noise-Contrastive Estimation (NCE) (Gutmann and Hyvärinen, 2010; Oord et al., 2018). First, we use the function $\sigma(x, y)$ to represent the term $\frac{P(x|y)}{P(x)}$ in Eqn. (2):

$$\sigma_{\text{global}}(x, y) \propto \frac{P(x|y)}{P(x)} \quad (3)$$

where $\sigma_{\text{global}}(x, y) : X \times Y \to \mathbb{R}$ is not a real probability and can be unnormalized. Here we use the notation "global" for the representation learning of a complete text or a complete image to distinguish from the local structures in section 3.2.

To compute the cross-modal mutual information, we first encode x and y to $f_{\text{global}}(x)$ and $g_{\text{global}}(y)$, respectively. Then we use the similarities of their encodings to model $\frac{P(x|y)}{P(x)}$. Note that $g_{\text{global}}(y)$ is a specific form of $g(x, y)$ in Eqn. (1). So $f_{\text{global}}(x)$ and $g_{\text{global}}(y)$ satisfy the form of f

and g in Eqn. (1). We will show how to incorporate the linguistic input when designing the encoder of local visual structures in section 3.3. We follow (Misra and van der Maaten, 2020) to compute the pretext-invariant energy function by the exponential function of their cosine similarity:

$$\begin{aligned}\sigma_{\text{global}}(x,y) &= d(f_{\text{global}}(x), g_{\text{global}}(y)) \\ &= \exp(\frac{cosine(f_{\text{global}}(x), g_{\text{global}}(y))}{\tau_\sigma})\end{aligned} \tag{4}$$

where τ_σ is a hyper-parameter of temperature.

To estimate $\sigma_{\text{global}}(x,y)$ and maximize the mutual information in Eqn. (2), the NCE loss (Oord et al., 2018) provides a valid toolkit. By taking the posterior probability $P(y|x)$, the NCE loss is defined as:

$$\begin{aligned}&\mathcal{L}^{\text{NCE}:P(y|x)}(X,Y) = -\mathbb{E}_{x,y\sim P(y|x)\tilde{P}(x)}\{\log \sigma_{\text{global}}(x,y) \\ &- \log \sum_{y'\sim P(y)} \sigma_{\text{global}}(x,y')\}\end{aligned} \tag{5}$$

where $\tilde{P}(x)$ denotes the real distribution of x, $P(y|x)\tilde{P}(x)$ denotes the distribution of y for given x, and $P(y)$ denotes the noise distribution of y. Thus minimizing Eqn. (5) can be seen as identifying the positive image $y \sim P(y|x)$ for given x from the noise image distribution $y \sim P(y)$.

It has been proved (Oord et al., 2018) that $\mathcal{L}^{\text{NCE}:P(y|x)}(X,Y)$ provides the lower bound of $\mathcal{I}(X,Y)$:

$$\mathcal{I}(X,Y) \geq \log N' - \mathcal{L}^{\text{NCE}:P(y|x)} \tag{6}$$

where N' denotes the number of noise samples and can be seen as a constant. So instead of maximizing $\mathcal{I}(X,Y)$ directly, we minimize $L^{\text{NCE}:P(y|x)}(X,Y)$ instead to maximize its lower bound.

Symmetrically, we also compute the NCE loss by taking the posterior probability $P(x|y)$. We define $\mathcal{L}^{\text{NCE}:P(x|y)}$ as:

$$\begin{aligned}\mathcal{L}^{\text{NCE}:P(x|y)}(X,Y) =& -\mathbb{E}_{x,y\sim P(x|y)\tilde{P}(y)}\{\log \sigma_{\text{global}}(x,y) \\ &- \log \sum_{x'\in P(x)} \sigma_{\text{global}}(x',y)\}\end{aligned} \tag{7}$$

Eqn. (7) can be seen as identifying the positive text $x \sim P(x|y)$ for given y from the noise text distribution $x \sim P(x)$.

By combining Eqn. (5) and Eqn. (7), we derive the loss for global MI maximization

$$\mathcal{L}^{\text{NCE}}_{\text{global}}(X,Y) = \mathcal{L}^{\text{NCE}:P(x|y)}(X,Y) + \mathcal{L}^{\text{NCE}:P(y|x)}(X,Y) \tag{8}$$

Here we say the MI is global, because it is over the complete text and the complete images, which are contrary to the local structures in section 3.2.

Negative sampling In practice, to compute $L^{\text{NCE}:P(y|x)}(X,Y)$, we need to construct noise samples for positive samples. We use all the $\{x_i, y_i\}$ pairs in the same minibatch from $\mathcal{D}_{t2i}$ as X, Y. Each y_i is the positive samples of x_i (i.e. $P(y_i|x_i) = 1$). For each $x_i \in X$, the noise y' in Eqn. (5) are sampled from Y. Likewise, to compute $\mathcal{L}^{\text{NCE}:P(x|y)}(X,Y)$ in Eqn. (7), we treat x_i as the positive sample for y_i, and other texts from the same minibatch as the noise samples.

3.2 MI Optimization for Local Structures

In this subsection, we incorporate the local information in multimodal contrastive learning. As demonstrated in DIM (Hjelm et al., 2019), local information plays a greater role in self-supervised learning than the global information.

We follow BERT (Devlin et al., 2018) and DIM to use the words and patches as the local structures for the text and the image, respectively. We maximize the MI between the cross-modal local/global structures. We denote a sentence x with L words as $x^{(1)} \cdots x^{(L)}$, and an image y with $M \times M$ patches as $y^{(1)} \cdots y^{(M^2)}$.

Similar to the objective of representation learning of global information, we use NCE as the objective of local information representation learning. The difference is that we use the local structure-based alignment to calculate the energy function, while there is no such objective in the representation learning of global information. This objective allows representation learning to emphasize the alignments of local structures between different modalities, such as the alignment between the word "piano" and the corresponding image patches.

Specifically, we use $\mathcal{L}^{\text{NCE}}_{\text{local}}(X,Y)$ to represent the loss of local information representation learning. The computation of $\mathcal{L}^{\text{NCE}}_{\text{local}}(X,Y)$ follows Eqn. (5)(7)(8), except that we replace σ_{global} with σ_{local} based on the local information alignment. We will elaborate on σ_{local} in section 3.3.

3.3 Alignment-based Local Energy Function and Representation Learning

In this subsection, we show the details of the local energy function σ_{local} and the encoders for local structures.

Following the form of Eqn. (1), we denote the encoders for the local structures of text as $f_{\text{word}}(x^{(i)})$. We denote the joint encoder for patches as $g_{\text{local}}(x, y^{(i)})$, which represents the linguistic information of patch $y^{(i)}$. Note that the

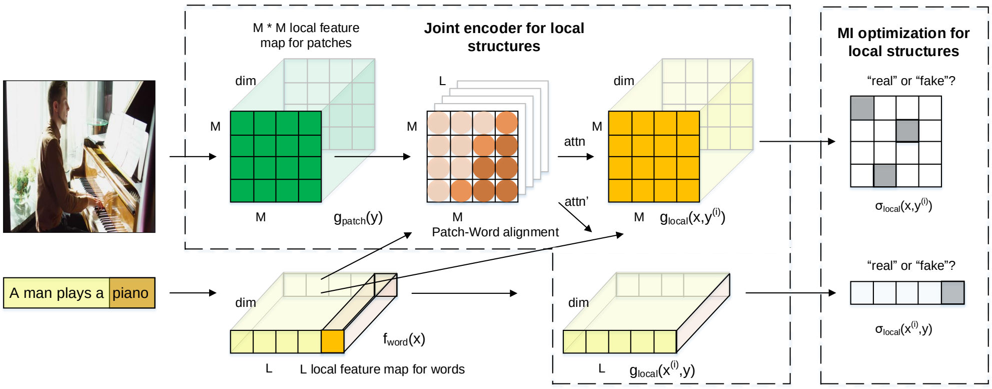


Figure 2: MI maximization for local structures. The local structures for images are joint encoded with text.

encoder $f_{\text{word}}(x^{(i)})$ is still decoupled and represents the local linguistic structures without taking image as input. On the other hand, the encoder $g_{\text{local}}(x, y^{(i)})$ for the local visual structure explicitly incorporate the linguistic information, which is more precise due to the discussion in section 2.1.

For a sentence x with L words $x^{(1)} \cdots x^{(L)}$, we represent its local information by encoding it into a local feature map $\boldsymbol{f}_{\text{word}}(x) = (f_{\text{word}}(x^{(1)}) \cdots f_{\text{word}}(x^{(L)})) \in \mathbb{R}^{dim \times L}$. For an image y with $M \times M$ patches $y^{(1)} \ldots y^{(M^2)}$, we represent its spatial locality by encoding it into a feature map $\boldsymbol{g}_{\text{patch}}(y) = \left(g_{\text{patch}}(y^{(1)}) \cdots g_{\text{patch}}(y^{(M^2)})\right)$.

The local information across modalities has obvious correlation characteristics (Xu et al., 2018). For example, a word is only related to some patches of the image, but not to other patches. As shown in Fig. 1c, our proposed image encoder is coupled with the text representation. Therefore we assign the local structures with different weights to achieve a more precise image encoder. This is achieved by the attention mechanism in the joint encoder:

$$g_{\text{local}}(x, y^{(i)}) = \frac{exp(attn_{i,j}/\tau_c)}{\sum_k exp(attn_{k,i}/\tau_c)} \boldsymbol{f}_{\text{word}}(x) \tag{9}$$

where τ_c denotes the temperature, $attn_{i,j}$ denotes the attention of the i-th word to the j-th patch:

$$attn_{i,j} = \frac{exp(f_{\text{word}}(x^{(i)})^\top g_{\text{patch}}(y^{(j)}))}{\sum_k exp(f_{\text{word}}(x^{(i)})^\top g_{\text{patch}}(y^{(k)}))} \tag{10}$$

We compute the alignment score for the local textual structures by:

$$\sigma_{\text{local}}(x, y^{(i)}) = d(g_{\text{patch}}(y^{(i)}), g_{\text{local}}(x, y^{(i)})) \tag{11}$$

Here we abuse the notation of σ_{local} since we will use $\sigma_{\text{local}}(x, y^{(i)})$ to compute $\sigma_{\text{local}}(x, y)$.

Symmetrically, we also compute the alignment score for the local visual structures by

$$\begin{aligned} attn'_{i,j} &= \frac{exp(f_{\text{word}}(x^{(i)})^\top g_{\text{patch}}(y^{(j)}))}{\sum_k exp(f_{\text{word}}(x^{(k)})^\top g_{\text{patch}}(y^{(j)}))} \\ g_{\text{local}}(x^{(i)}, y) &= \frac{exp(attn'_{i,j}/\tau_c)}{\sum_k exp(attn'_{i,k}/\tau_c)} \boldsymbol{g}_{\text{patch}}(y) \\ \sigma_{\text{local}}(x^{(i)}, y) &= d(f_{\text{word}}(x^{(i)}), g_{\text{local}}(x^{(i)}, y)) \end{aligned} \tag{12}$$

We compute the energy function of x and y based on local structure alignments by:

$$\begin{aligned} \sigma_{\text{local}}(x, y) = &\log \sum_{i=1}^{L} exp(\sigma_{\text{local}}(x^{(i)}, y)) \\ &+ \log \sum_{i=1}^{M^2} exp(\sigma_{\text{local}}(x, y^{(i)})) \end{aligned} \tag{13}$$

How the model uses the attention mechanism to represent the interactions among local structures and how the energy function is computed is shown in Fig. 2.

3.4 Anchor Text via Lifelong Learning

In this subsection, we illustrate how to solve the catastrophic forgetting problem by the lifelong learning regularization.

If we only use the loss in Eqn. (8), the text encoder $f(x; \theta_f)$ will tend to only learn vision-related features for text. Since our downstream problem is over the plain text, NLI still relies more on textual features instead of visual features. Compared with the single modality unsupervised natural language representation learning (Devlin et al., 2018), the multimodal model will even perform worse. Similar phenomena called catastrophic forgetting or negative transfer (Sun et al., 2020) often occurs in multi-task learning.

To avoid the catastrophic forgetting, we keep the model's representation for general text while ensuring that it learns visual features. More generally, since there are only data of a certain modality (i.e. plain text) in the downstream task, we anchor this modality in the multimodal SSL phase. We add lifelong learning regularization (Li and Hoiem, 2017) to achieve modality anchoring. For the text encoder, we keep its original textual representation (e.g. by masked language model (MLM) and next sentence prediction in BERT) while learning new visual knowledge. To do this, we follow (Li and Hoiem, 2017) and introduce the distance from the existing text encoder to the original text encoder as the training loss.

Specifically, we use BERT (Devlin et al., 2018) to initialize our text encoder $f(x)$. During multimodal SSL, we keep the textual representation consistent with the original BERT. According to the ablation study in DistilBERT (Sanh et al., 2019), we use the knowledge distillation loss (Hinton et al., 2015) and cosine loss as regularization:

$$\begin{aligned}&\mathcal{L}_{anchor}(X) = \mathbb{E}_{x\sim\tilde{P}(x)}[\\&-\epsilon\sum_{i=1}^{dim}\frac{f_i(x)^{1/\tau'}}{\sum_j f_j(x)^{1/\tau'}}\log\frac{f'_i(x)^{1/\tau'}}{\sum_j f'_j(x)^{1/\tau'}}\\&-(1-\epsilon)cosine(f(x),f'(x))]\end{aligned} \tag{14}$$

where $f'(x)$ denotes the textual representation by the original BERT encoder, $f_i(x)$ denotes the i-th dimension of $f(x)$, and τ' is the temperature.

By combing the lifelong learning regularization, we obtain the final loss for SSL:

$$\begin{aligned}\hat{\theta}_x,\hat{\theta}_y,\hat{\theta}_\alpha = &\underset{\theta_f,\theta_g}{\text{argmax}}\,\gamma\mathcal{L}^{\text{NCE}}_{\text{global}}(X,Y)\\&+\beta\mathcal{L}^{\text{NCE}}_{\text{local}}(X,Y)+(1-\gamma-\beta)\mathcal{L}_{anchor}(X)\end{aligned} \tag{15}$$

4 Experiments

4.1 Setup

All the experiments run over a computer with 4 Nvdia Tesla V100 GPUs.

Datasets We use Flickr30k (Young et al., 2014) and COCO (Lin et al., 2014) as the text2image dataset $\mathcal{D}_{t2i}$ for self-supervised learning. We use STS-B (Cer et al., 2017) and SNLI (Bowman et al.) as the downstream NLI tasks for evaluation. STS-B is a collection of sentence pairs, each of which has a human-annotated similarity score from 1 to 5. The task is to predict these scores. We follow GLUE (Wang et al., 2018) and use Pearson and Spearman correlation coefficients as metrics. SNLI is a collection of human-written English sentence pairs, with manually labeled categories *entailment*, *contradiction*, and *neutral*. Note that for STS-B, some sentence pairs drawn from image captions overlap with Flickr30k. So in order to avoid the potential information leak, we remove all sentence pairs drawn from image captions in STS-B to construct a new dataset *STS-B-filter*. Similarly, we remove all sentence pairs in SNLI whose corresponding images occur in the training split of $\mathcal{D}_{t2i}$ to construct *SNLI-filter*.

The statistics of these datasets are shown in Table 1. In addition, Flickr30k has 22248 images for training, 9535 images for development. COCO has 82783 images for training, 40504 images for development.

	Type	#Text		
		Train	Dev	Test
Flickr30k	Text2Image	111240	47675	-
COCO	Text2Image	414113	202654	-
STS-B	Text Similarity	5749	1500	1379
STS-B-filter	Text Similarity	3749	875	754
SNLI	NLI	549367	9842	9824
SNLI-filter	NLI	157284	3321	3207

Table 1: Statistics of datasets.

4.2 Model Details

Encoder details We use BERT-base as the text encoder f_{global}. The local information $f_{\text{word}}(x^{(i)})$ is the feature vector of the i-th word through BERT. We use *Resnet-50* as the image encoder g_{global}. We use the encoding before the final pooling layer as the representations of M^2 patches $g_{\text{patch}}(y^{(i)})$. To guarantee that the image encoder and the text encoder are in the same space, we project the feature vectors of the image encoder to the dimension of 768, which is the dimension of BERT.

Unsupervised NLI We compute the similarity of two sentences via the cosine of their representations learned by MACD. For STS-B, such similarities are directly used to compute the Pearson and Spearman correlation coefficients. For SNLI, we make inferences based on whether the similarity reaches a certain threshold. More specifically, if the similarity $>= \psi_1$, we predict “entailment”. If the similarity $< \psi_2$, we predict “contradiction”. Otherwise we predict “neutral”.

Competitors We compare MACD with the single-modal pre-training model BERT, and multimodal pre-training model LXMERT (Tan and Bansal, 2019) and VilBert (Lu et al., 2019). Both LXMERT and VilBert use the network architecture as in Fig. 1b. We extract the lower layer text encoder for unsupervised representation and fine-tuning. We also compare MACD with classical NLP models, including BiLSTM and BiLSTM+ELMO (Peters et al., 2018).

Hyper-parameters We list the hyper-parameters below. For ψ_1 and ψ_2, we use the best set of values chosen in the grid search from range $\{-1, -0.95, -0.9, \cdots, 1\}$. For τ_σ and τ_c, we use the best set of values chosen in the grid search from range $\{0.01, 0.1, 1\}$. For τ', ϵ, γ and β, we follow their settings in DistilBert (Sanh et al., 2019).

τ_σ	τ_c	τ'	ϵ	γ	β
0.1	1	2	5/6	1/3	1/3
Batch Size	lr	Epochs	Grad Acc	ψ_1	ψ_2
64	1e-4	10	8	0.80	0.55

Table 2: Hyper-parameters for self-supervised learning. "lr" means learning rate.

4.3 Main Results

We evaluate MACD by unsupervised NLI. Table 3 shows the results on STS-B. MACD achieves significantly higher effectiveness than single-modal pre-trained model BERT and multimodal pre-trained model LXMERT and VilBert. Note that LXMERT and VilBert use more text2image corpora for pre-training than MACD. This verifies that the joint encoder in previous multimodal SSL cannot represent visual knowledge well in their text encoder. So their adaptations to the single-modal problem are limited.

To our surprise, the unsupervised MACD even outperforms fully-supervised models such as BiLSTM and BiLSTM+ELMO. Here the results of BiLSTM and BiLSTM+ELMO for STS-B are directly derived from GLUE (Wang et al., 2018). This verifies the effectiveness of MACD.

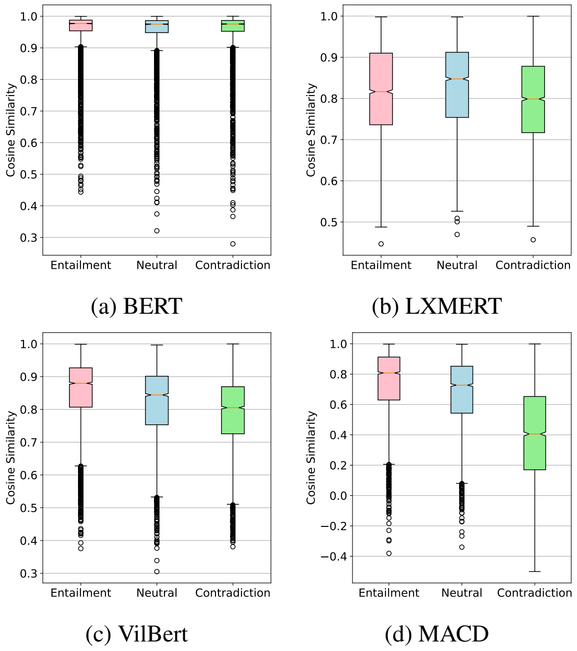
	STS-B		STS-B-filter	
	P.	S.	P.	S.
BiLSTM (sup.)	66.0	62.8	47.0	43.2
BiLSTM+ELMO (sup.)	64.0	60.2	33.3	30.7
BERT	1.7	6.4	5.5	12.5
LXMERT	42.7	47.2	35.9	40.0
VilBert	55.8	57.1	45.9	46.3
MACD + COCO	70.1	70.2	55.1	52.4
MACD + Flickr30k	**71.5**	**72.1**	**55.8**	**54.8**

Table 3: Effectiveness of unsupervised learning on STS. Baselines with "(sup.)" mean they are trained by supervised labels. Other methods are unsupervised. "P." and "S." mean Pearson and Spearman correlation coefficients, respectively.

	SNLI Acc	SNLI-filter Acc
BERT	35.09	35.45
LXMERT	39.03	40.29
VilBert	43.13	43.83
MACD + COCO	**52.63**	53.15
MACD + Filckr30k	52.27	**53.20**

Table 4: Effectiveness on SNLI. All approaches are unsupervised.

We also report the results of MACD on SNLI under the unsupervised setting in Table 4. MACD outperforms its competitors by a large margin. This verifies the effectiveness of our approach for unsupervised NLI. The experimental results suggest that *we achieve natural language inference via multimodal self-supervised learning without any supervised inference labels.* Since MACD+Filckr30k performs better than MACD+COCO in most cases, we will only evaluate MACD+Filckr30k in the rest experiments.



(a) BERT

(b) LXMERT

(c) VilBert

(d) MACD

Figure 3: Categorial distribution visualization.

We visualize the distribution of the cosine similarities for samples of different labels in SNLI in Fig. 3 by boxplot. We found obvious distribution patterns by MACD. In contrast, the distributions of other pre-training models have lower correlations with NLI labels.

4.4 Fine-tuning

We also evaluated the effectiveness of MACD when fine-tuned under the semi-supervised learning setting. More specifically, we first initialize the parameters of the text encoder as in MACD, then fine-tune it by the supervised training samples of the downstream tasks. The results are shown in Table 5.

MACD also outperforms other approaches. For example, for SNLI-filter, the accuracy of MACD increases by 0.97 compared to the best competitor (i.e. BERT). Note that MACD is the only multimodal method that performs better than BERT. Other multimodal approaches (i.e. LXMERT and VilBert) perform even worse than the original BERT, although they also initialize their text encoders by BERT, and use more text2image data for SSL than MACD. This verifies the effectiveness of the proposed decoupled contrastive learning model.

	STS-B	STS-B-filter	SNLI	SNLI-filter
	P. / S.	P. / S.	Acc	Acc
BERT	85.0/83.6	75.8/74.6	89.37	87.15
LXMERT	63.3/59.2	37.3/28.3	87.80	83.57
VilBert	78.8/77.2	63.9/62.2	88.49	85.69
MACD	**87.1/86.4**	**79.5/78.0**	**90.01**	**88.12**

Table 5: Effectiveness of fine-tuning over STS-B and SNLI. "P." and "S." mean Pearson and Spearman correlation coefficients, respectively.

To further verify the natural language representation learned by the self-supervised learning and get rid of the influence of its neural network architecture (i.e., BERT), Hjelm et al. (2019) suggest training models directly over the features learned by SSL. By following its settings (Hjelm et al., 2019), we use a linear classifier (SVM) and a nonlinear classifier (a single layer perception neural network, marked as SLP) over the features by SSL. The results are shown in Table 6.

	STS-B	STS-B-filter	SNLI	SNLI-filter
	P. / S.	P. / S.	Acc	Acc
SVM+BERT	69.8 / 68.3	57.1 / 53.3	58.77	58.87
SVM+LXMERT	33.0 / 31.3	10.2 / 13.2	52.28	50.98
SVM+VilBert	52.4 / 50.0	36.7 / 35.9	55.93	55.22
SVM+MACD	**70.0 / 68.4**	**62.2 / 59.3**	**61.64**	**62.58**
SLP+BERT	56.2 / 53.5	47.3 / 42.0	55.07	54.19
SLP+LXMERT	36.5 / 33.4	16.1 / 12.3	52.41	50.42
SLP+VilBert	49.6 / 46.0	29.1 / 26.5	54.86	51.82
SLP+MACD	**72.3 / 69.7**	**63.4 / 59.5**	**61.31**	**60.80**

Table 6: Effectiveness of the learned representations.

MACD outperforms the competitors by a large margin. Similar to the results in Table 5, although MACD, LXMERT, and VilBert are all trained by multimodal data, only MACD performs better than the original text encoder (i.e. BERT).

4.5 Ablations

In addition to the decoupled contrastive learning model, we propose two optimizations by adding the local structures into account, and by regularizing the model on the text mode via lifelong learning. In order to verify the effectiveness of the two optimizations, we compare MACD with its ablations. The results of unsupervised NLI are shown in Table 7. The results show that the effectiveness decreases when the proposed optimizations are removed.

	STS-B	STS-B-filter
	P. / S.	P. / S.
MACD	**71.5 / 72.1**	**55.8 / 54.8**
-local	71.0 / 70.9	55.0 / 52.6
-lifelong	70.7 / 70.8	54.9 / 52.3
-local -lifelong	69.6 / 69.7	53.0 / 52.0

Table 7: Ablations.

4.6 Case studies: Nearest-neighbor analysis

To give a deeper insight into the learned representation, we analyze the k nearest neighbors over the representations. For the query sentence randomly sampled from Flickr30k, we show the results of the 3 nearest sentences according to their L1 distances in Table 8. The results of MACD are more interpretable than BERT.

Query	Someone is wearing a large white dress in a crowd.
MACD No.1	Lady dressed in white on blanket in middle of crowd.
MACD No.2	Women in white robes, dancing with half their face painted.
MACD No.3	A group of women dressed in white are dancing in the street.
BERT No.1	A man is standing alone in a boat.
BERT No.2	A bald man is standing in a crowd.
BERT No.3	A woman is taking a picture of a man.

Table 8: Nearest-neighbor on the encoded text.

5 Conclusion

In this paper, we study the multimodal self-supervised learning for unsupervised NLI. The major flaw of previous multimodal SSL methods is that they use a joint encoder for representing the cross-modal correlations. This prevents us from integrating visual knowledge into the text encoder. We propose the multimodal aligned contrastive decoupled learning (MACD), which learns to represent visual knowledge while using only texts as inputs. In the experiments, our proposed approach steadily surpassed other methods by a large margin.

Acknowledgments

This paper was supported by National Natural Science Foundation of China (No. 61906116, No. 61732004), by Shanghai Sailing Program (No. 19YF1414700).

References

Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. 2019. Fusion of detected objects in text for visual question answering. In *EMNLP-IJCNLP*, pages 2131–2140.

Moshe Bar. 2004. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629.

Moshe Bar. 2007. The proactive brain: using analogies and associations to generate predictions. *Trends in cognitive sciences*, 11(7):280–289.

Moshe Bar. 2009. The proactive brain: memory for predictions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1521):1235–1243.

Piotr Bojanowski, Armand Joulin, David Lopez-Pas, and Arthur Szlam. 2018. Optimizing the latent space of generative networks. In *ICML*, pages 600–609.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. In *EMNLP*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *SemEval-2017*, pages 1–14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Karol Gregor and Yann LeCun. 2010. Learning fast approximations of sparse coding. In *ICML*, pages 399–406.

Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *AISTATS*, pages 297–304.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. *CVPR*.

Donald Olding Hebb. 2005. *The organization of behavior: A neuropsychological theory*. Psychology Press.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network.

R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio. 2019. Learning deep representations by mutual information estimation and maximization. In *ICLR*.

Donald Homa, Sharon Sterling, and Lawrence Trepel. 1981. Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning and Memory*, 7(6):418.

Douwe Kiela, Alexis Conneau, Allan Jabri, and Maximilian Nickel. 2018. Learning visually grounded sentence representations. In *NAACL-HLT*, pages 408–418.

WB Levy and O Steward. 1983. Temporal contiguity requirements for long-term associative potentiation/depression in the hippocampus. *Neuroscience*, 8(4):791–797.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and Ming Zhou. 2020. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. *AAAI*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Zhizhong Li and Derek Hoiem. 2017. Learning without forgetting. *TPAMI*, 40(12):2935–2947.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *ECCV*, pages 740–755.

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. *CoRR*, abs/1907.11692.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pages 13–23.

Ishan Misra and Laurens van der Maaten. 2020. Self-supervised learning of pretext-invariant representations. *CVPR*.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL*.

Virginia R de Sa. 1994. Learning classification with unlabeled data. In *NeurIPS*, pages 112–119.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations. In *ICLR*.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *ICCV*, pages 7464–7473.

Tianxiang Sun, Yunfan Shao, Xiaonan Li, Pengfei Liu, Hang Yan, Xipeng Qiu, and Xuanjing Huang. 2020. Learning sparse sharing architectures for multiple tasks. In *AAAI*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP-IJCNLP*, pages 5103–5114.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2019. Contrastive multiview coding. *arXiv preprint arXiv:1906.05849*.

Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology*, 5(2):207–232.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *EMNLP*, page 353.

Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *CVPR*, pages 1316–1324.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2:67–78.