embedding would simply learn the degenerate solution of mapping every point to the same representation.

A particular contrastive learning algorithm is defined predominantly by how the positive and negative pairs are chosen, which is how we use our prior knowledge to specify what a good representation should be. For example, consider the problem of learning representations of images. Here, a common choice is to create positive pairs by corrupting the input images in ways that should preserve the semantic information of the image while greatly altering the image in the pixel space (Wu *et al.*, 2018; He *et al.*, 2019; Chen, Kornblith, *et al.*, 2020). Corruptions are closely related to *data augmentations*, and examples include rotation, translation, and colour shifts. Other images from the data set can then be used to create the negative pairs. This approach to contrastive learning is known as *instance discrimination*.

Section 9.1.3

If, however, we have access to class labels, then we can use images of the same class as positive pairs and images of different classes as negative pairs. This relaxes the reliance on specifying the augmentations that the representation should be invariant to and also avoids treating two semantically similar images as a negative pair. This is referred to as supervised contrastive learning (Khosla *et al.*, 2020) because of the reliance on the class labels, and it can often yield better results than simply learning the representation using cross-entropy classification.

The members of positive and negative pairs do not necessarily have to come from the same data modality. In contrastive-language image pretraining, or CLIP (Radford *et al.*, 2021), a positive pair consists of an image and its corresponding text caption, and two separate functions, one for each modality, are used to map the inputs to the same representation space. Negative pairs are then mismatched images and captions. This is often referred to as *weakly supervised*, as it relies on captioned images, which are often easier to obtain by scraping data from the internet than by manually labelling images with their classes. The loss function in this case is given by

$$E(\mathbf{w}) = -\frac{1}{2}\ln\frac{\exp\{\mathbf{f}_{\mathbf{w}}(\mathbf{x}^+)^{\mathrm{T}}\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{y}^+)\}}{\exp\{\mathbf{f}_{\mathbf{w}}(\mathbf{x}^+)^{\mathrm{T}}\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{y}^+)\} + \sum_{n=1}^{N}\exp\{\mathbf{f}_{\mathbf{w}}(\mathbf{x}_n^-)^{\mathrm{T}}\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{y}^+)\}} - \frac{1}{2}\ln\frac{\exp\{\mathbf{f}_{\mathbf{w}}(\mathbf{x}^+)^{\mathrm{T}}\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{y}^+)\}}{\exp\{\mathbf{f}_{\mathbf{w}}(\mathbf{x}^+)^{\mathrm{T}}\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{y}^+)\} + \sum_{m=1}^{M}\exp\{\mathbf{f}_{\mathbf{w}}(\mathbf{x}^+)^{\mathrm{T}}\mathbf{g}_{\boldsymbol{\theta}}(\mathbf{y}_m^-)\}} \tag{6.21}$$

where $\mathbf{x}^+$ and $\mathbf{y}^+$ represent a positive pair in which $\mathbf{x}$ is an image and $\mathbf{y}$ is its corresponding text caption, $\mathbf{f}_{\mathbf{w}}$ represents the mapping from images to the representation space, and $\mathbf{g}_{\boldsymbol{\theta}}$ is the mapping from text input to the representation space. We also require a set $\{\mathbf{x}_1^-, \ldots, \mathbf{x}_N^-\}$ of other images from the data set, for which we can assume the text caption $\mathbf{y}^+$ is inappropriate, and a set $\{\mathbf{y}_1^-, \ldots, \mathbf{y}_M^-\}$ of text captions that are similarly mismatched to the input image $\mathbf{x}$. The two terms in the loss function ensure that (a) the representation of the image is close to its text caption representation relative to other image representations and (b) the text caption representation is close to the representation of the image it describes relative to other representations of text captions. Although CLIP uses text and image pairs, any data