Lecture 3:
Vision Representation Learners

Announcements

- More information about A1 and project will be coming on Wednesday
- We have confirmed some Google Cloud credits to help with the project (~$50/person for each coupon, may be able to be refreshed), for those who would like to use this.

Today's agenda

- Modern visual representation learning methods
  - Methods that rely only on visual data

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Two major classes of vision foundation models

Representation Learners

Train on huge amounts of data

Objective: learn powerful generalist feature extractors that can be used for downstream computer vision tasks

- Vision Only (e.g. DINOv2)
- Vision-Language (e.g. CLIP)

Generative Models

Train on huge amounts of data

Objective: learn to generate images and/or text

- Text -> Image (e.g. StableDiffusion, DALL-E)
- Text, Image -> Text (e.g. Flamingo, GPT4-V)

Representation learning for images

- **Goal:** learn to extract lower-dimensional feature **representations** from images that effectively capture meaningful semantics in an image and can be used for varied downstream tasks

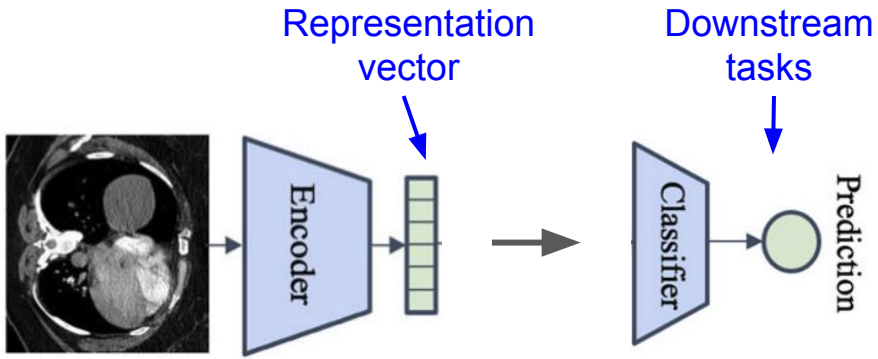


Figure credit: Huang et al. 2021.

Representation learning for images

- **Goal:** learn to extract lower-dimensional feature **representations** from images that effectively capture meaningful semantics in an image and can be used for varied downstream tasks

- Another common term for these representations is "**embeddings**", and the neural networks that produce the representations are often also referred to as "**embedding models**" or "**encoders**"

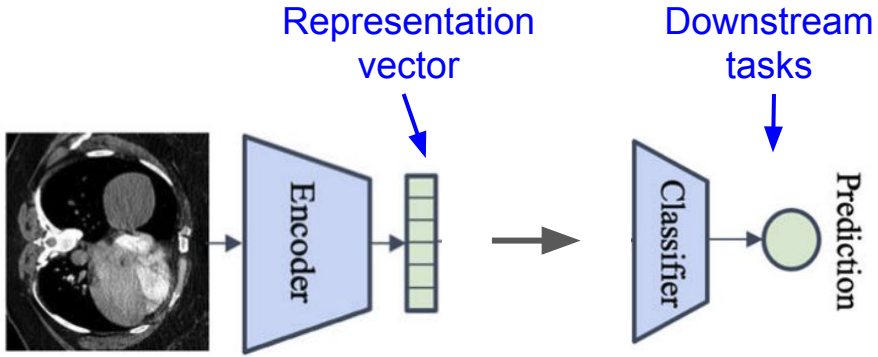
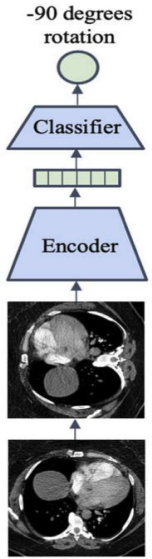


Figure credit: Huang et al. 2021.

Different representation learning paradigms

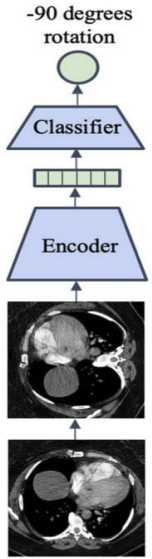
Different representation learning paradigms



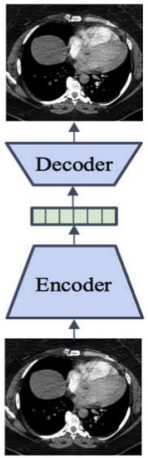
Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

Figure credit: Huang et al. 2023.

Different representation learning paradigms



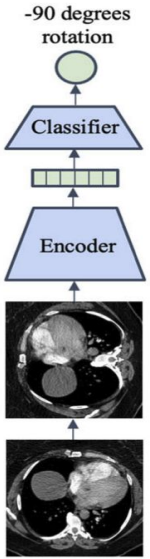
Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image



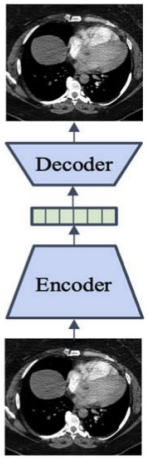
Generative objective
Compress and then reconstruct input image (e.g. autoencoders)

Figure credit: Huang et al. 2023.

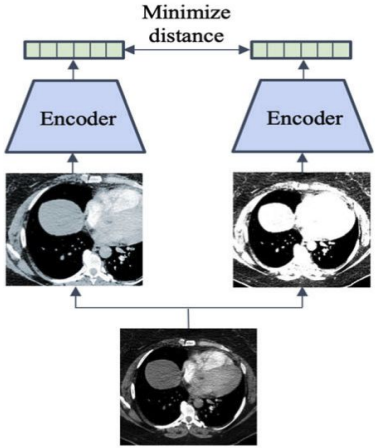
Different representation learning paradigms



Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image



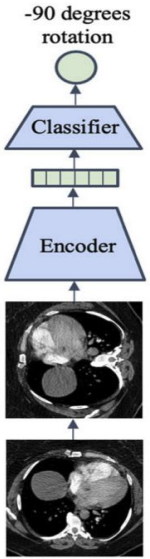
Generative objective
Compress and then reconstruct input image (e.g. autoencoders)



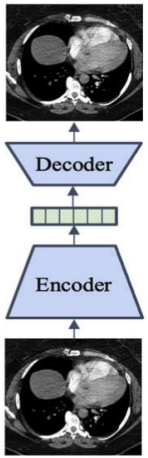
Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

Figure credit: Huang et al. 2023.

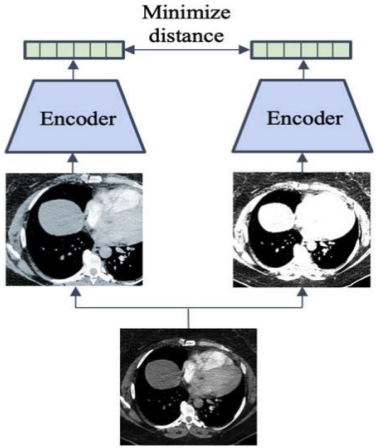
Different representation learning paradigms



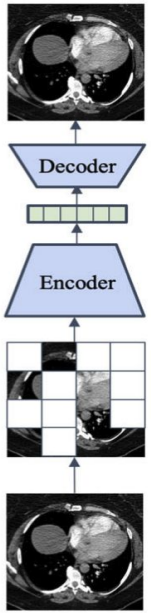
Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image



Generative objective
Compress and then reconstruct input image (e.g. autoencoders)



Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input



Self-prediction objective
Mask parts of input data and predict these parts

Figure credit: Huang et al. 2023.

Different representation learning paradigms

- Note that these are all examples of self-supervised learning (SSL): there are no explicit, externally provided training labels from humans. Instead, training labels are generated by the inherent structure or properties of the data.

- Although supervised learning also produces learned representations, SSL is dominant for representation learning where the goal is subsequent transfer learning to varied downstream tasks. It typically achieves higher transfer learning performance, and can learn from very large amounts of diverse, unlabeled data.

Innate **objective**

E.g., predict rotation angle (or some other innate property) of an image

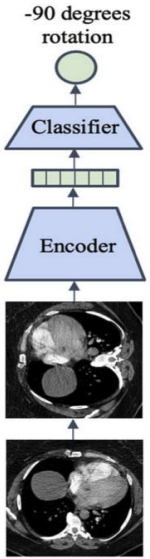
Compress and then reconstruct input image (e.g. autoencoders)

Different views of the same input should have more similar representation to each other than with a different input

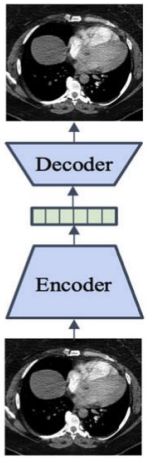
Mask parts of input data and predict these parts

Figure credit: Huang et al. 2023.

Different representation learning paradigms

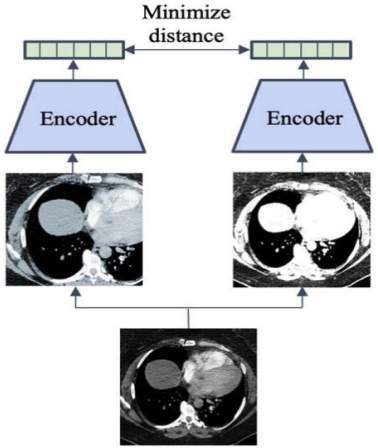


Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

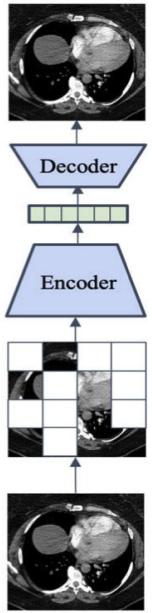


Generative objective
Compress and then reconstruct input image (e.g. autoencoders)

Popular state-of-the-art approaches



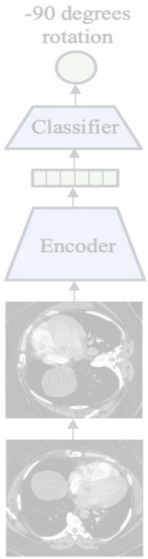
Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input



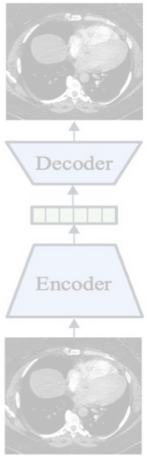
Self-prediction objective
Mask parts of input data and predict these parts

Figure credit: Huang et al. 2023.

Different representation learning paradigms

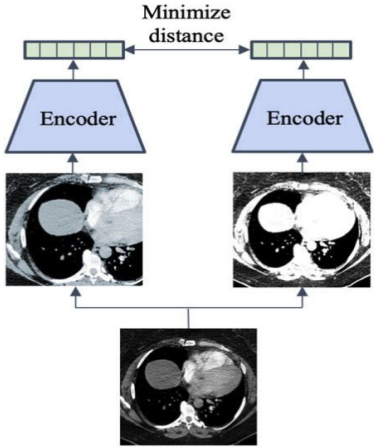


Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

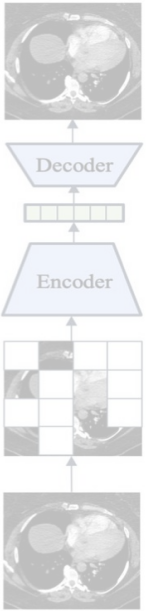


Generative objective
Compress and then reconstruct input image (e.g. autoencoders)

Popular state-of-the-art approaches



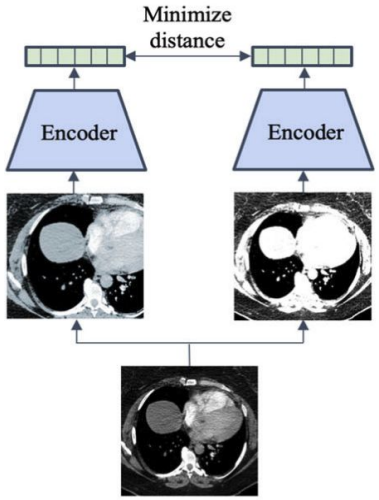
Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input



Self-prediction objective
Mask parts of input data and predict these parts

Figure credit: Huang et al. 2023.

SimCLR: a foundational method for contrastive self-supervised learning

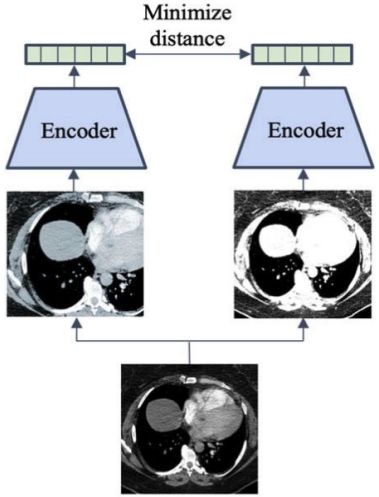


Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input

SimCLR: "Simple Framework for Contrastive Learning of Visual Representations"

Chen et al. 2020

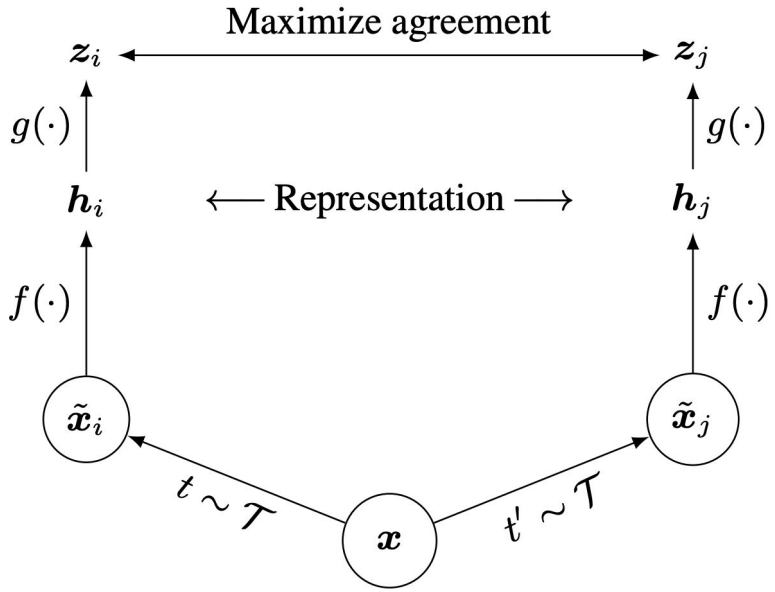
SimCLR: a foundational method for contrastive self-supervised learning



Contrastive objective

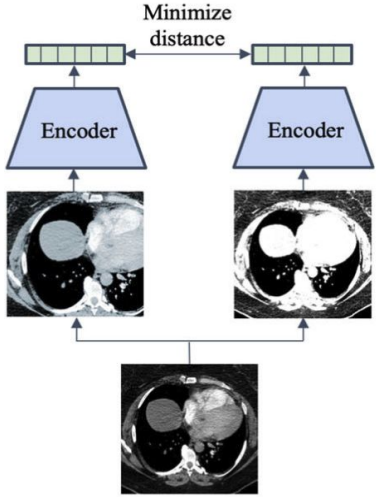
Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation



Chen et al. 2020

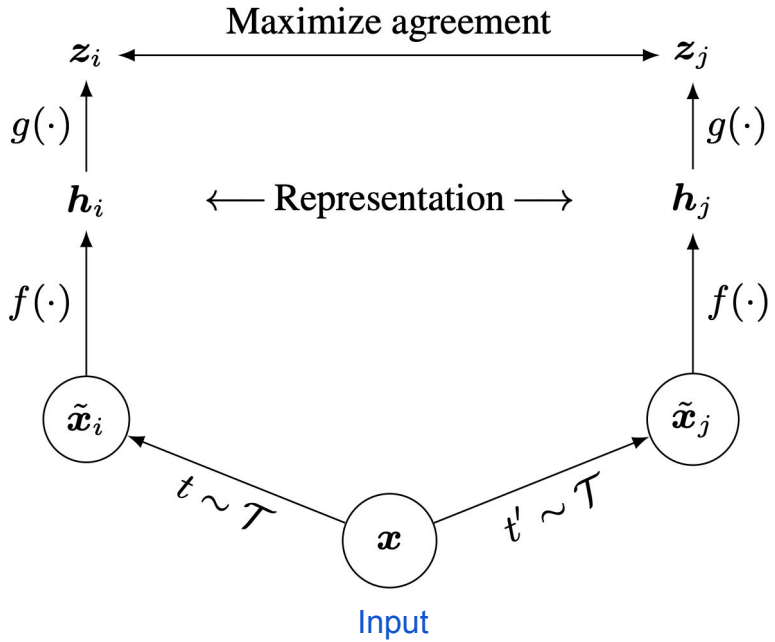
SimCLR: a foundational method for contrastive self-supervised learning



Contrastive objective

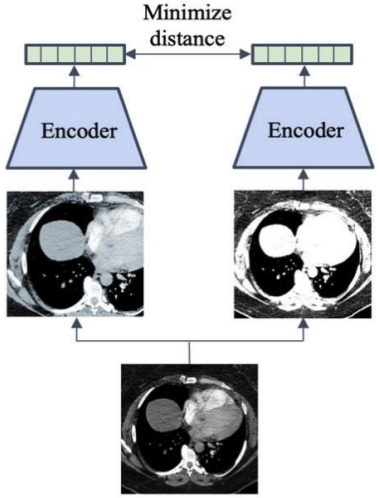
Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation



Chen et al. 2020

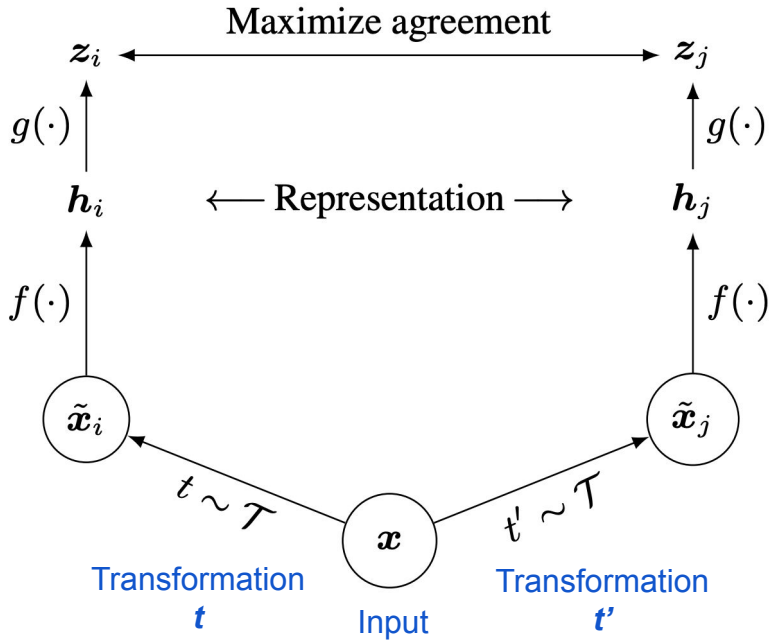
SimCLR: a foundational method for contrastive self-supervised learning



Contrastive objective

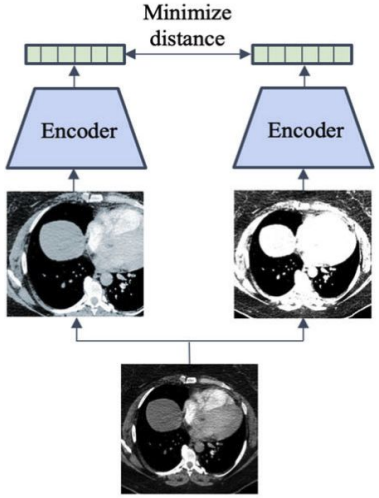
Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation



Chen et al. 2020

SimCLR: a foundational method for contrastive self-supervised learning

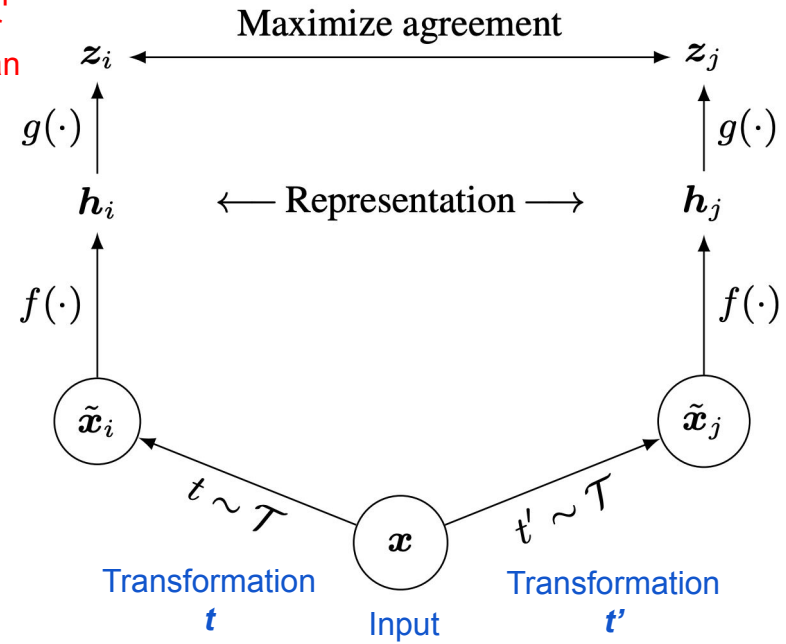


Contrastive objective

Different views of the same input should have more similar representation to each other than with a different input

Transformation set: random crop (w/ flip and resize), color distortion, Gaussian blur

SimCLR formulation

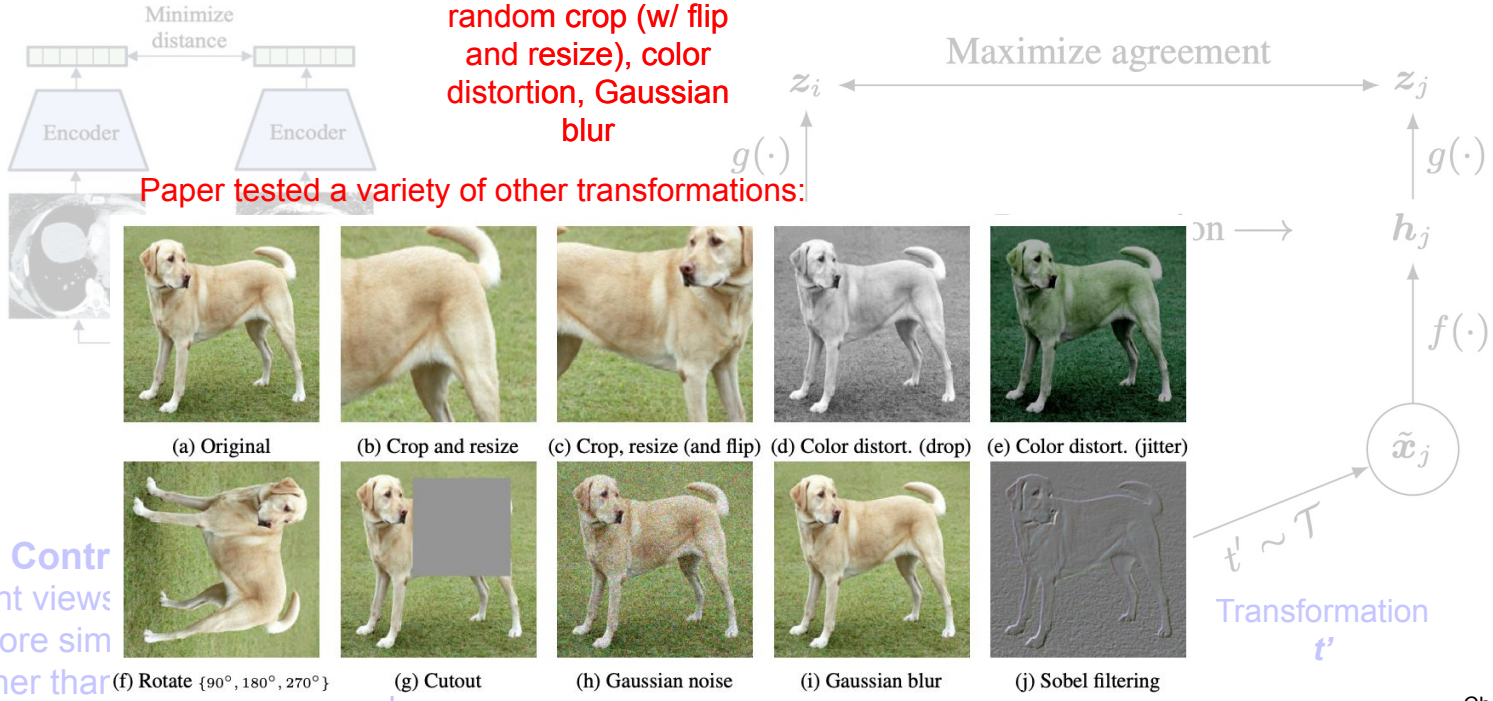


Chen et al. 2020

SimCLR: a foundational method for contrastive self-supervised learning

Transformation set: random crop (w/ flip and resize), color distortion, Gaussian blur

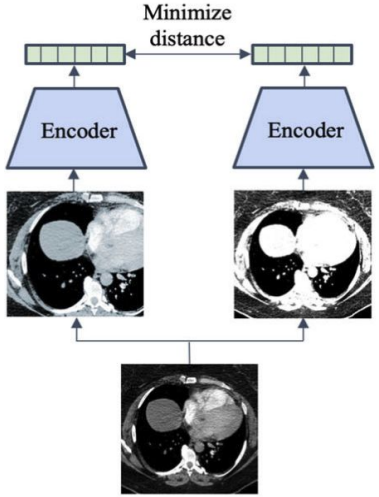
Paper tested a variety of other transformations:



Contr
Different views
have more sim
other than

Chen et al. 2020

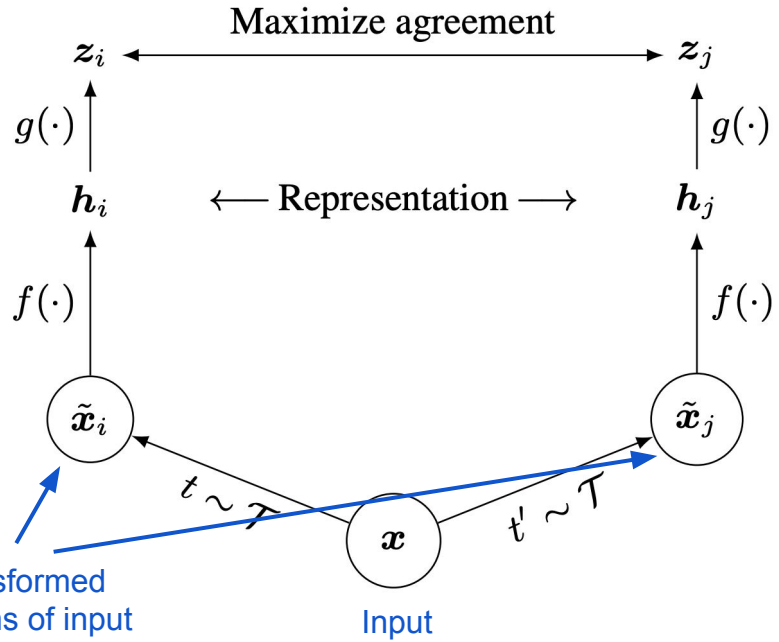
SimCLR: a foundational method for contrastive self-supervised learning



Contrastive objective

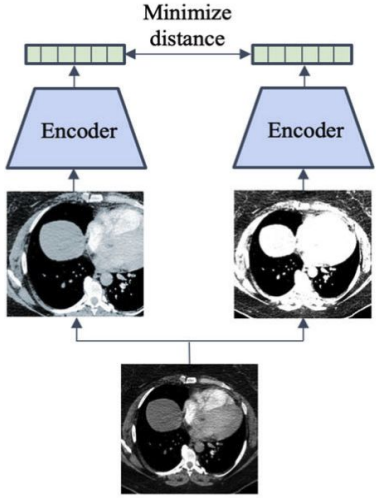
Different views of the same input should have more similar representation to each other than with a different input

SimCLR formulation



Chen et al. 2020

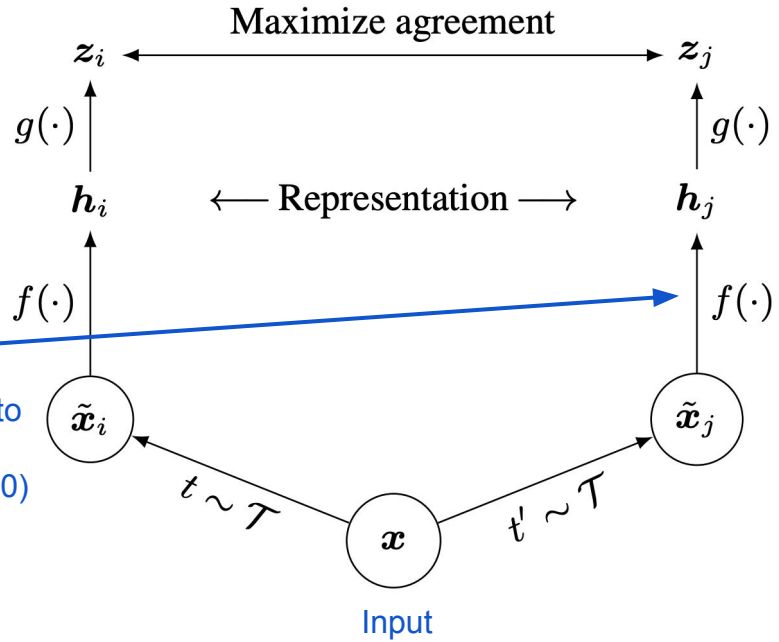
SimCLR: a foundational method for contrastive self-supervised learning



Contrastive objective

Different views of the same input should have more similar representation to each other than with a different input

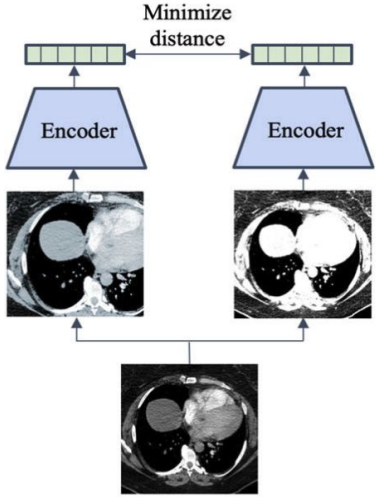
SimCLR formulation



Encoder network f (same model applied to both transformed images, e.g. ResNet-50)

Chen et al. 2020

SimCLR: a foundational method for contrastive self-supervised learning

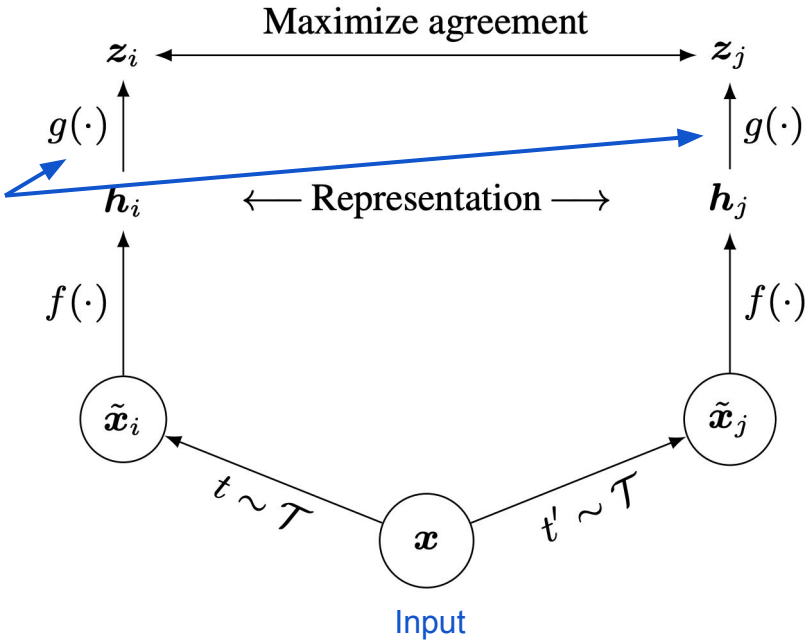


Contrastive objective

Different views of the same input should have more similar representation to each other than with a different input

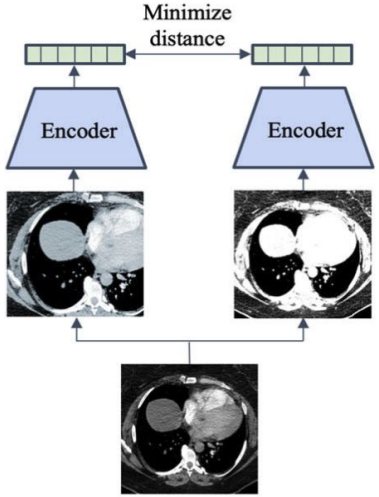
SimCLR formulation

Projection head (MLP w/ one hidden layer), same network applied to both representations h_i, h_j



Chen et al. 2020

SimCLR: a foundational method for contrastive self-supervised learning

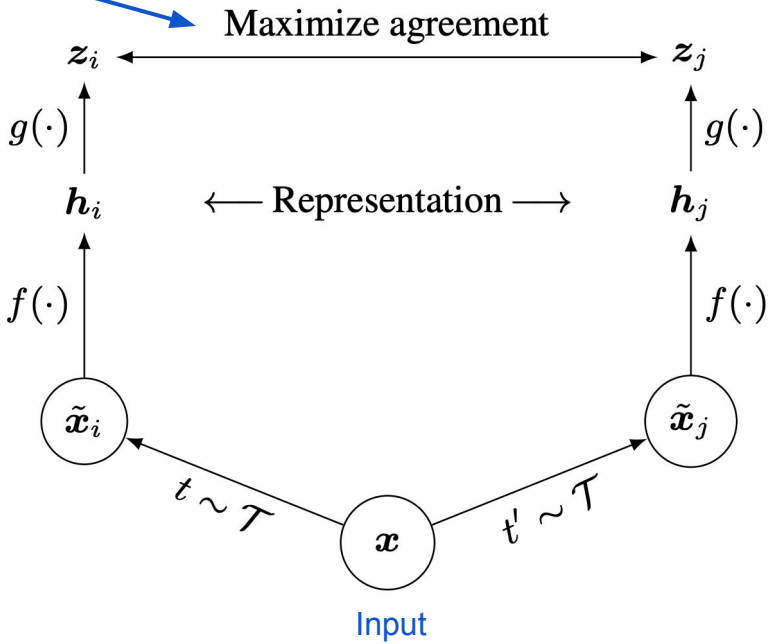


Contrastive objective

Different views of the same input should have more similar representation to each other than with a different input

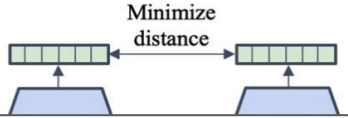
Contrastive loss

SimCLR formulation



Chen et al. 2020

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

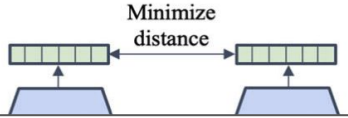
Maximize agreement

z_i z_j

Contrastive loss can take the form of a familiar cross-entropy loss!

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

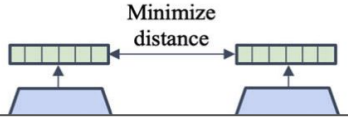
z_i z_j

Contrastive loss can take the form of a familiar cross-entropy loss!

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

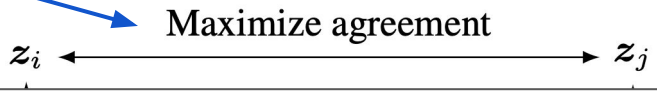
Compute loss over a minibatch of N examples. Generate two augmented views of each example, resulting in 2N data points total. Now in the contrastive loss, we wish for a pair of data points (i,j) corresponding to augmentations of the same example to have closer representation similarity than with other data points generated from different examples.

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation



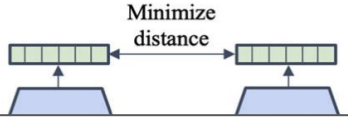
Contrastive loss can take the form of a familiar cross-entropy loss!

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Compute loss over a minibatch of N examples. Generate two augmented views of each example, resulting in 2N data points total. Now in the contrastive loss, we wish for a pair of data points (i,j) corresponding to augmentations of the same example to have closer representation similarity than with other data points generated from different examples.

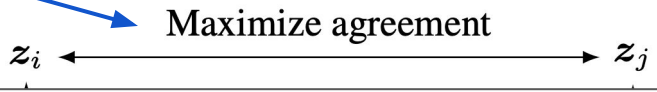
Use a cross-entropy formulation: given data point i, similarity with data point j should have higher score than with all other points, such that it is “correctly classified”!

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation



Contrastive loss can take the form of a familiar cross-entropy loss!

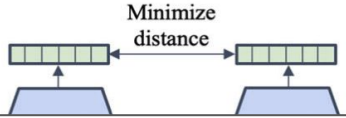
Loss for a pair of data points (i,j)

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Compute loss over a minibatch of N examples. Generate two augmented views of each example, resulting in 2N data points total. Now in the contrastive loss, we wish for a pair of data points (i,j) corresponding to augmentations of the same example to have closer representation similarity than with other data points generated from different examples.

Use a cross-entropy formulation: given data point i, similarity with data point j should have higher score than with all other points, such that it is “correctly classified”!

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

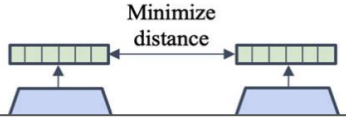
z_i z_j

Contrastive loss can take the form of a familiar cross-entropy loss!

Similarity score between final-layer representations of i and j

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

z_i ⟷ z_j

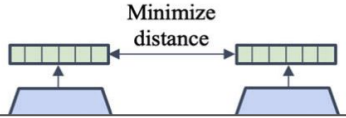
Contrastive loss can take the form of a familiar cross-entropy loss!

Similarity score between final-layer representations of i and j

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Use cosine similarity $\text{sim}(\boldsymbol{u}, \boldsymbol{v}) = \boldsymbol{u}^\top \boldsymbol{v} / \|\boldsymbol{u}\| \|\boldsymbol{v}\|$

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

z_i ⟷ z_j

Contrastive loss can take the form of a familiar cross-entropy loss!

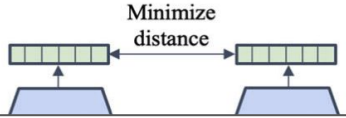
Similarity score between final-layer representations of i and j

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

From here, looks very similar to softmax loss (generalized cross entropy to multiple classes)

$$L_{Softmax} = \frac{1}{M} \sum_i -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

z_i z_j

Contrastive loss can take the form of a familiar cross-entropy loss!

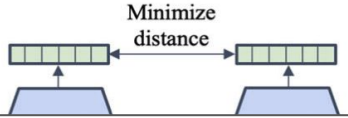
Exponentiate

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

From here, looks very similar to softmax loss (generalized cross entropy to multiple classes)

$$L_{Softmax} = \frac{1}{M} \sum_i -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

z_i z_j

Contrastive loss can take the form of a familiar cross-entropy loss!

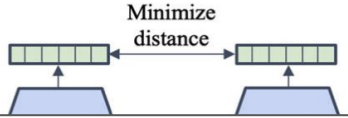
$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Detail: Loss uses a temperature hyperparameter, controls peakiness of final probability distribution for better learning dynamics

From here, looks very similar to softmax loss (generalized cross entropy to multiple classes)

$$L_{Softmax} = \frac{1}{M} \sum_i -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

z_i z_j

Contrastive loss can take the form of a familiar cross-entropy loss!

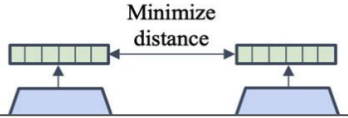
$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Normalize over scores of similarity between i and all other data points in the minibatch (2N total)

From here, looks very similar to softmax loss (generalized cross entropy to multiple classes)

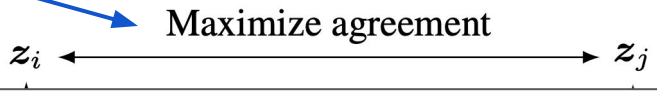
$$L_{Softmax} = \frac{1}{M} \sum_i -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation



Contrastive loss can take the form of a familiar cross-entropy loss!

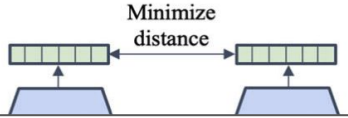
Negative log likelihood, as in softmax / cross-entropy

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

From here, looks very similar to softmax loss (generalized cross entropy to multiple classes)

$$L_{Softmax} = \frac{1}{M} \sum_i -\log\left(\frac{e^{s_{y_i}}}{\sum_j e^{s_j}}\right)$$

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation

Maximize agreement

z_i z_j

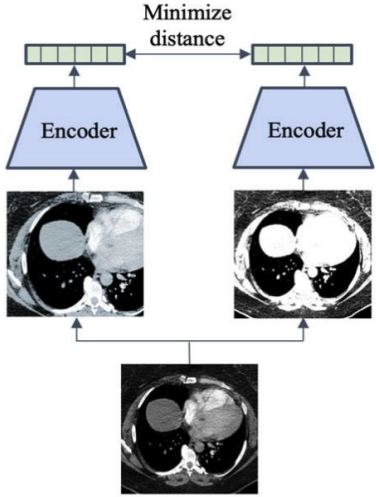
Contrastive loss can take the form of a familiar cross-entropy loss!

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

For a minibatch of N examples (2N augmented data points), compute this loss over all corresponding pairs (i,j), as well as (j,i) for symmetry of the loss, and then average these individual loss terms (2N terms total)

$$\mathcal{L} = \frac{1}{2N} \sum_{k=1}^{N} [\ell(2k-1, 2k) + \ell(2k, 2k-1)]$$

SimCLR: a foundational method for contrastive self-supervised learning

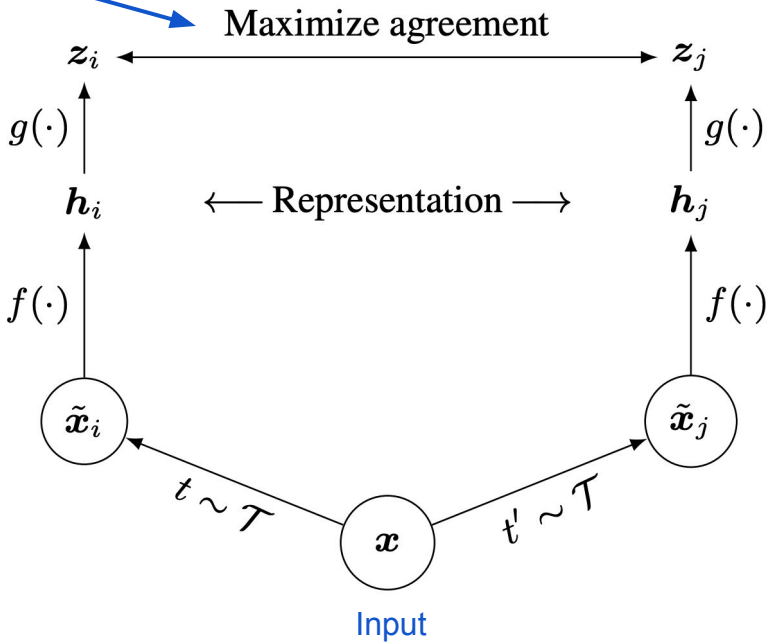


Contrastive objective

Different views of the same input should have more similar representation to each other than with a different input

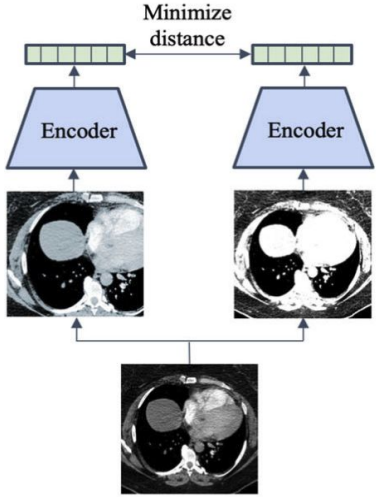
Contrastive loss

SimCLR formulation



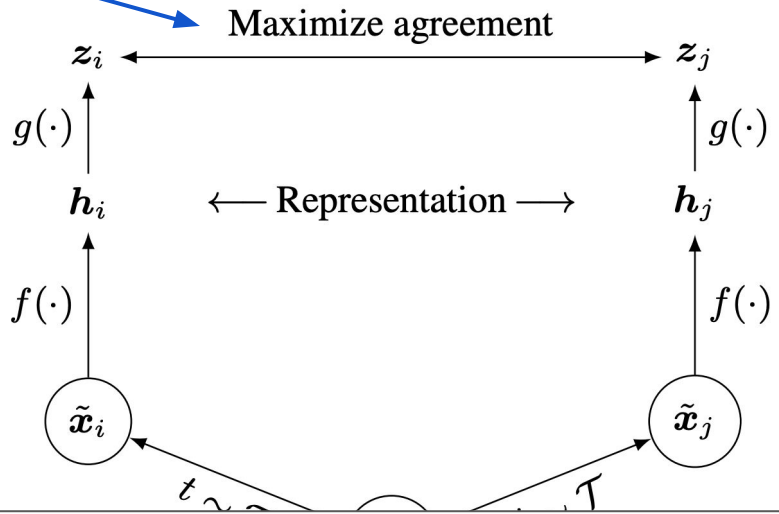
Chen et al. 2020

SimCLR: a foundational method for contrastive self-supervised learning



Contrastive loss

SimCLR formulation



After self-supervised training, can fine-tune the encoder *f* on smaller labeled datasets. Can also directly extract learned representations h for downstream tasks.

Quantitative evaluation of SimCLR

- SimCLR significantly outperforms a supervised baseline on ImageNet when labels are provided for only 1% or 10% of images

Method	Architecture	Label fraction (Top 5) 1%	10%
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(∗)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

Chen et al. 2020

Quantitative evaluation of SimCLR

- SimCLR significantly outperforms a supervised baseline on ImageNet when labels are provided for only 1% or 10% of images
- SimCLR also outperforms the supervised baseline in transfer learning using fine-tuning across different datasets

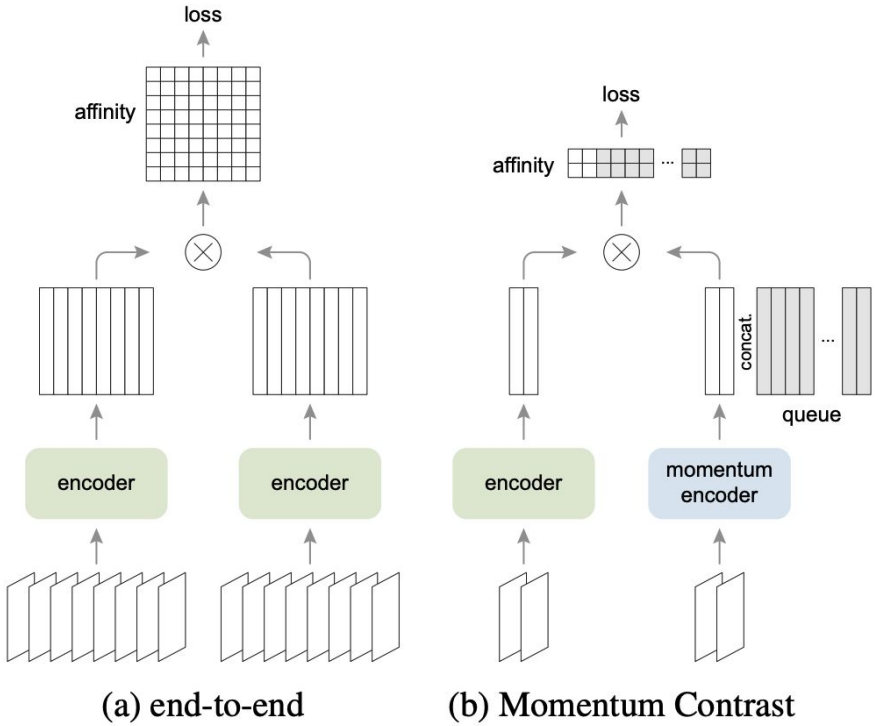
Method	Architecture	Label fraction (Top 5) 1%	10%
Supervised baseline	ResNet-50	48.4	80.4
Methods using other label-propagation:			
Pseudo-label	ResNet-50	51.6	82.4
VAT+Entropy Min.	ResNet-50	47.0	83.4
UDA (w. RandAug)	ResNet-50	-	88.5
FixMatch (w. RandAug)	ResNet-50	-	89.1
S4L (Rot+VAT+En. M.)	ResNet-50 (4×)	-	91.2
Methods using representation learning only:			
InstDisc	ResNet-50	39.2	77.4
BigBiGAN	RevNet-50 (4×)	55.2	78.8
PIRL	ResNet-50	57.2	83.8
CPC v2	ResNet-161(*)	77.9	91.2
SimCLR (ours)	ResNet-50	75.5	87.8
SimCLR (ours)	ResNet-50 (2×)	83.0	91.2
SimCLR (ours)	ResNet-50 (4×)	**85.8**	**92.6**

	Food	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
SimCLR (ours)	**76.9**	**95.3**	80.2	48.4	**65.9**	60.0	61.2	**84.2**	**78.9**	89.2	**93.9**	**95.0**
Supervised	75.2	**95.7**	**81.2**	**56.4**	64.9	**68.8**	**63.8**	83.8	**78.7**	**92.3**	**94.1**	94.2
Fine-tuned:												
SimCLR (ours)	**89.4**	**98.6**	**89.0**	**78.2**	**68.1**	**92.1**	**87.0**	**86.6**	**77.8**	92.1	**94.1**	97.6
Supervised	88.7	98.3	**88.7**	**77.8**	67.0	91.4	**88.0**	86.5	**78.8**	**93.2**	**94.2**	**98.0**
Random init	88.3	96.0	81.9	**77.0**	53.7	91.3	84.8	69.4	64.1	82.7	72.5	92.5

Chen et al. 2020

MoCo: Alleviates the batch size limitation of SimCLR

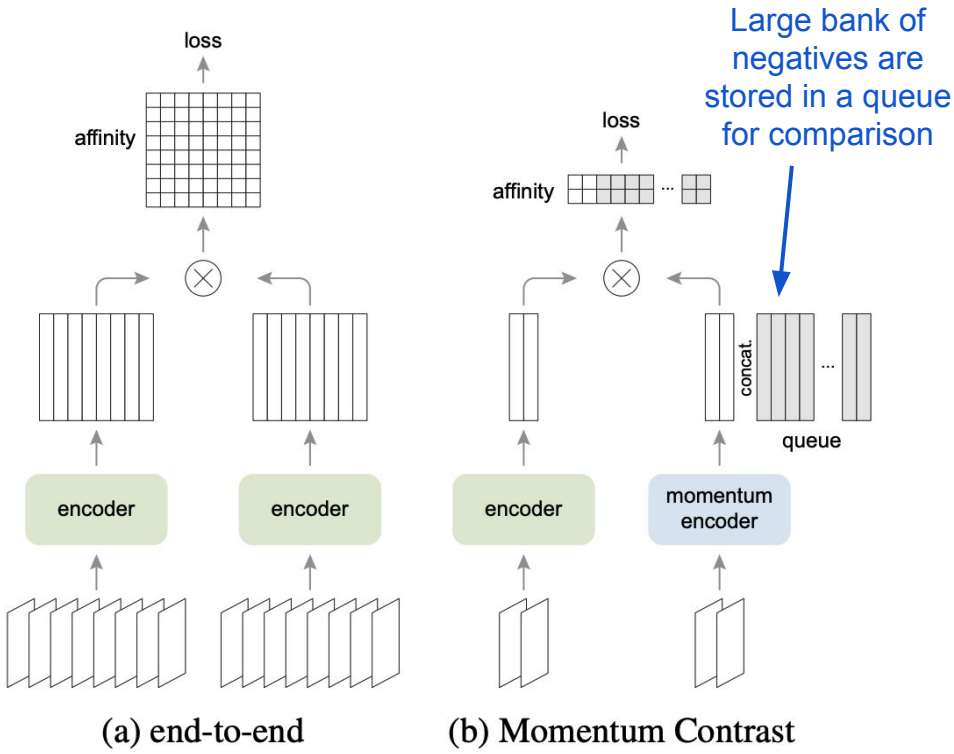
- SimCLR's relies on a large batch size to generate a sufficient number of negative pairs for effective contrastive learning. This creates significant computational and memory burden.
- MoCo (Momentum Contrast) and MoCo v2, v3 alleviates this by using a momentum-updated queue that allows incorporating many negative pairs without increasing batch size.



He et al. 2020
Chen et al. 2020

MoCo: Alleviates the batch size limitation of SimCLR

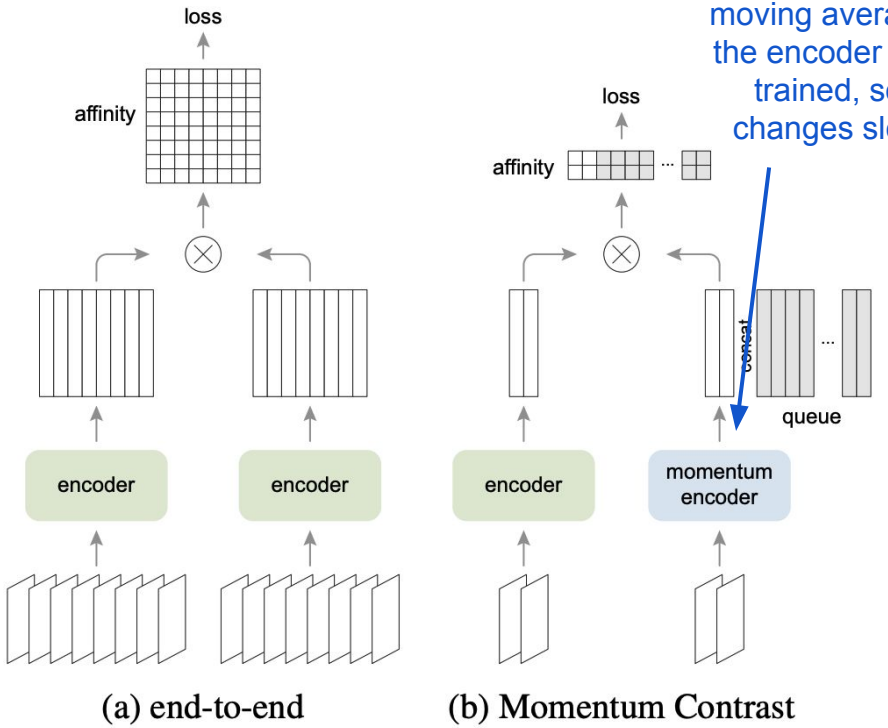
- SimCLR's relies on a large batch size to generate a sufficient number of negative pairs for effective contrastive learning. This creates significant computational and memory burden.
- MoCo (Momentum Contrast) and MoCo v2, v3 alleviates this by using a momentum-updated queue that allows incorporating many negative pairs without increasing batch size.



He et al. 2020
Chen et al. 2020

MoCo: Alleviates the batch size limitation of SimCLR

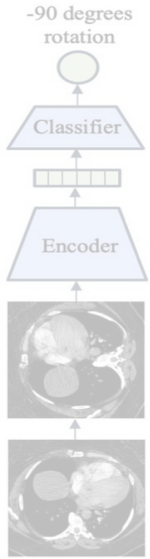
- SimCLR's relies on a large batch size to generate a sufficient number of negative pairs for effective contrastive learning. This creates significant computational and memory burden.
- MoCo (Momentum Contrast) and MoCo v2, v3 alleviates this by using a momentum-updated queue that allows incorporating many negative pairs without increasing batch size.



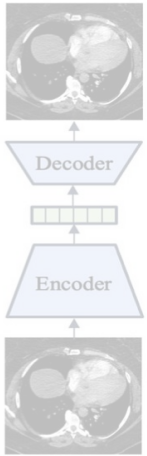
Momentum encoder is updated as an exponential moving average of the encoder being trained, so it changes slowly

He et al. 2020
Chen et al. 2020

Major image-based representation learning paradigms

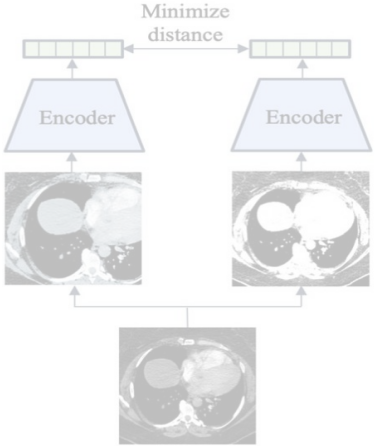


Innate relationship objective
E.g., predict rotation angle (or some other innate property) of an image

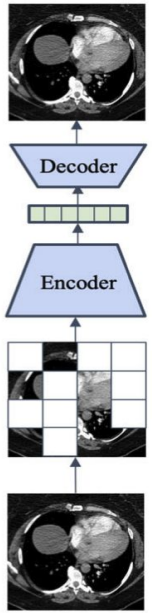


Generative objective
Compress and then reconstruct input image (e.g., variational autoencoders)

State-of-the-art approaches typically fall into these categories



Contrastive objective
Different views of the same input should have more similar representation to each other than with a different input



Self-prediction objective
Mask parts of input data and predict these parts

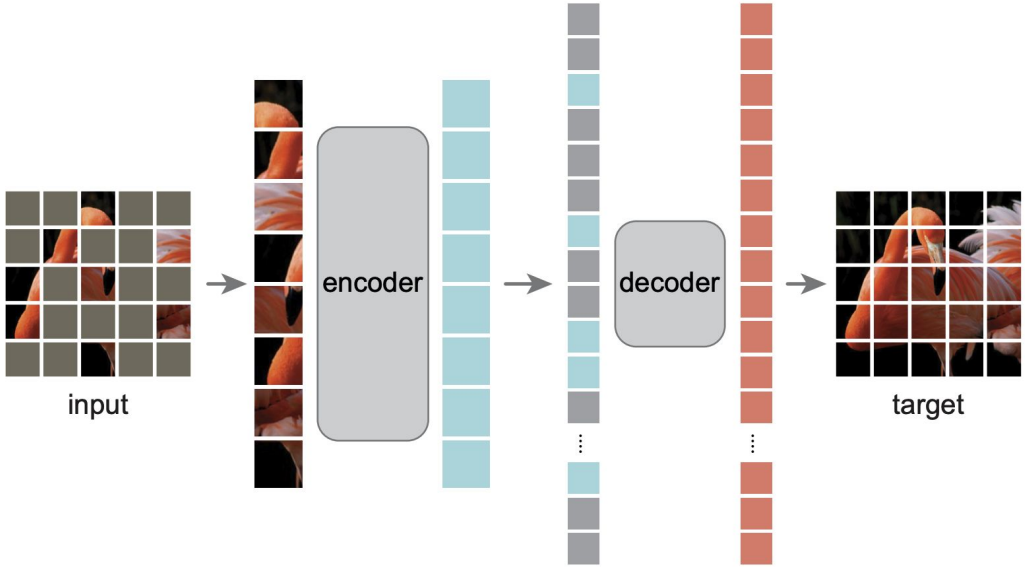
Figure credit: Huang et al. 2021.

Masked Autoencoders (MAE)

- Key idea: mask substantial parts of the input, train the model to reconstruct (predict) these parts
- Inspired by major self-supervised representation learning paradigm in NLP (e.g. BERT), that masks tokens in sentences and trains models to reconstruct them
- Intuition: Transformer architecture is well-suited to this objective

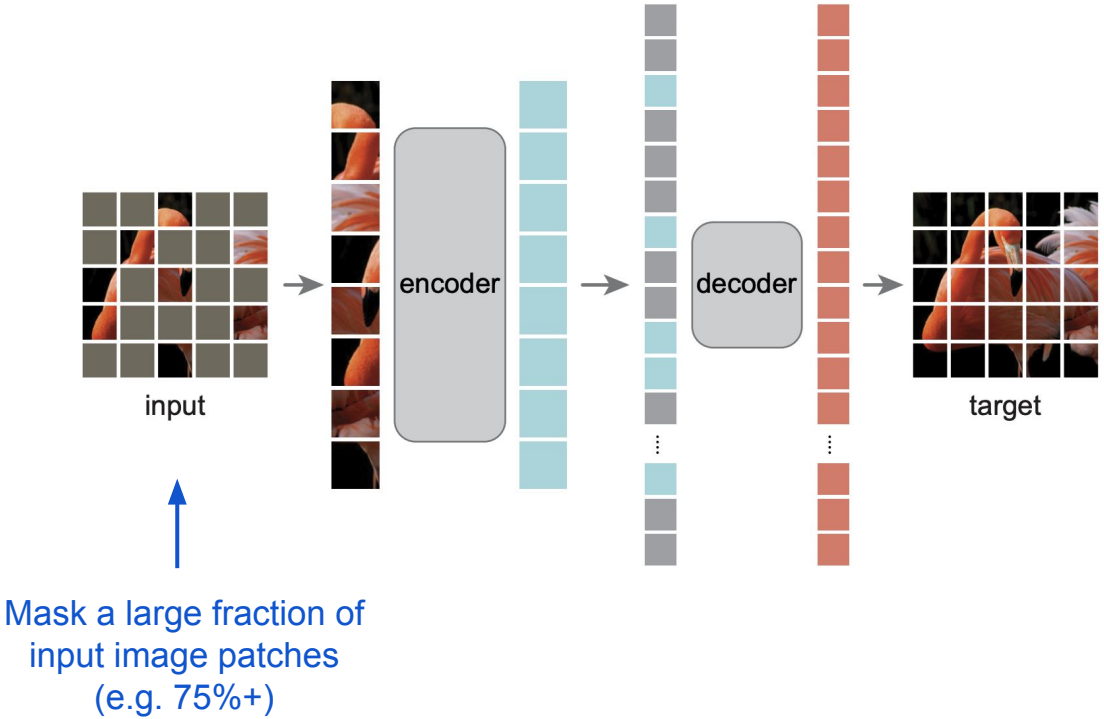
He et al. 2021

Masked Autoencoders (MAE)



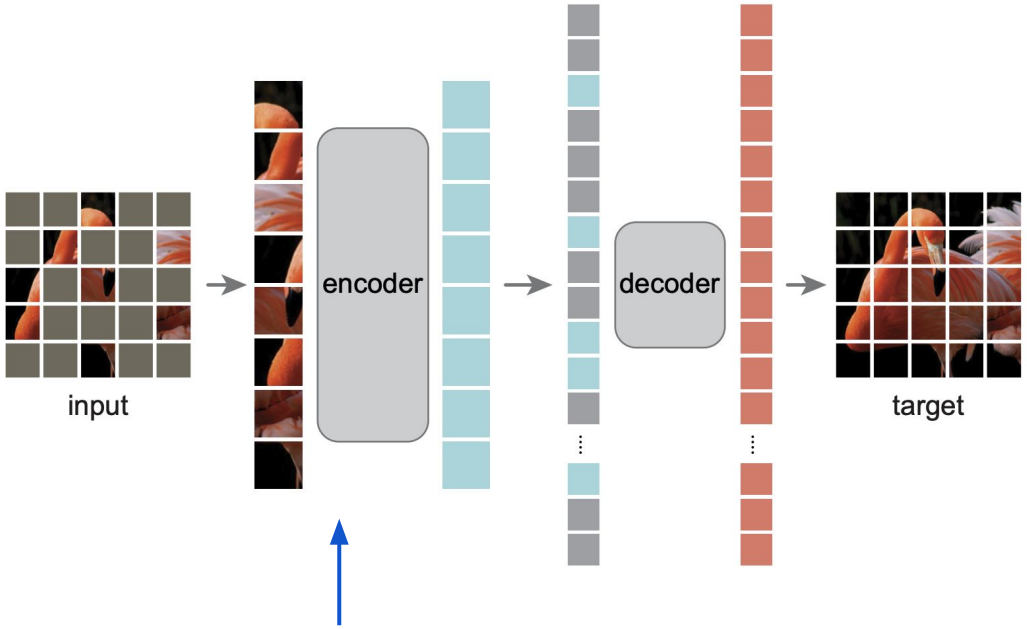
He et al. 2021

Masked Autoencoders (MAE)



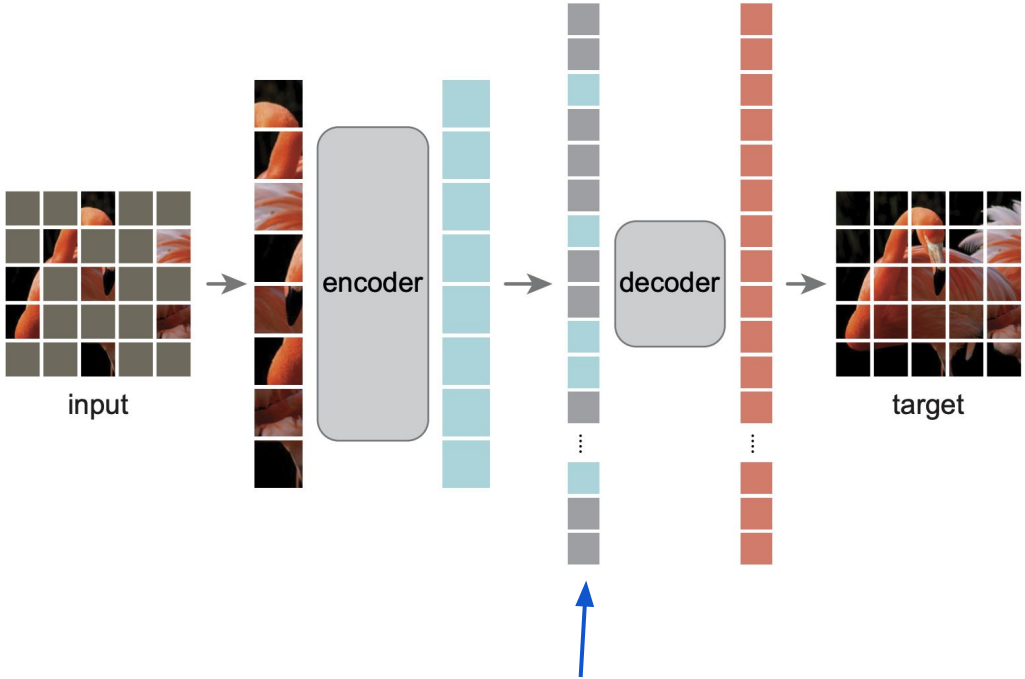
He et al. 2021

Masked Autoencoders (MAE)



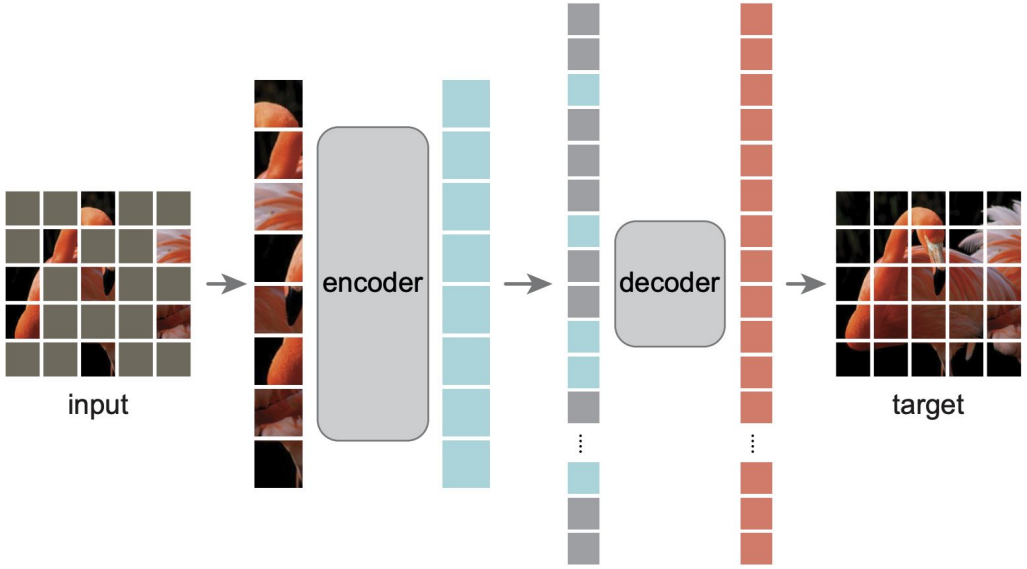
He et al. 2021

Masked Autoencoders (MAE)



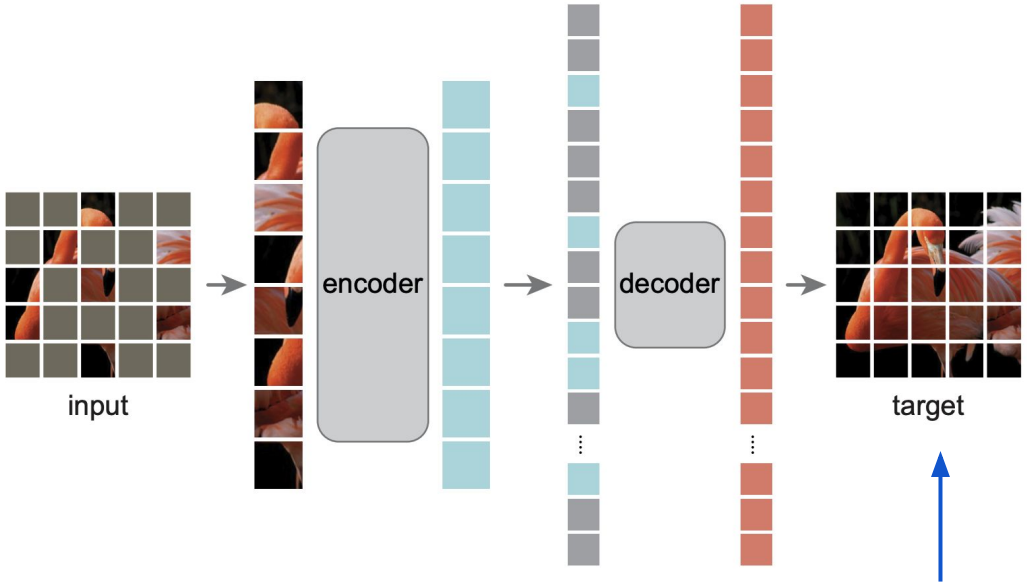
He et al. 2021

Masked Autoencoders (MAE)



He et al. 2021

Masked Autoencoders (MAE)



MAE is trained with a reconstruction loss, MSE (means squared error) between output and target image

He et al. 2021

Comparing these SSL methods

- Comparison often requires careful inspection of the reported results across different settings, and there can be subtle differences.
- For example, MoCo v3 has higher linear probing accuracy than MAE (on an evaluation task, attach one learned linear layer to the encoder output), but MAE gets stronger results when fine-tuning a few last layers of the encoder.
- Suggests that MAE representations are less linearly separable, but are stronger non-linear features and could perform better under fine-tuning of a non-linear head in transfer learning.

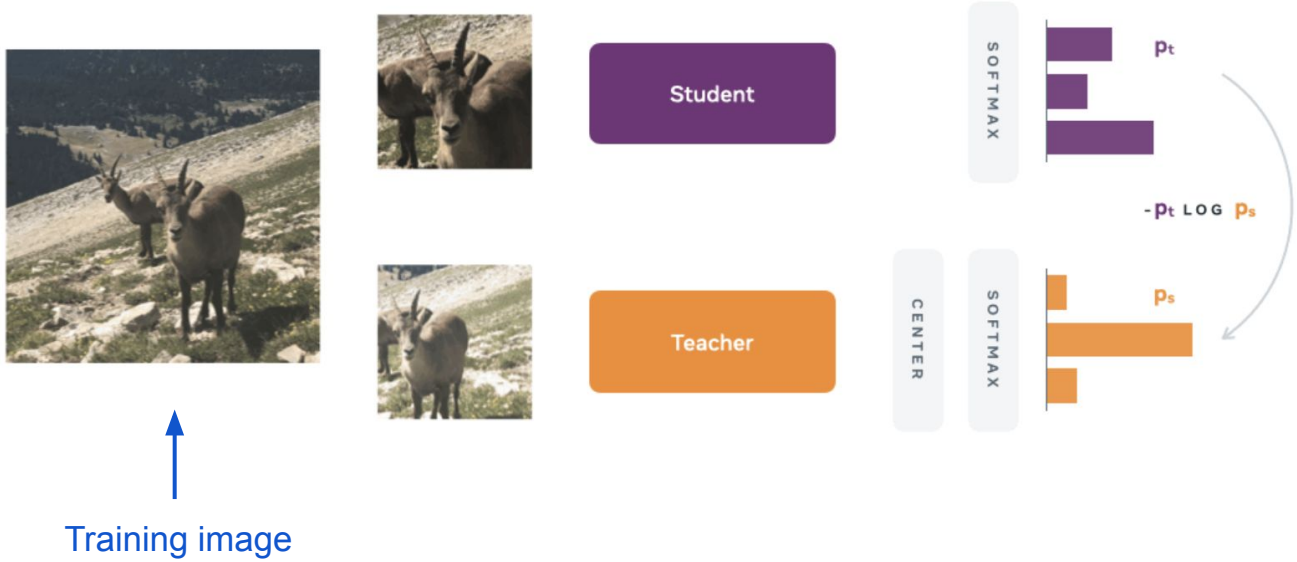
He et al. 2020
Chen et al. 2020

One more SSL approach: DINO (Self-**Di**stillation with **No** Labels)

- Not an example of a contrastive learning objective for self-supervised learning, but related!
- Builds on the notion of matching representations of augmented views of the same image, but no longer uses negative samples. Instead, proposes a **teacher-student framework**.
- A slowly updated, stable teacher model generates “soft labels” for representations from unlabeled data, and a student model is trained to match these over different augmentations, encouraging the student to learn meaningful, invariant representations.
- Uses ViT as the student and teacher model architectures.

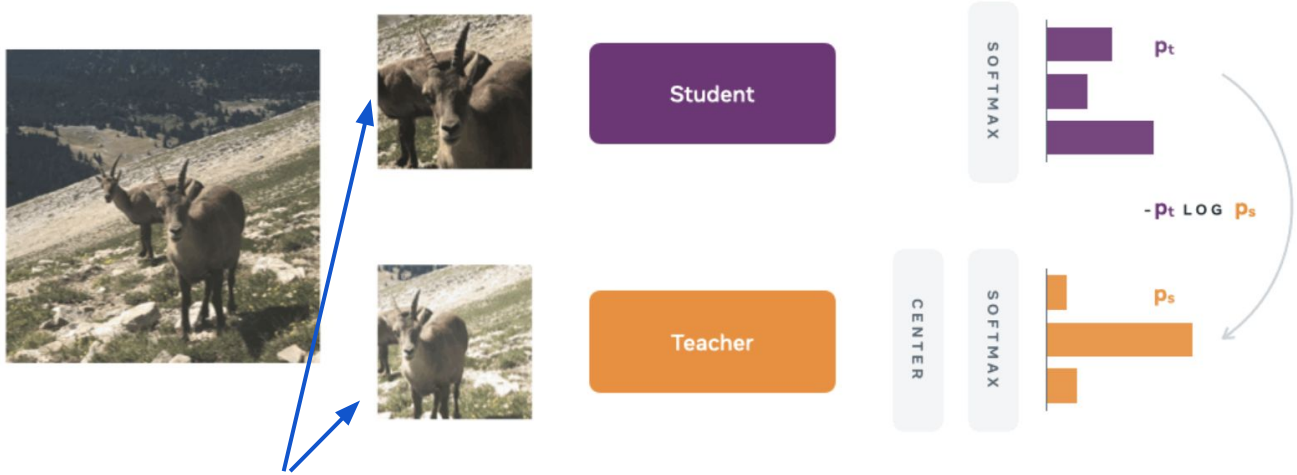
Caron et al. 2021
Oquab et al. 2024

DINO: Self-**Di**stillation with **No** Labels



Caron et al. 2021
Oquab et al. 2024

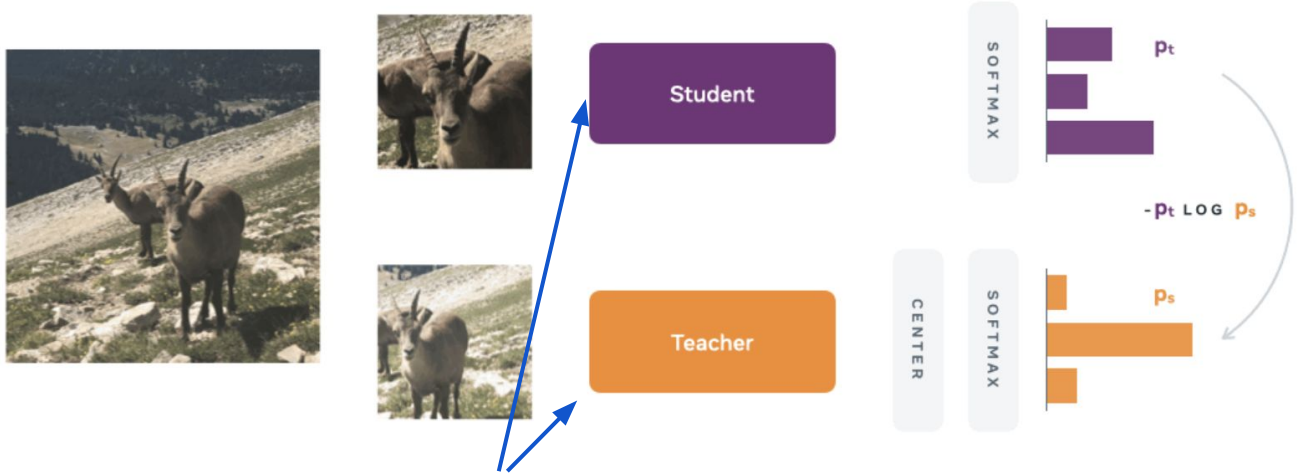
DINO: Self-**Di**stillation with **No** Labels



Augmented “views” of the input are passed to both the student and teacher networks.
Student views: more diverse and aggressive augmentations, to learn representations that generalize across augmentations
Teacher views: less aggressive augmentations (including larger crops), to provide a more stable target for the student to follow.

Caron et al. 2021
Oquab et al. 2024

DINO: Self-**Di**stillation with **No** Labels

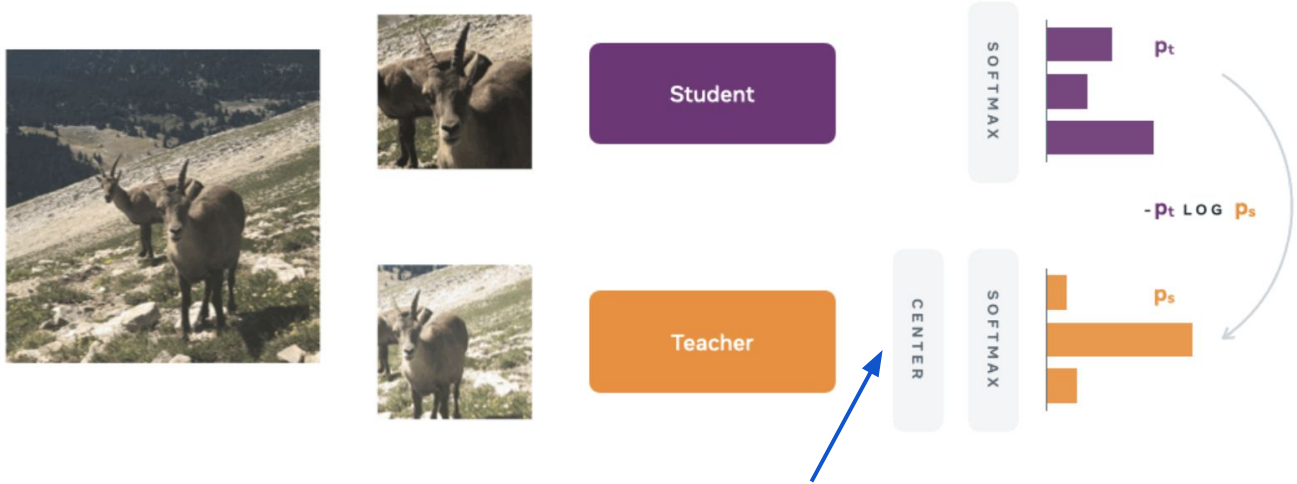


Student network (ViT): actively trained, will produce the final representations once trained

Teacher network: not directly trained with backpropagation. Instead, updated as an exponential moving average of the student's parameters, to provide a stable target for the student

Caron et al. 2021
Oquab et al. 2024

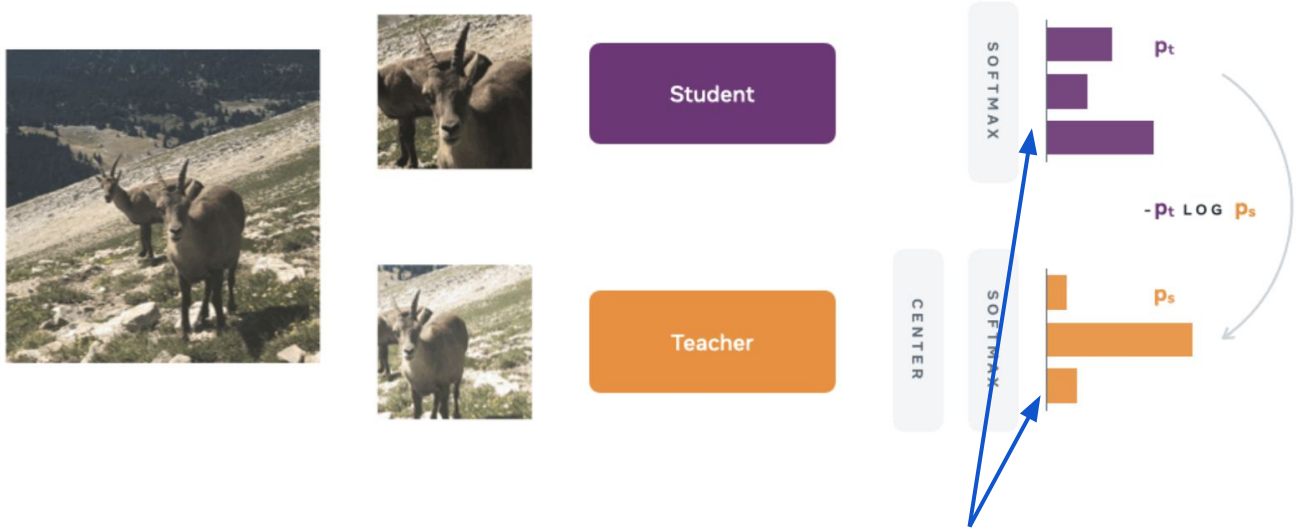
DINO: Self-**Di**stillation with **No** Labels



Teacher output logits are **centered** by subtracting running mean of teacher outputs. Ensures output distribution remains well-balanced and does not collapse to a single or a few classes.

Caron et al. 2021
Oquab et al. 2024

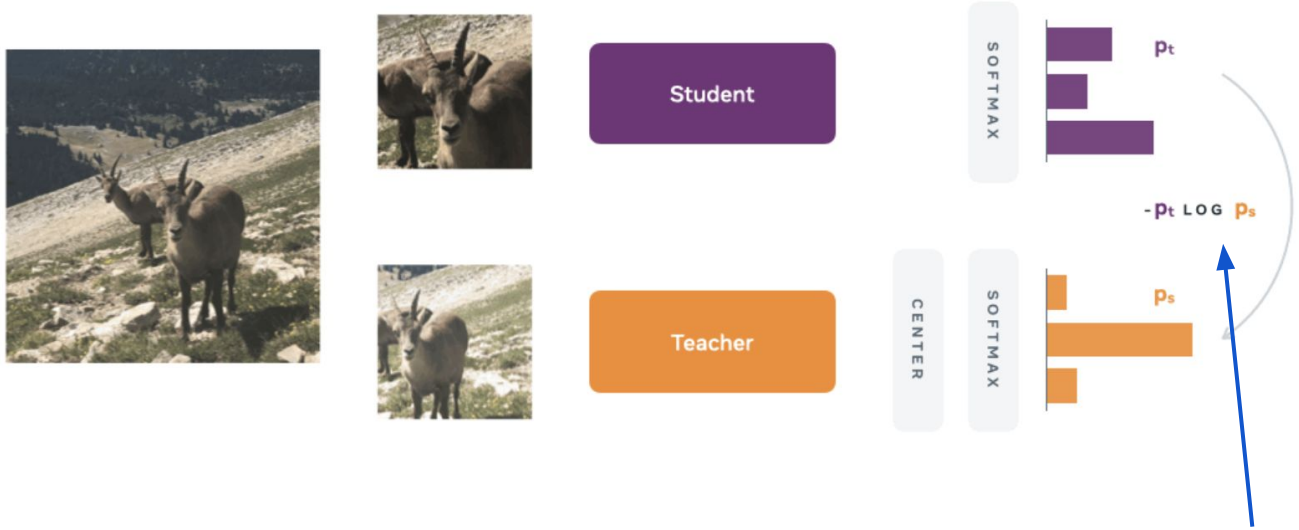
DINO: Self-**Di**stillation with **No** Labels



Student and teacher output softmax-normalized probabilities for image features

Caron et al. 2021
Oquab et al. 2024

DINO: Self-**Di**stillation with **No** Labels

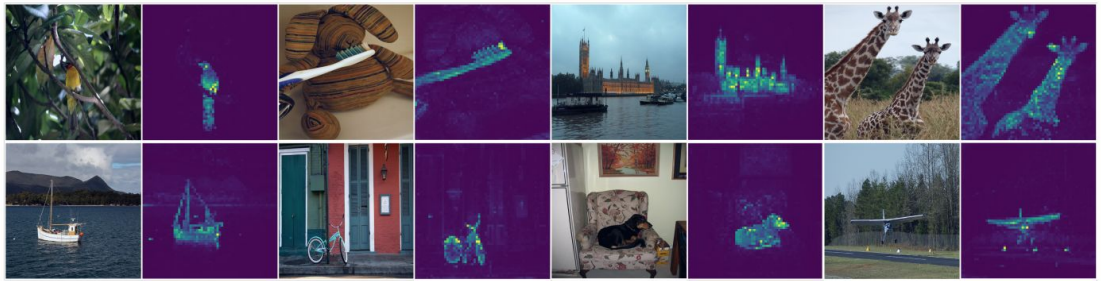


Cross-entropy loss compares the two probability distributions

Caron et al. 2021
Oquab et al. 2024

DINO demonstrates strong features for dense tasks such as segmentation

- Visualizing self-attention maps from the last encoder layer shows strong class-specific information that can be used for object segmentation

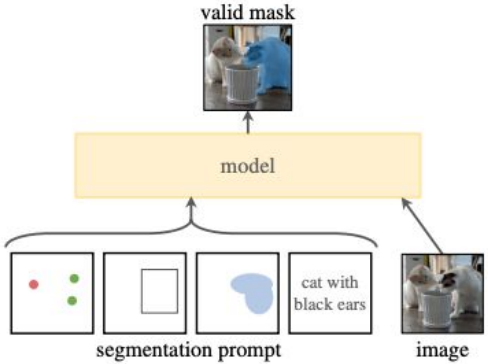


- DINOv2 made further implementation refinements (including adding a second masked patch prediction objective) and trained on 142 million images
- An example of a vision "**foundation model**": a model trained on a very large amount of data that can be subsequently used or tuned for diverse tasks. DINOv2 was demonstrated to be a state-of-the-art backbone for many tasks.

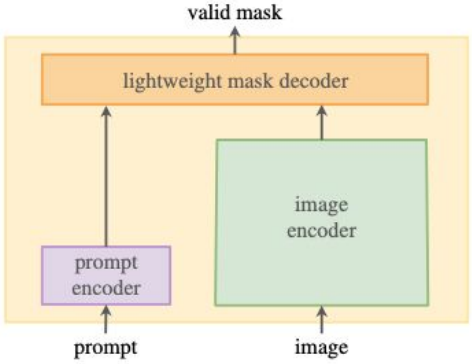
Caron et al. 2021, Oquab et al. 2023

Segment Anything Model (SAM): A foundation model targeted for segmentation

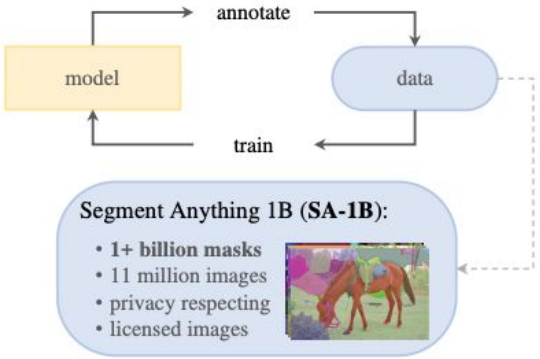
- Foundation model for promptable segmentation, based on a Transformer encoder-decoder architecture. Generalizes to many segmentation tasks.
- Not all representation learning needs to be through self-supervised learning. Here, a supervised paradigm that also achieves powerful representation learning.
- Trained on 1 billion masks from 11 million images, using the model in a data collection loop.



(a) **Task**: promptable segmentation



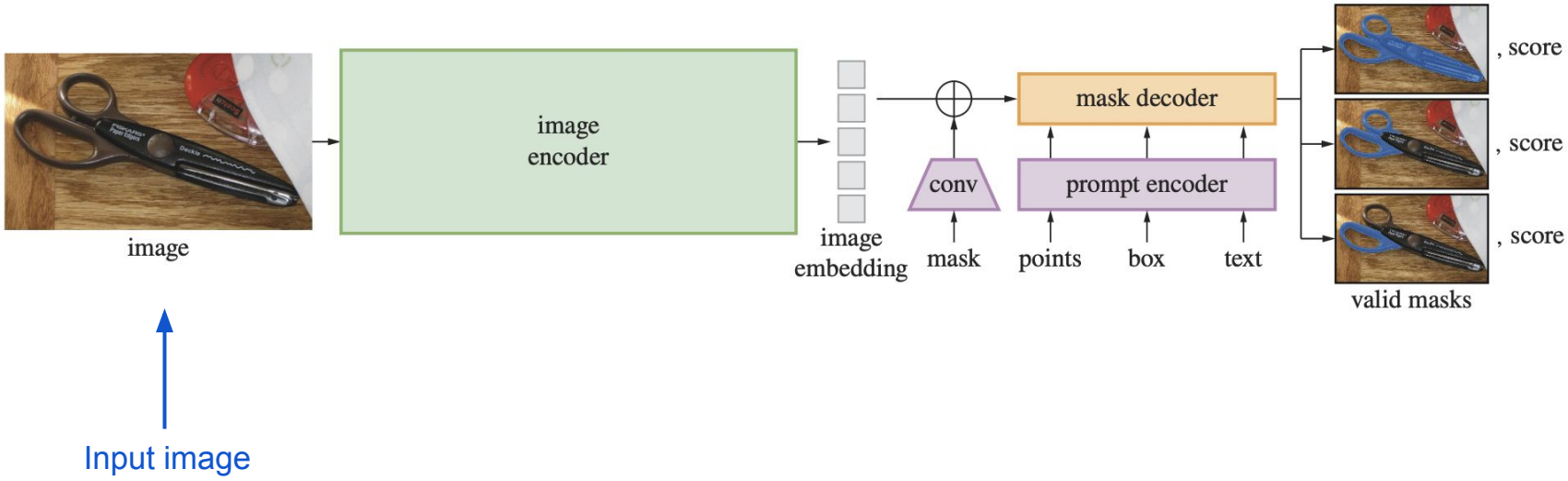
(b) **Model**: Segment Anything Model (**SAM**)



(c) **Data**: data engine (top) & dataset (bottom)

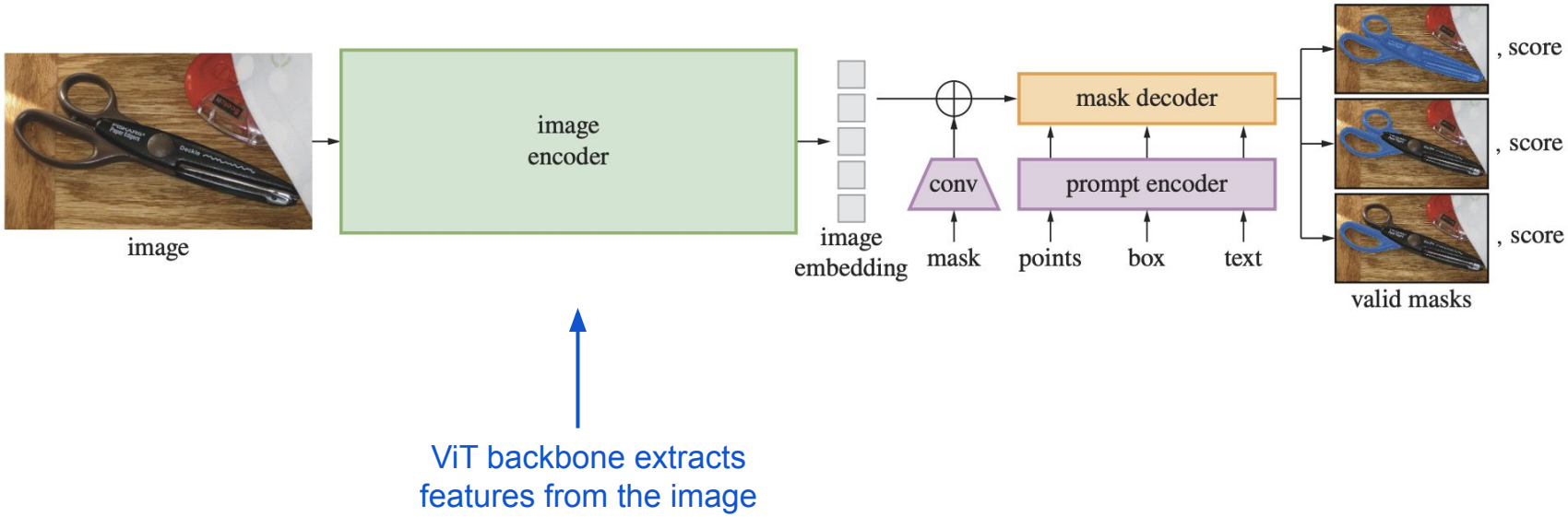
Kirrilov et al. 2023.

Segment Anything Model (SAM): A foundation model targeted for segmentation



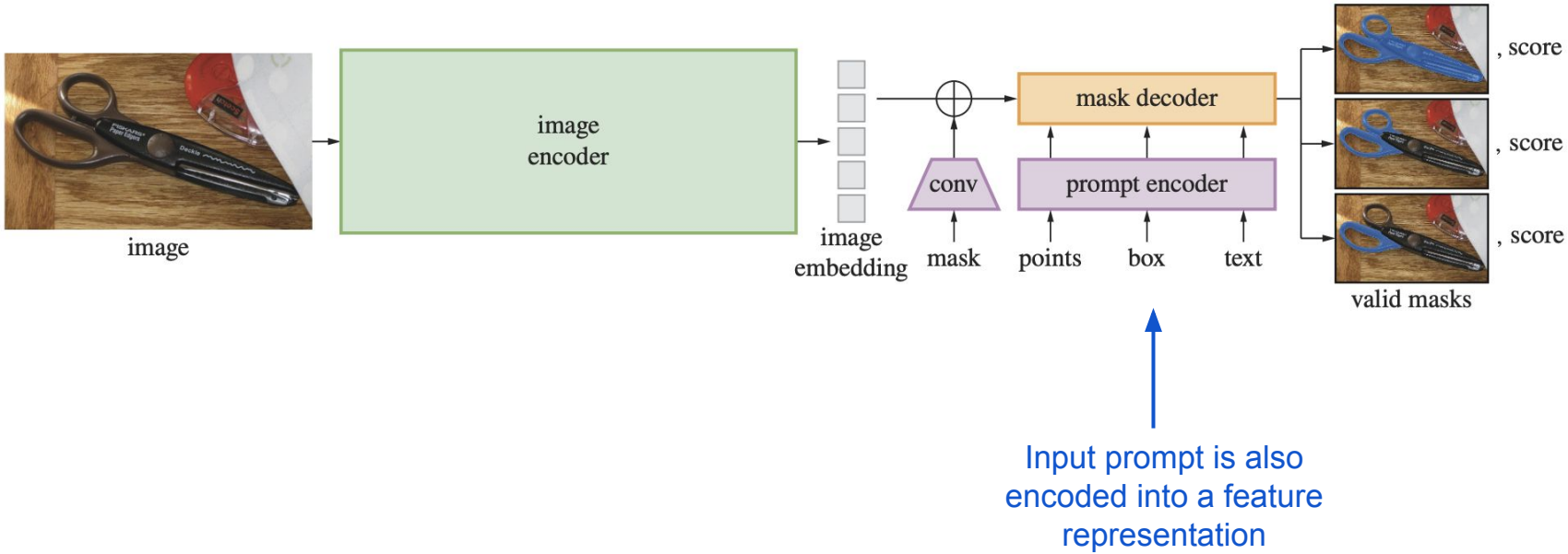
Kirrilov et al. 2023.

Segment Anything Model (SAM): A foundation model targeted for segmentation



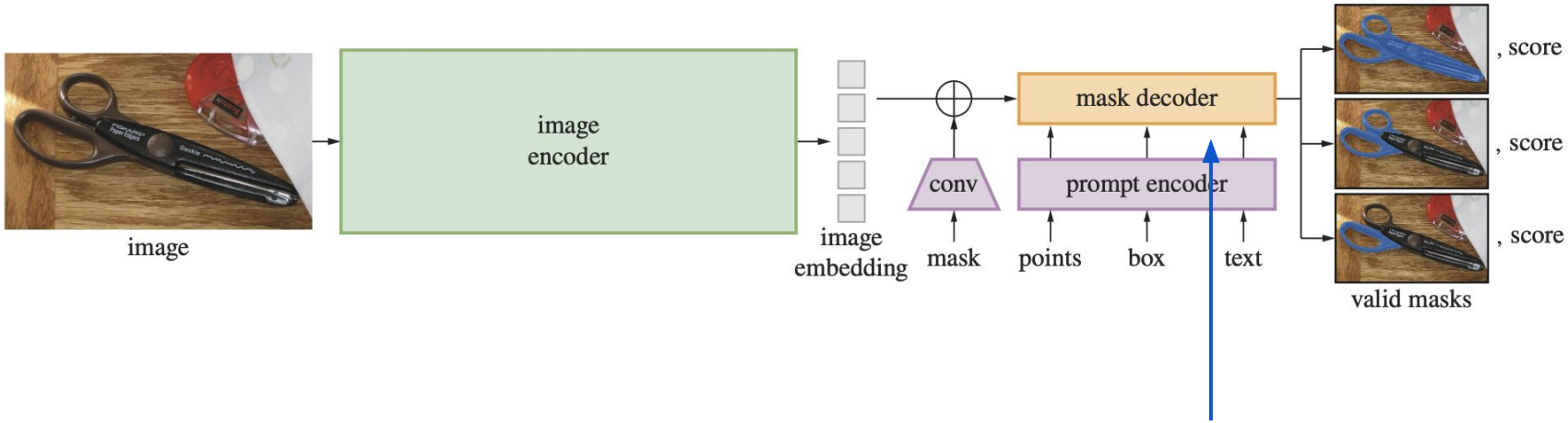
Kirrilov et al. 2023.

Segment Anything Model (SAM): A foundation model targeted for segmentation



Kirrilov et al. 2023.

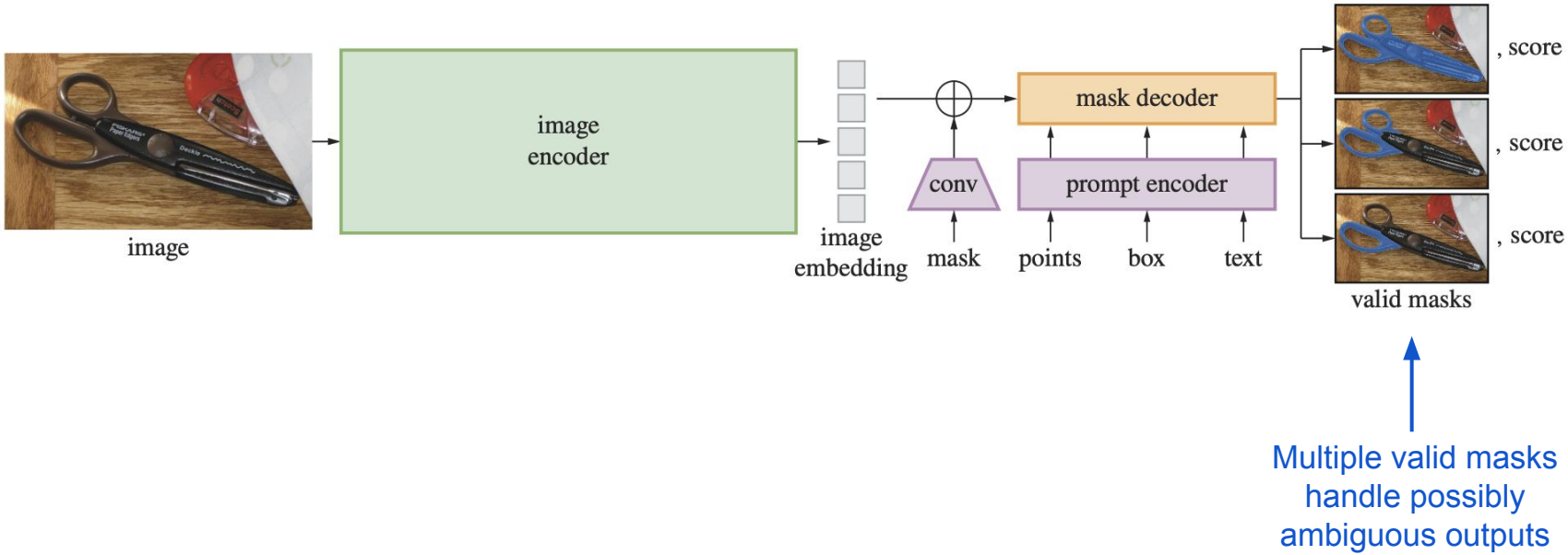
Segment Anything Model (SAM): A foundation model targeted for segmentation



Lightweight, two-layer mask decoder includes cross-attention with image embedding and transposed convolutions for upsampling

Kirrilov et al. 2023.

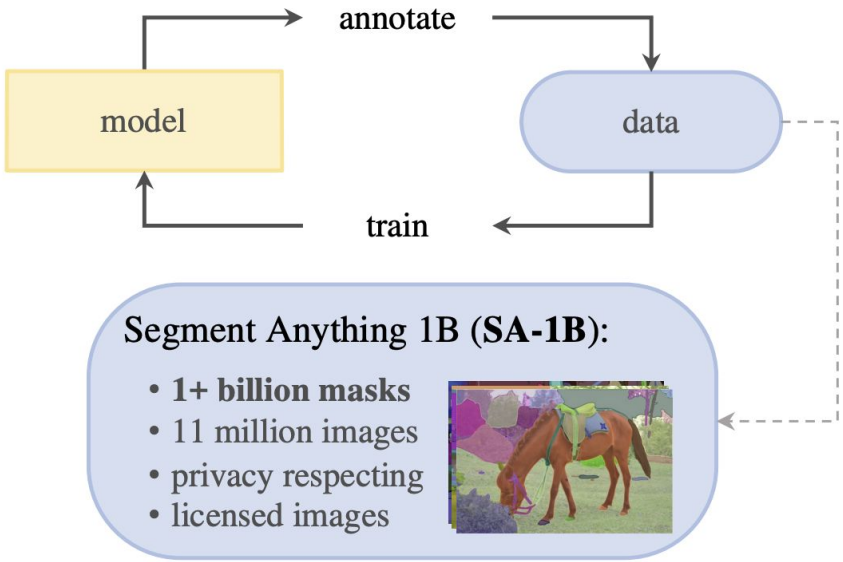
Segment Anything Model (SAM): A foundation model targeted for segmentation



Kirrilov et al. 2023.

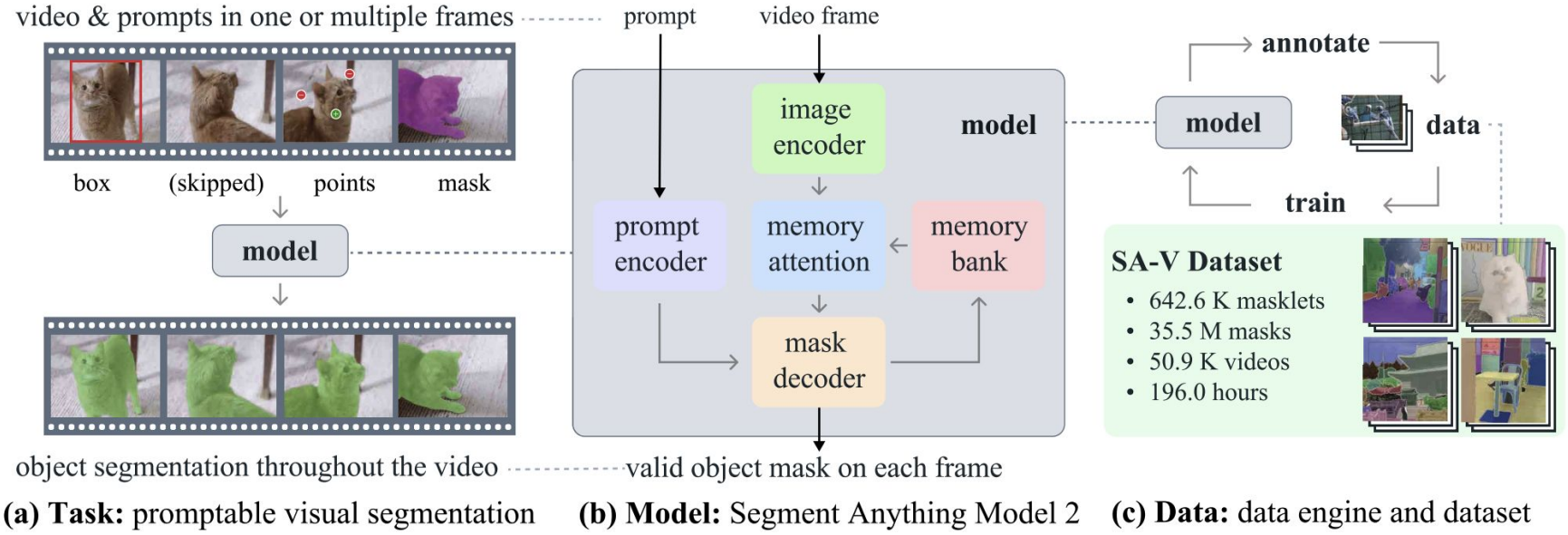
SAM data collection loop

- **Assisted-manual stage:** SAM is initially trained on existing annotated segmentation datasets. This is used as a starting point for interactive annotation by human annotators on new images. Produced 4.3M masks from 120K images.
- **Semi-automatic stage:** Images are prefilled with confident masks detected by SAM. Humans are asked to annotate additional unannotated objects. Produced 5.9M more masks from 180K images.
- **Fully automated stage:** Prompt SAM with dense grid of point prompts, and use post-processing and filtering to identify confident masks that can be used as training labels. Produced 1.1B masks from 11M images.



Kirrilov et al. 2023.

Segment Anything Model 2 (SAM 2): Extending to videos



(a) Task: promptable visual segmentation **(b) Model:** Segment Anything Model 2 **(c) Data:** data engine and dataset

Ravi et al. 2024

Next time

- Vision representation learners in biomedicine