Self-supervised learning

Christian Lessig
christian.lessig@ecmwf.int

Introduction

"Self-supervised learning: The dark matter of intelligence"

https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/

Introduction

Motivation--two sides of the same coin:

Introduction

Motivation--two sides of the same coin:

- Overcome limits imposed by requiring labelled data for training
  - Train on unlabelled data, i.e. data as it can be found ”in the wild”

Introduction

Motivation--two sides of the same coin:

- Overcome limits imposed by requiring labelled data for training
  - Train on unlabelled data, i.e. data as it can be found ”in the wild”
- Train a neural network that is useful for a wide range of tasks
  - Training strategy and problem formulation that goes beyond supervised, task specific learning

Why can this work at all?

- Small neural (e.g. a 10,000 parameter MLP) are interpolation "engines".

Why can this work at all?

- Small neural (e.g. a 10,000 parameter MLP) are interpolation “engines”.

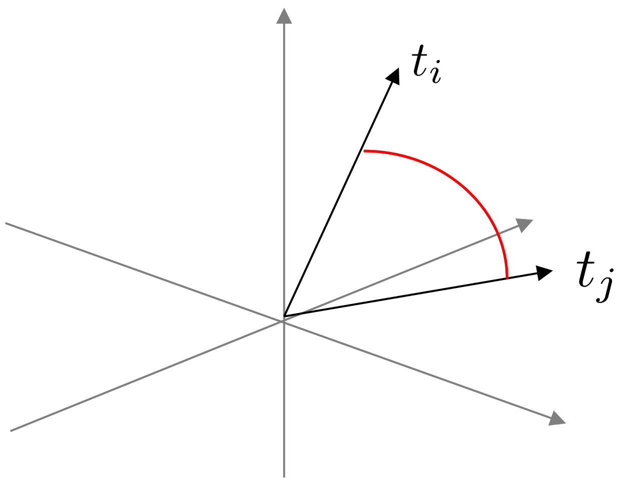
- Well-trained networks with 100s of millions or billions of parameters behave qualitatively differently
  - LLMs can answer a wide range of questions not seen during training
  - Pangu-Weather, GraphCast, AIFS provide skillful predictions multiple years past their training data set

Two different perspectives

1. Feature spaces:

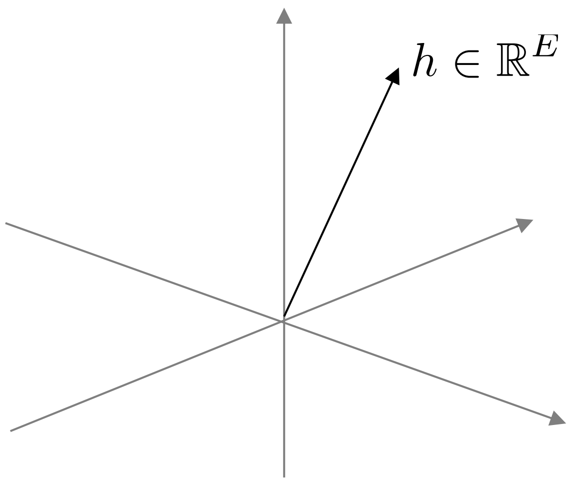
Two different perspectives

1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$



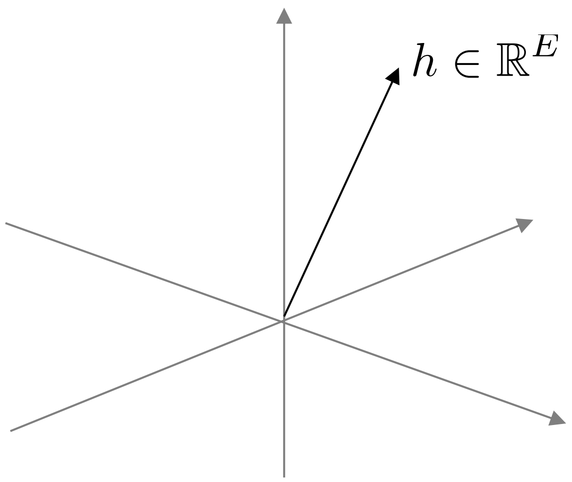
Two different perspectives

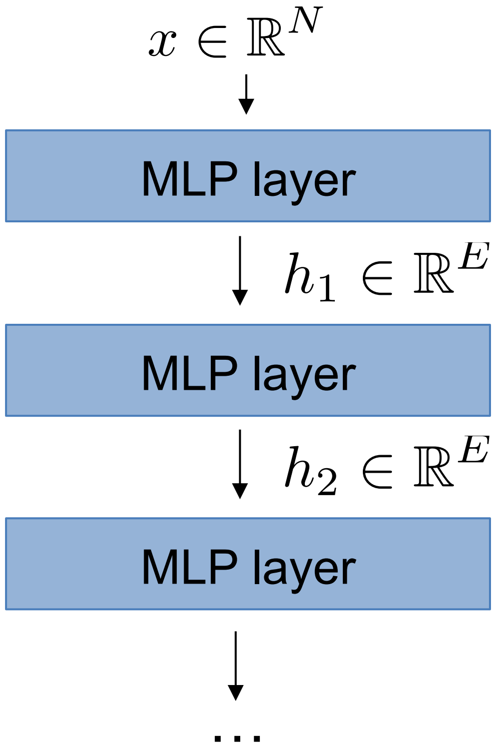
1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$



Two different perspectives

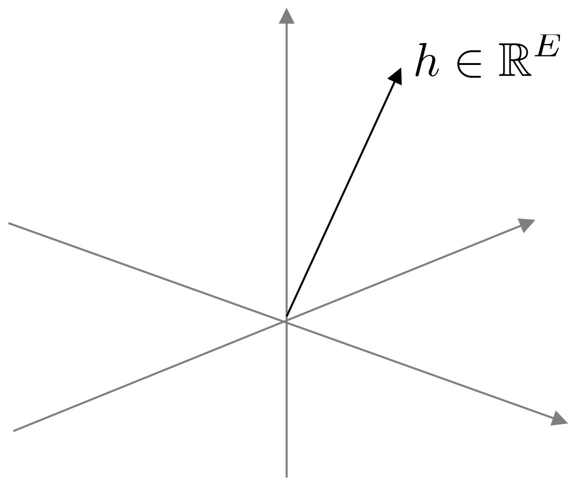
1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$

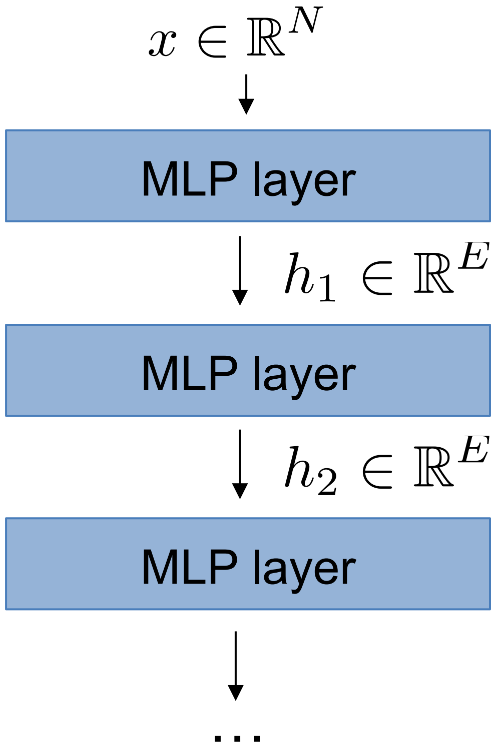




Two different perspectives

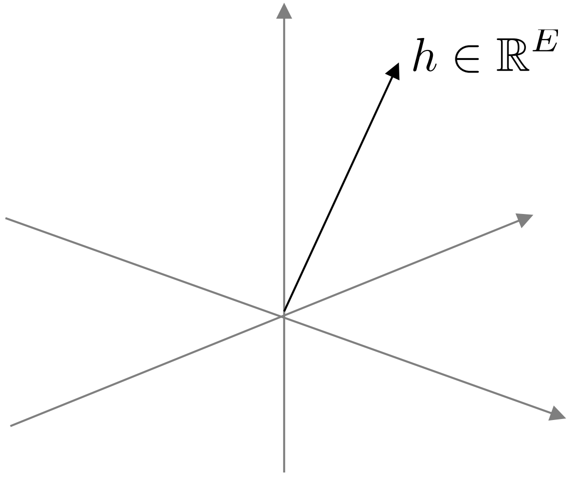
1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$
   - Feature spaces reveal important structures and remove noise

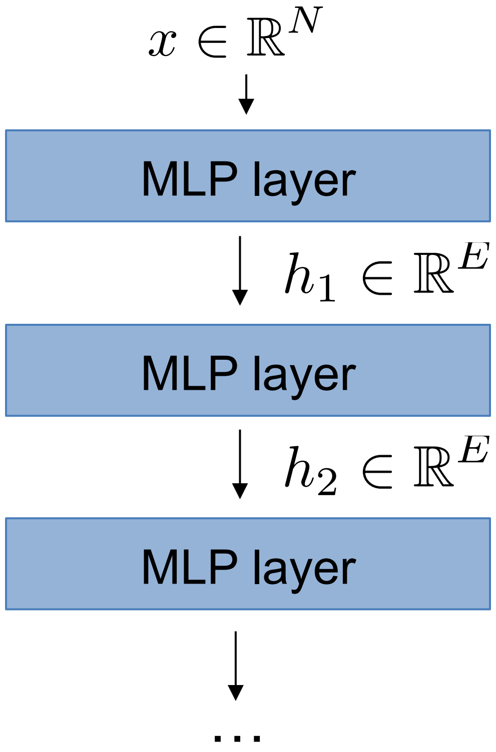




Two different perspectives

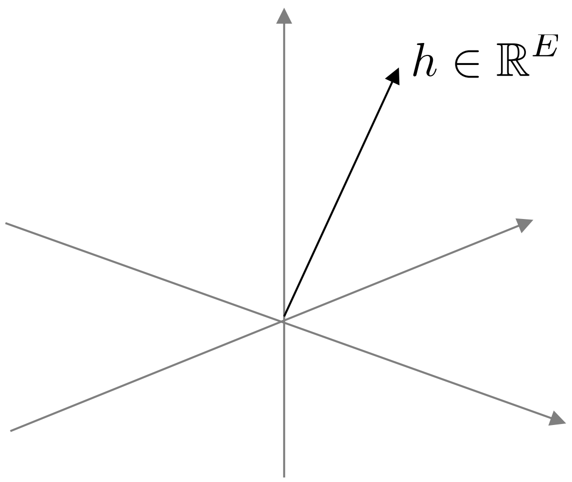
1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$
   - Feature spaces reveal important structures and remove noise
     - Analogous to Fourier domain, POD/PCA, …

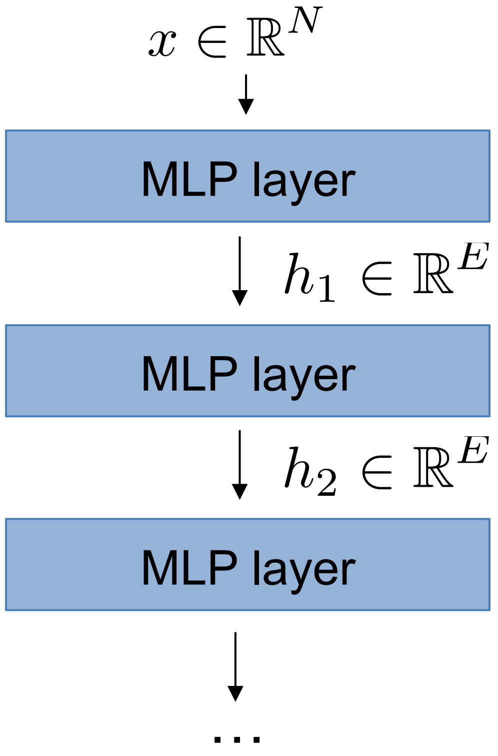




Two different perspectives

1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$
   - Feature spaces reveal important structures and remove noise
     - Analogous to Fourier domain, POD/PCA, …
     - BUT: learned and nonlinear





Two different perspectives

2. Learn general probabilistic model $p(y, x)$
   - When x, y are defined sufficiently general than this is task independent

Two different perspectives

2. Learn general probabilistic model $p(y, x)$
   - When x, y are defined sufficiently general than this is task independent
   - E.g. $p(y, x)$ the joint distribution over natural language

Two different perspectives

2. Learn general probabilistic model $p(y,x)$
   - When x, y are defined sufficiently general than this is task independent
   - E.g. $p(y,x)$ the joint distribution over natural language
     - Chat bot: x = question, y = answer
     - Translation: x = language A, y = language B
     - Spell/grammar correction: x = incorrect, y = corrected
     - Creative writing: x = content outline, y = long text form
     - …

Two different perspectives

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```
Translate English to French:
cheese =>
```

task description

prompt

Brown et al., Language Models are Few-Shot Learners, 2020, https://arxiv.org/pdf/2005.14165.pdf

Two different perspectives

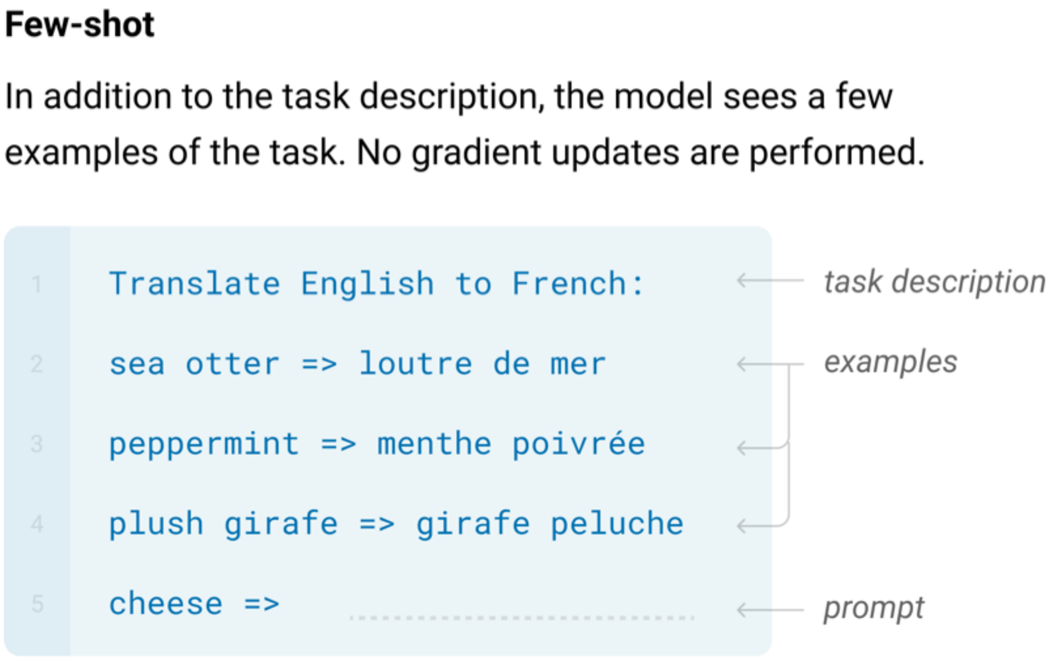
One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```
Translate English to French:      ← task description
sea otter => loutre de mer        ← example
cheese =>                         ← prompt
```

Brown et al., Language Models are Few-Shot Learners, 2020, https://arxiv.org/pdf/2005.14165.pdf

Two different perspectives



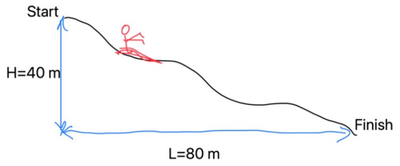
Brown et al., Language Models are Few-Shot Learners, 2020, https://arxiv.org/pdf/2005.14165.pdf

Two different perspectives

Prompt

Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?



Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

Start: $E = mgL$

End: $E = \frac{1}{2} mv^2$

$\Rightarrow mgL = \frac{1}{2} mv^2$

$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6 \, m.s^{-1}$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.

Response

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by mgh, where m is the mass of the skier, g is the acceleration due to gravity, and h is the height of the slope. The student used the length of the slope, L, instead of the height, H, in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by mgh.

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2} mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2} mv^2$$

Canceling out the mass m on both sides, we get:

$$gh = \frac{1}{2} v^2$$

Solving for v, we get:

$$v = \sqrt{2gh}$$

Substituting the given values, we get:

$$v = \sqrt{2 \times 9.81 \times 40} = 28.01 \text{ m/s}$$

Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.

Gemini technical report, https://arxiv.org/pdf/2312.11805.pdf

Two different perspectives

Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

Brown et al., Language Models are Few-Shot Learners, 2020, https://arxiv.org/pdf/2005.14165.pdf

Two different perspectives

State-of-the-art use 4 steps of fine-tuning for chat models

Brown et al., Language Models are Few-Shot Learners, 2020, https://arxiv.org/pdf/2005.14165.pdf

Two different perspectives

2. Learn general probabilistic model
   - When x, y are defined sufficiently general than this is task independent
   - E.g. $p(y, x)$ Is the joint distribution over atmospheric states

Two different perspectives

2. Learn general probabilistic model
   - When x, y are defined sufficiently general than this is task independent
   - E.g. $p(y, x)$ Is the joint distribution over atmospheric states
     - Forecasting: x = current state, y = future state
     - Downscaling: x = coare res. state, y = fine res. state
     - Spatial interpolation: x = incomplete state, y = completed state
     - Counterfactual/scenario: x = initial condition in scenario A, y = forecast in scenario B
     - …

Self-supervised learning tasks

Self-supervised learning: define a training task from a dataset without an explicit set of labels

Self-supervised learning tasks

Self-supervised learning: define a training task from a dataset without an explicit set of labels

=> “hide” some information from the network during input and network predicts this information

Self-supervised learning tasks

Self-supervised learning: define a training task from a dataset without an explicit set of labels

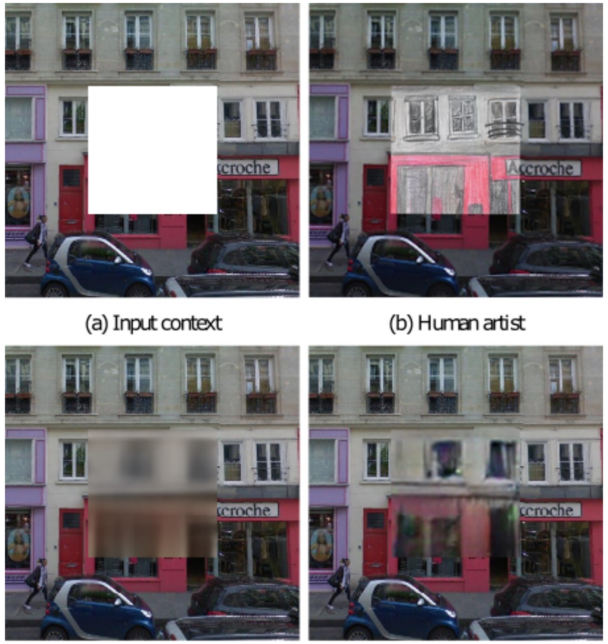
=> “hide” some information from the network during input and network predicts this information

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

Self-supervised learning tasks

Self-supervised learning: define a training task from a dataset without an explicit set of labels

=> “hide” some information from the network during input and network predicts this information



(a) Input context (b) Human artist

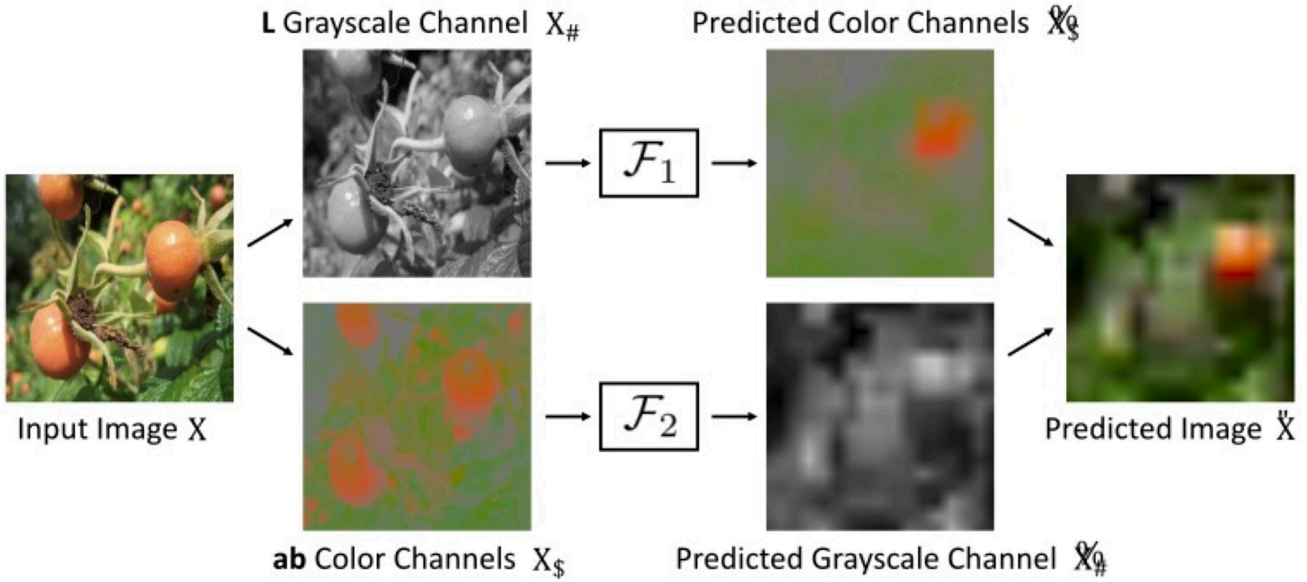
(c) Context Encoder (L 2 loss) (d) Context Encoder (L 2 +Adversarial loss)

D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

Self-supervised learning tasks

Self-supervised learning: define a training task from a dataset without an explicit set of labels

=> “hide” some information from the network during input and network predicts this information



R. Zhang, P. Isola, and A. A. Efros. Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), July 2017.

Self-supervised learning tasks

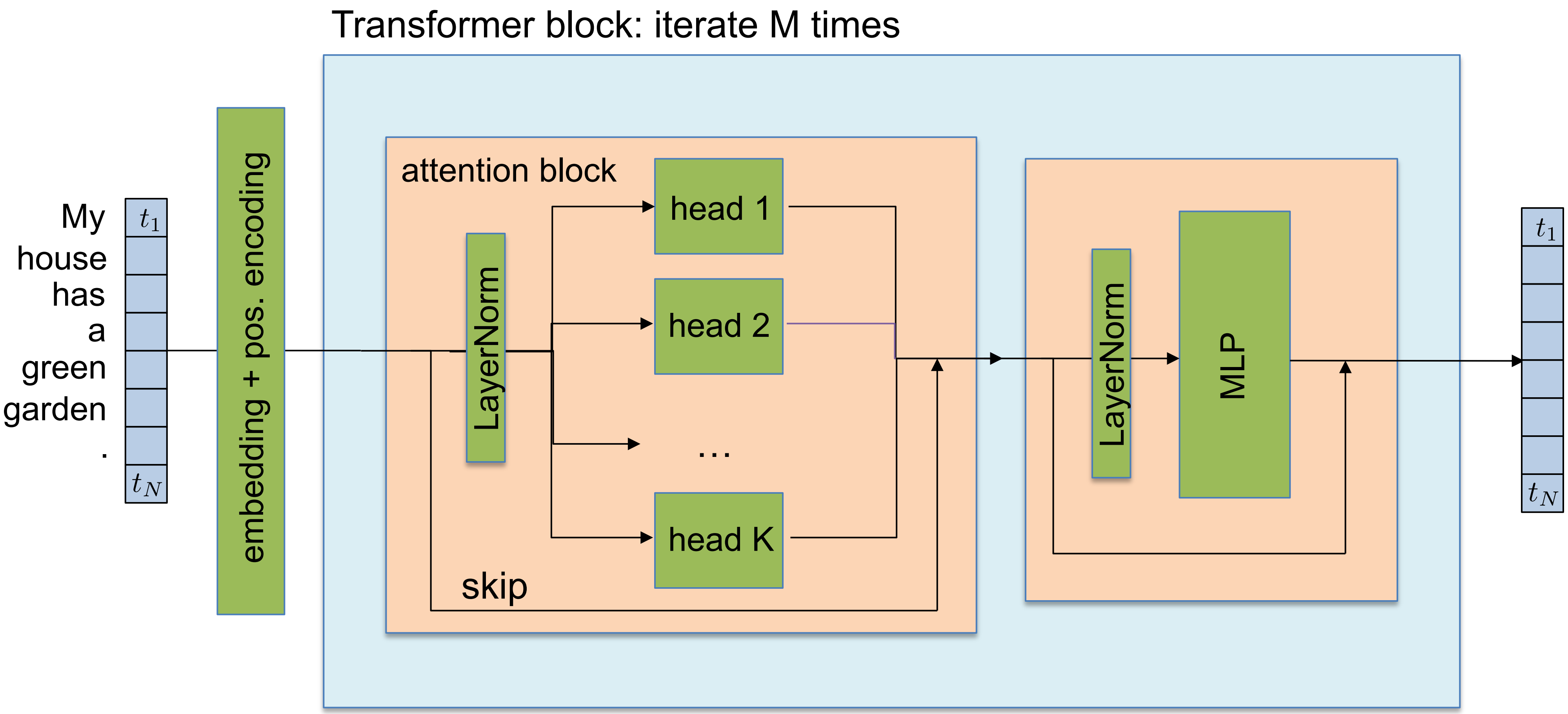
Self-supervised learning: define a training task from a dataset without an explicit set of labels

=> "hide" some information from the network during input and network predicts this information

Transformer takes sequence of words as input ((sub-)words, image patches, local atmospheric states, …)

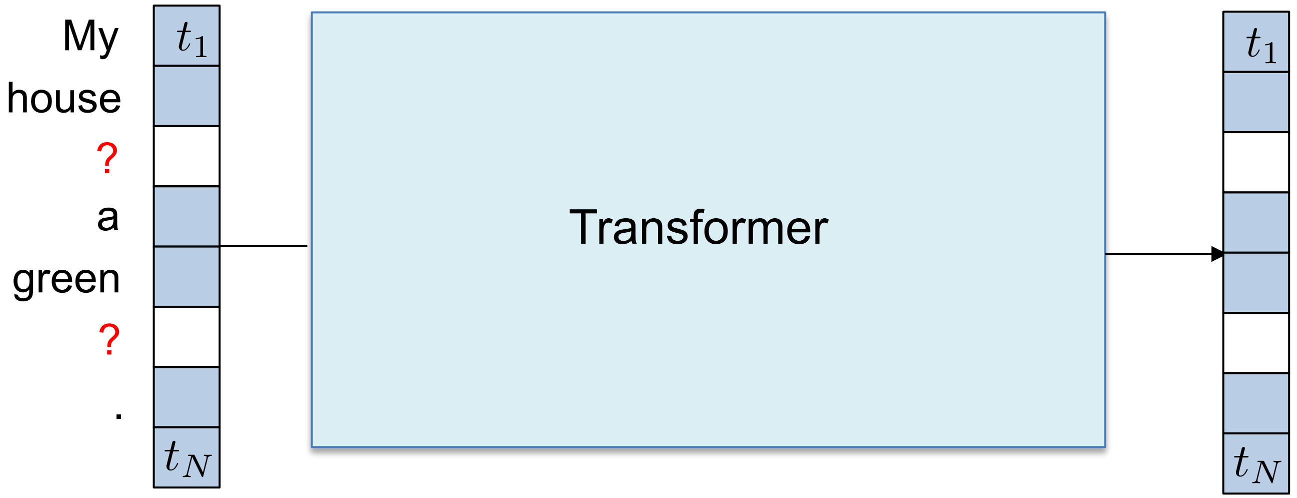
=> mask some of the patches from the network during input (or remove them entirely) and network predicts these

Self-supervised learning tasks

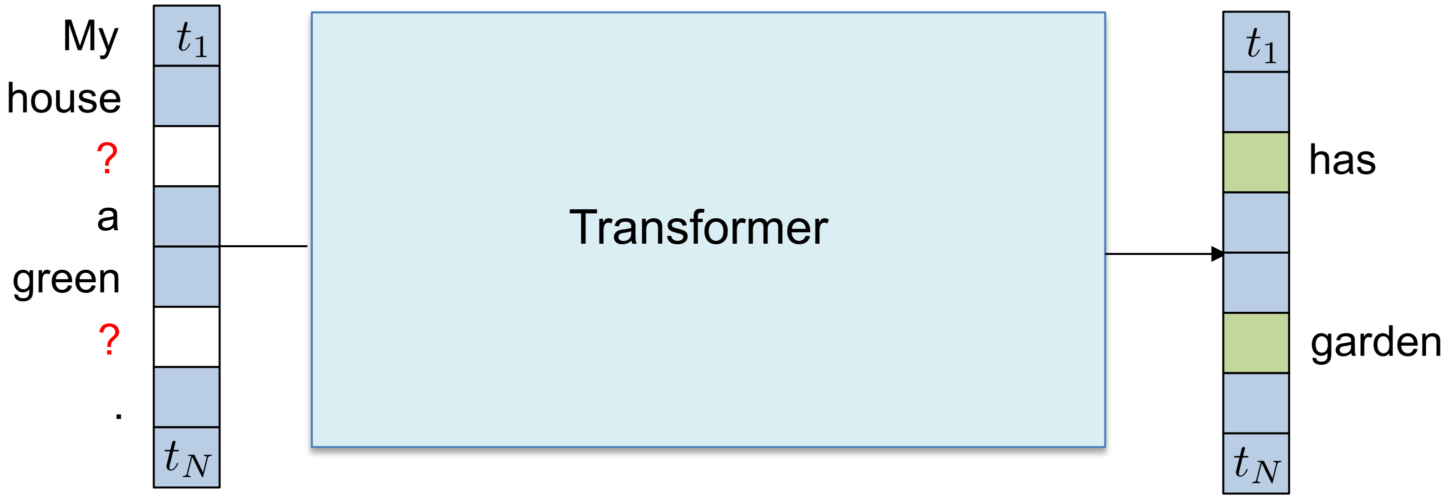


Self-supervised learning tasks

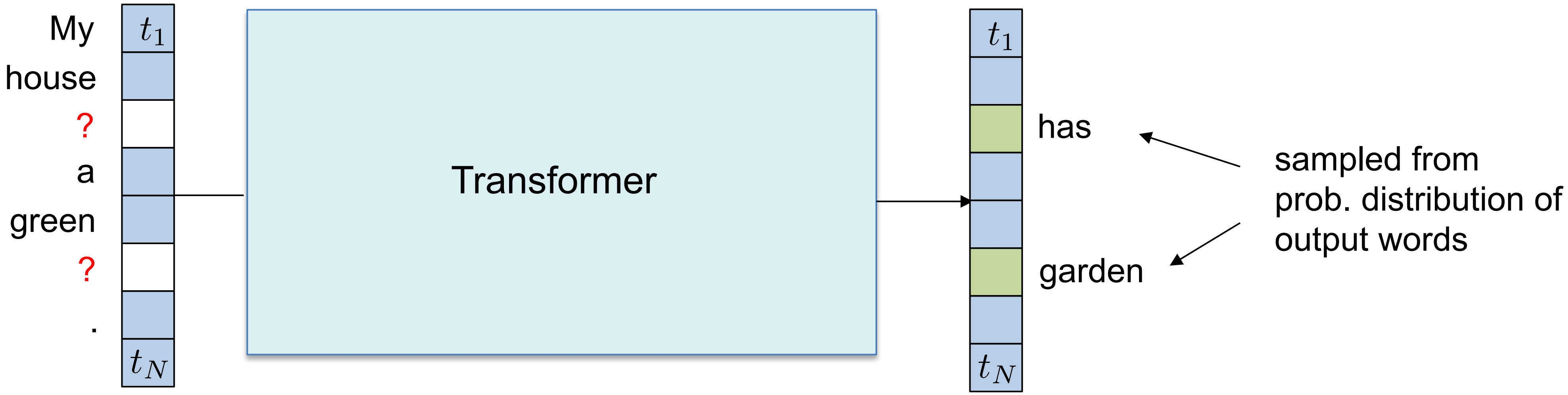
Self-supervised learning tasks



Self-supervised learning tasks

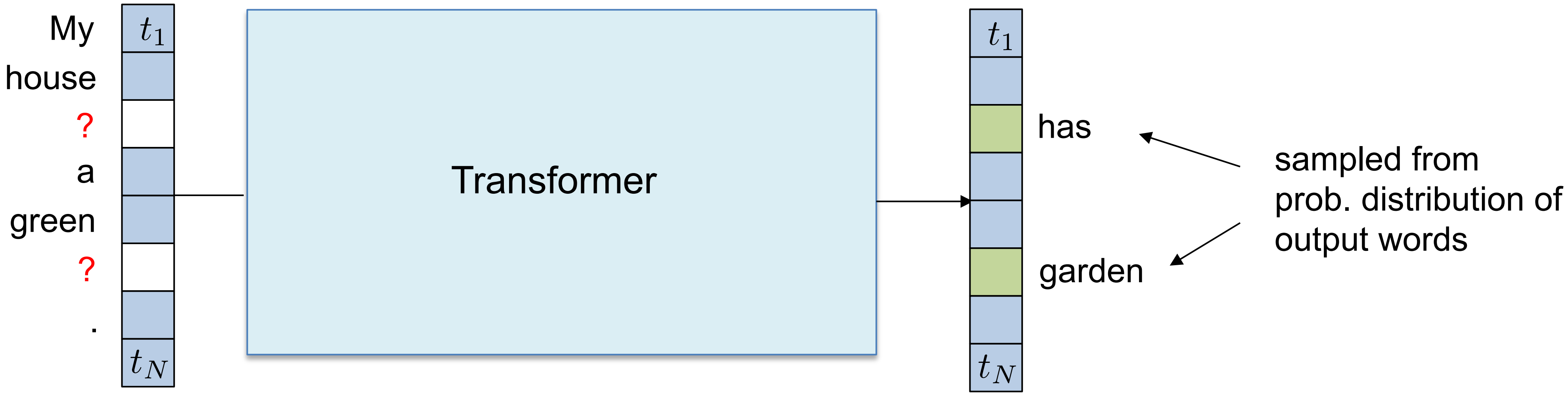


Self-supervised learning tasks



Self-supervised learning tasks

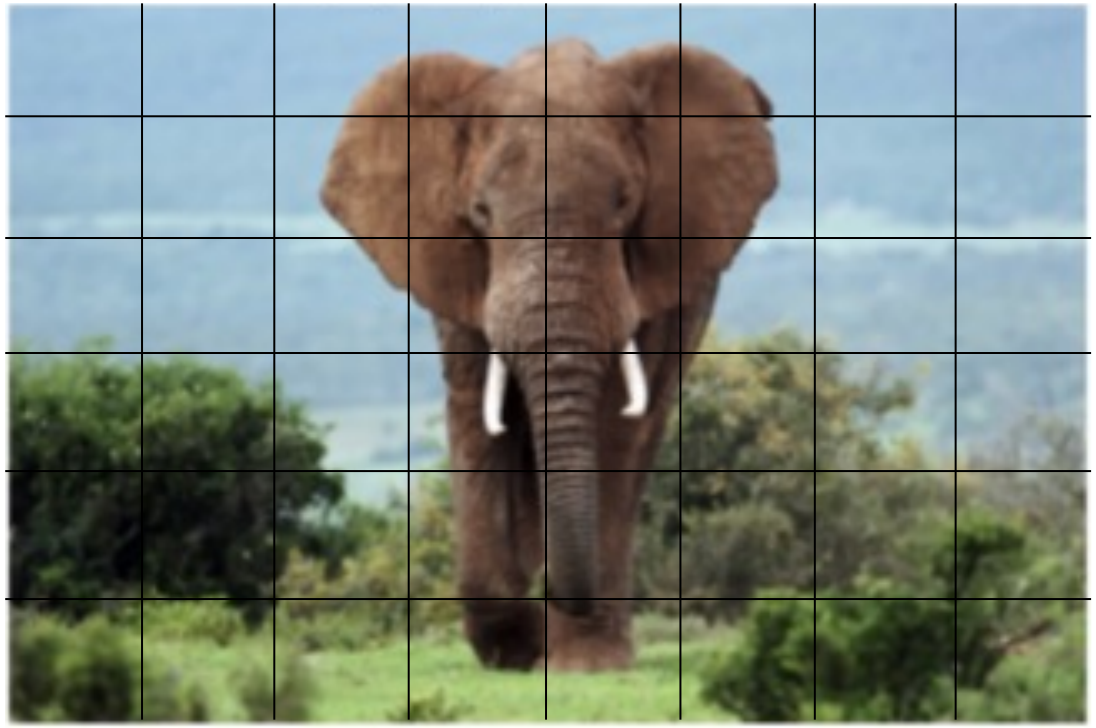
- BERT (Google):[1] randomly mask words from a sequence (and add some random distortions)
- Predictive masking (OpenAI):[2] always mask subsequent words



[1] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, https://arxiv.org/abs/1810.04805
[2] Radford et al. Improving Language Understanding by Generative Pre-Training, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

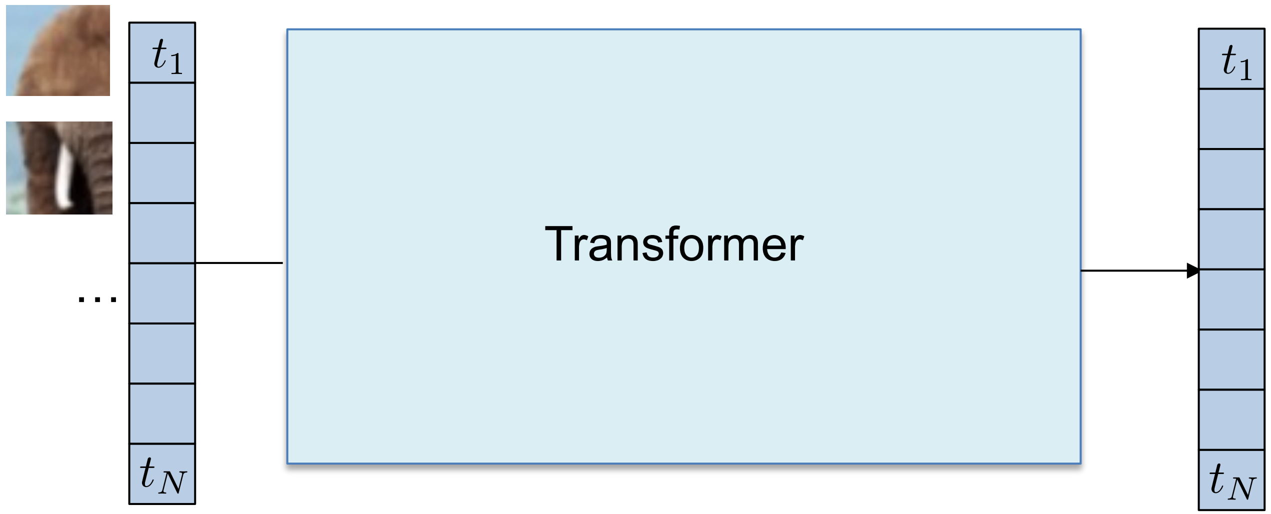
Self-supervised learning tasks

Vision transformer: image is a small patch



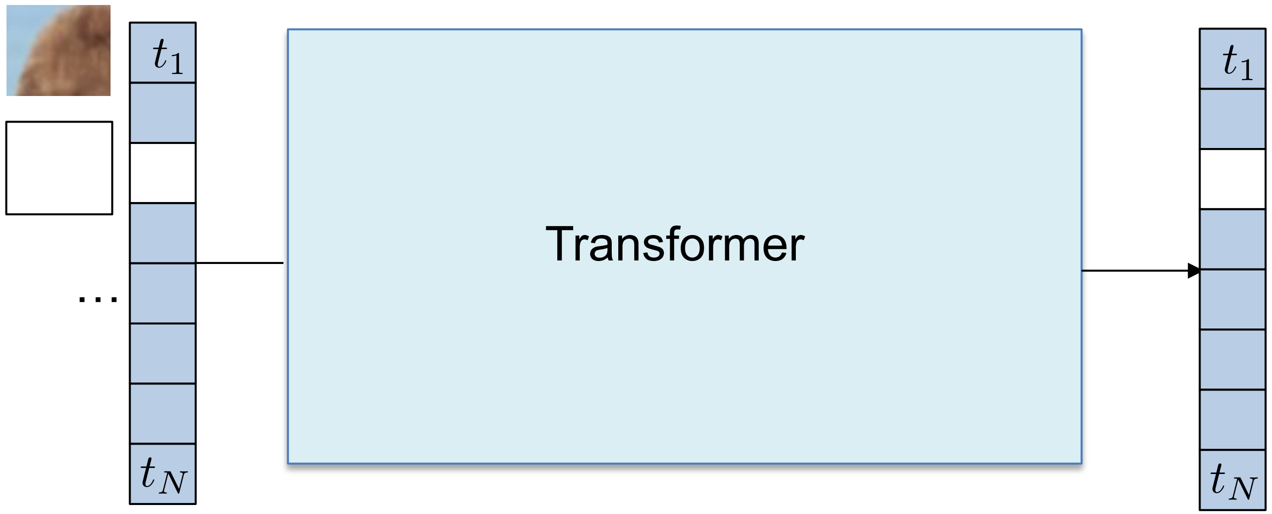
Dosovitskiy et al., An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, 2021, https://arxiv.org/abs/2010.11929

Self-supervised learning tasks



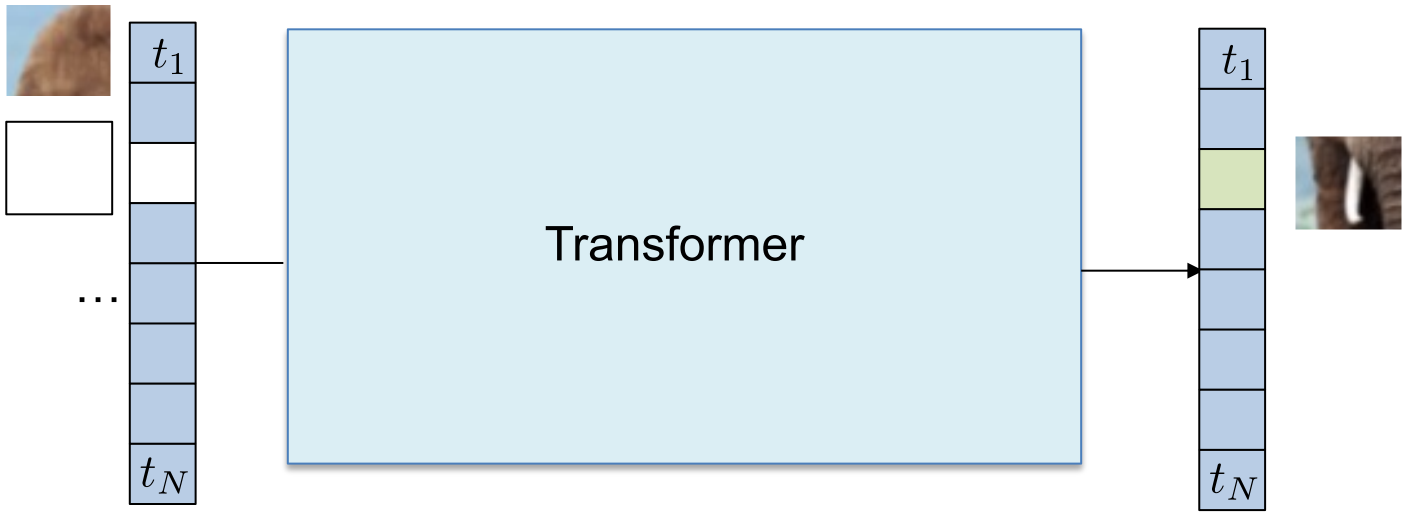
He et al., Masked Autoencoders Are Scalable Vision Learners, 2021, https://arxiv.org/abs/2111.06377

Self-supervised learning tasks



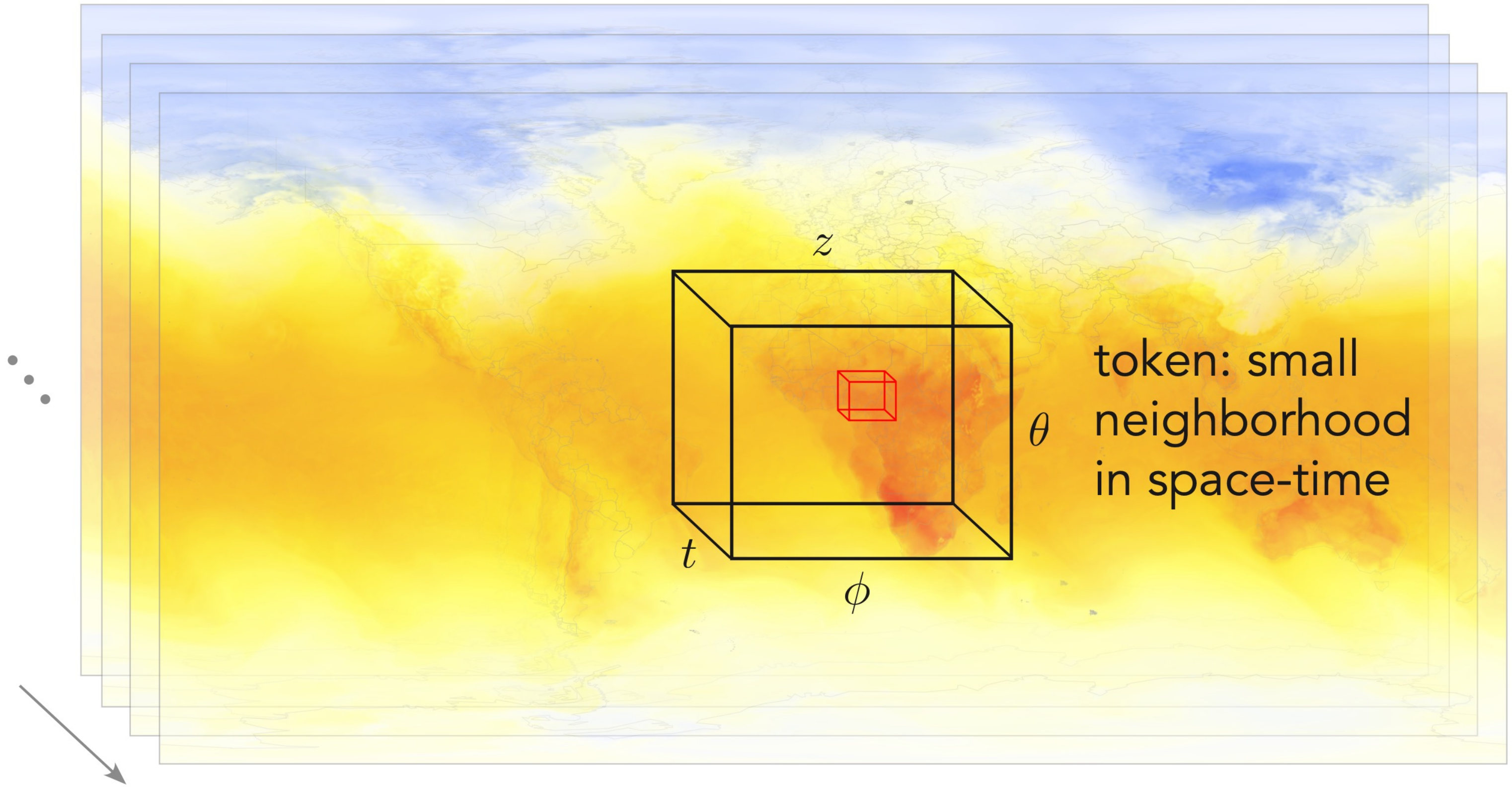
He et al., Masked Autoencoders Are Scalable Vision Learners, 2021, https://arxiv.org/abs/2111.06377

Self-supervised learning tasks



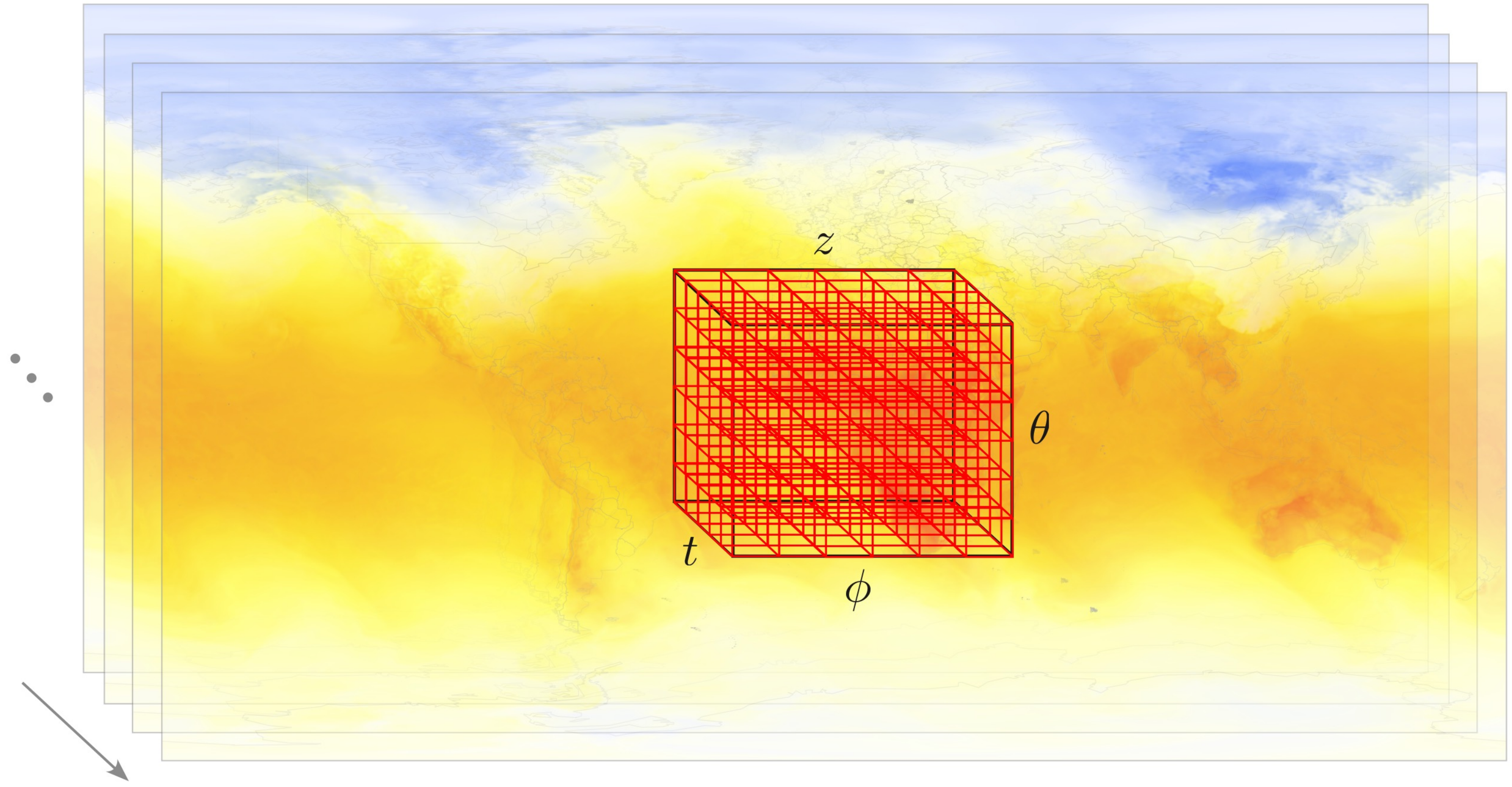
He et al., Masked Autoencoders Are Scalable Vision Learners, 2021, https://arxiv.org/abs/2111.06377

Self-supervised learning tasks



Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

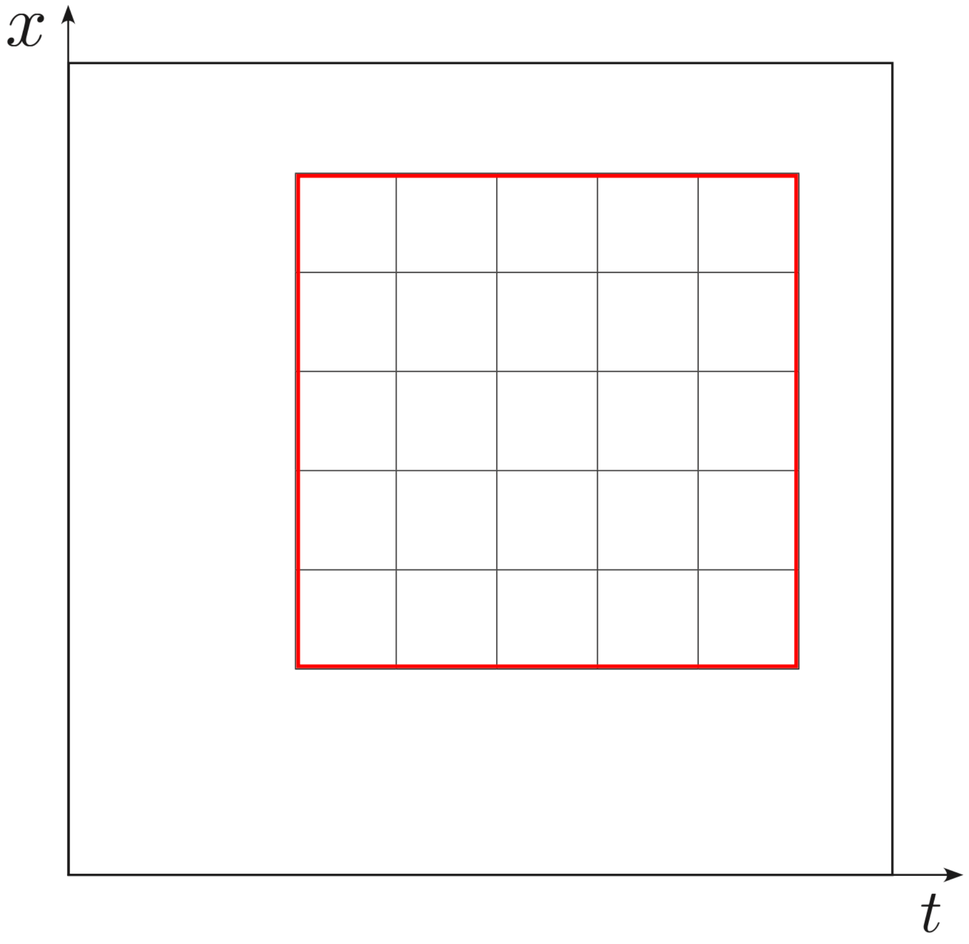
Self-supervised learning tasks



Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks

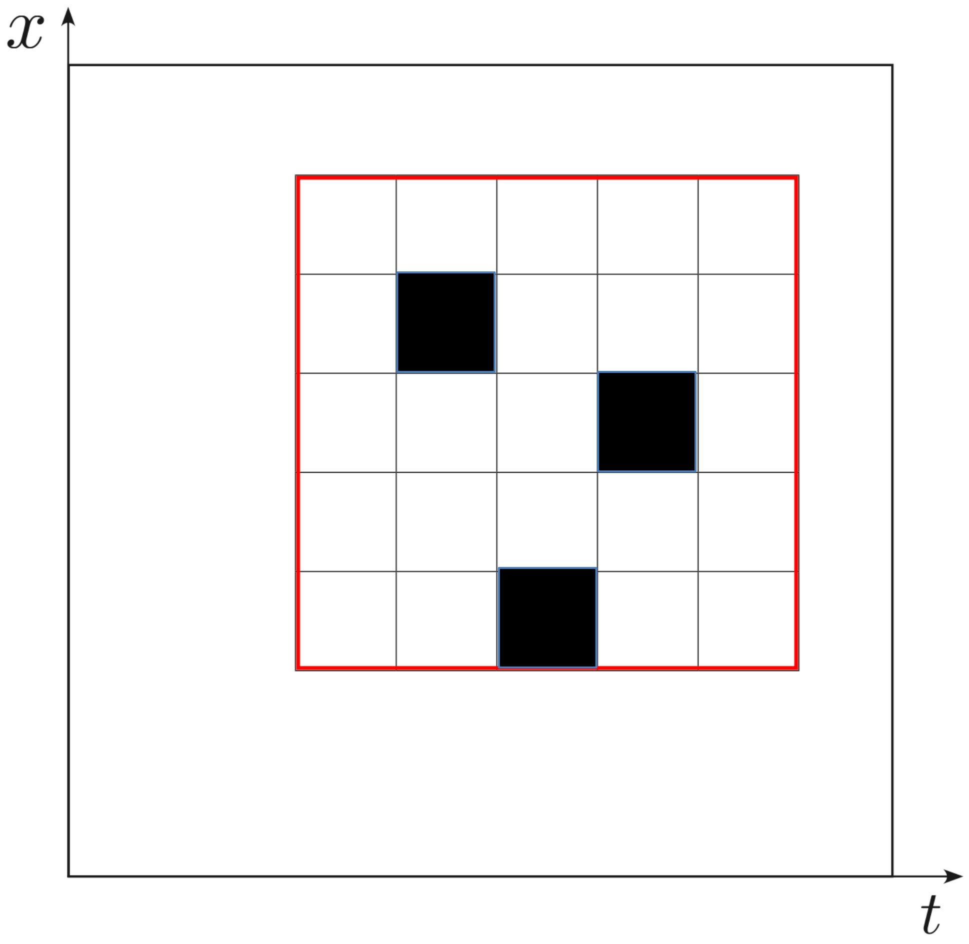
Flatland view



Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

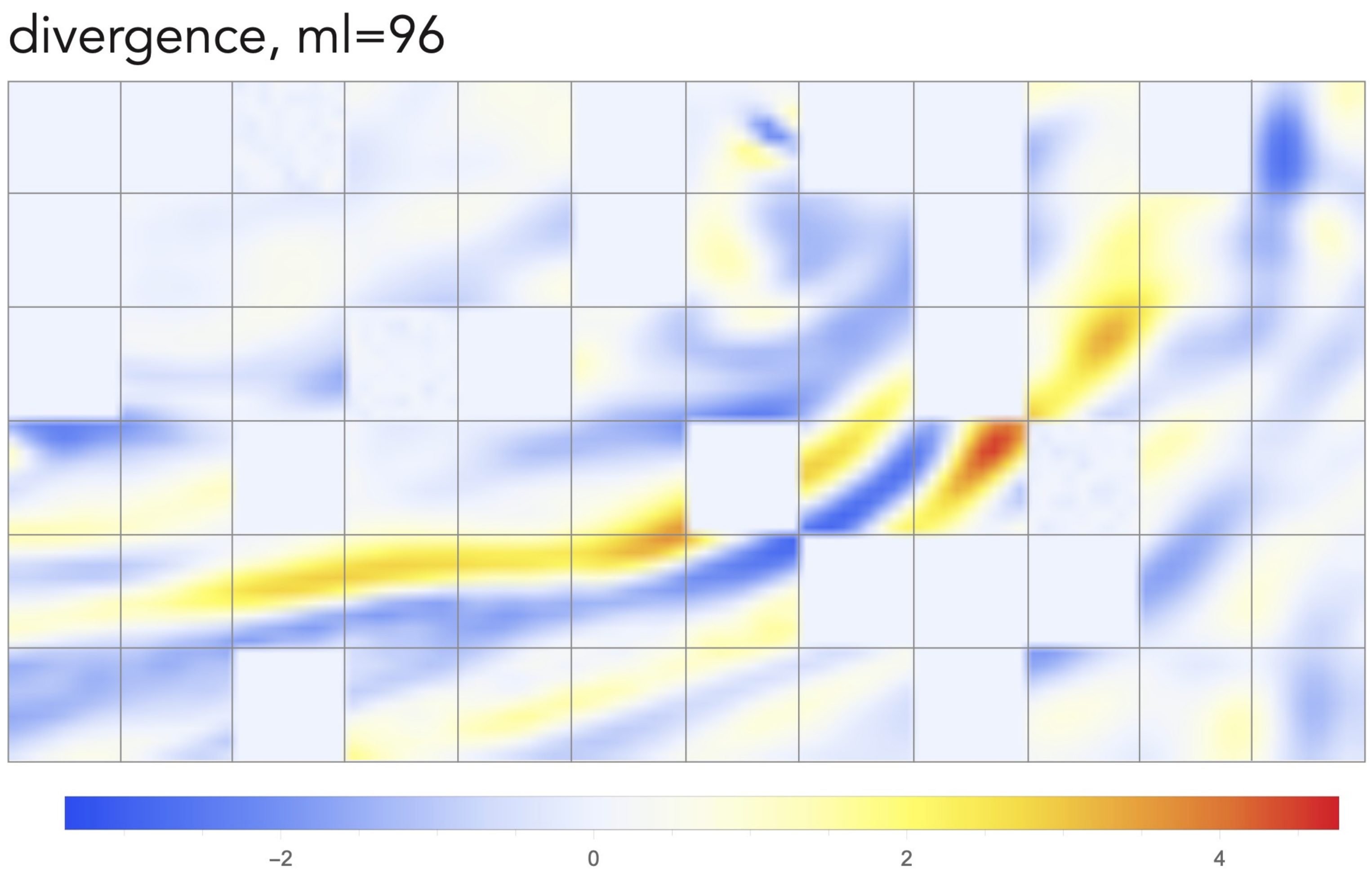
Self-supervised learning tasks

Flatland view



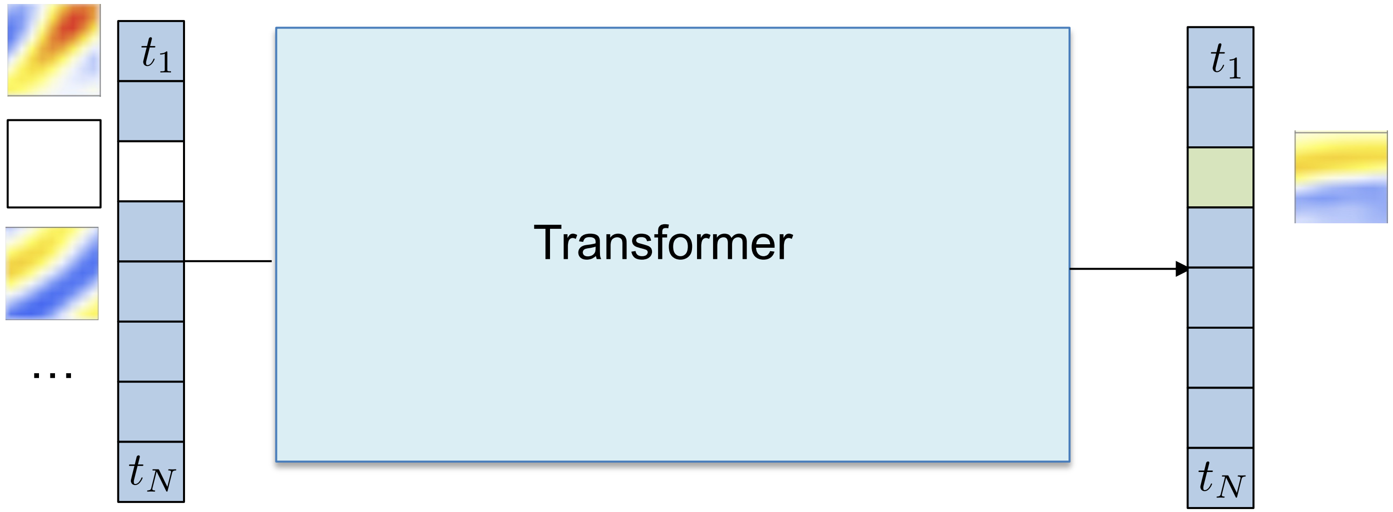
Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks



Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

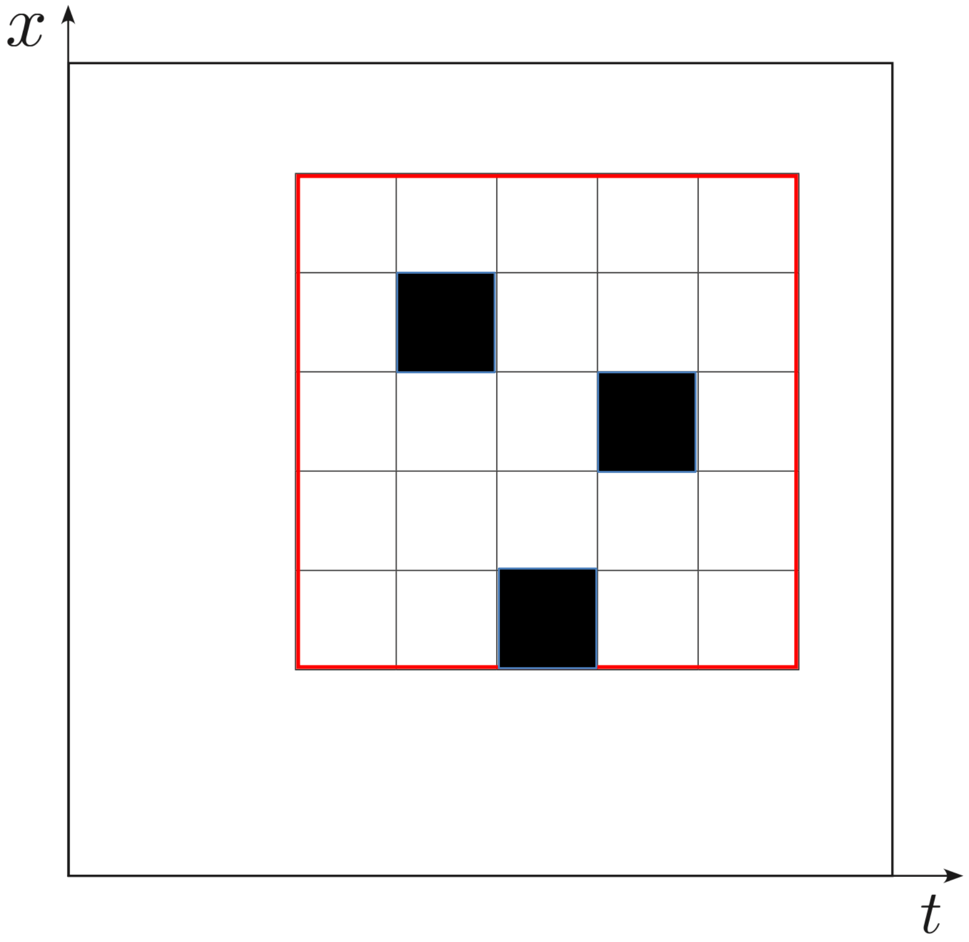
Self-supervised learning tasks



Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks

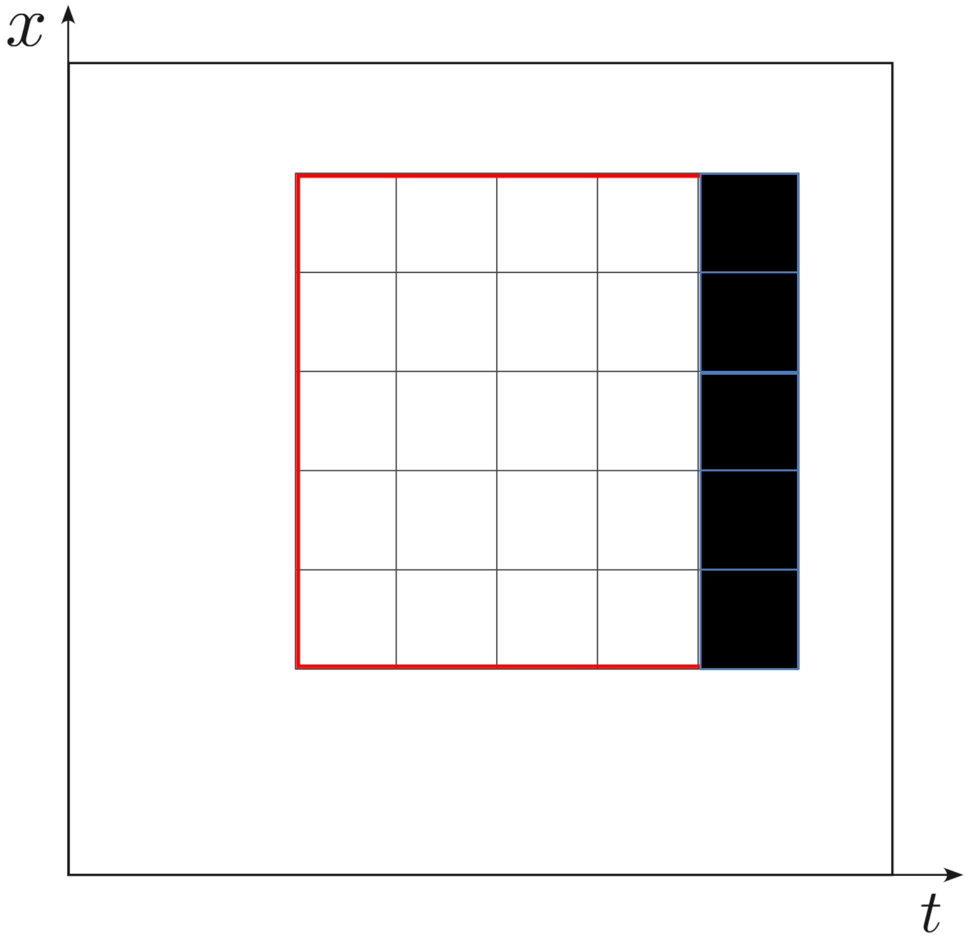
Zero-shot capabilities



Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks

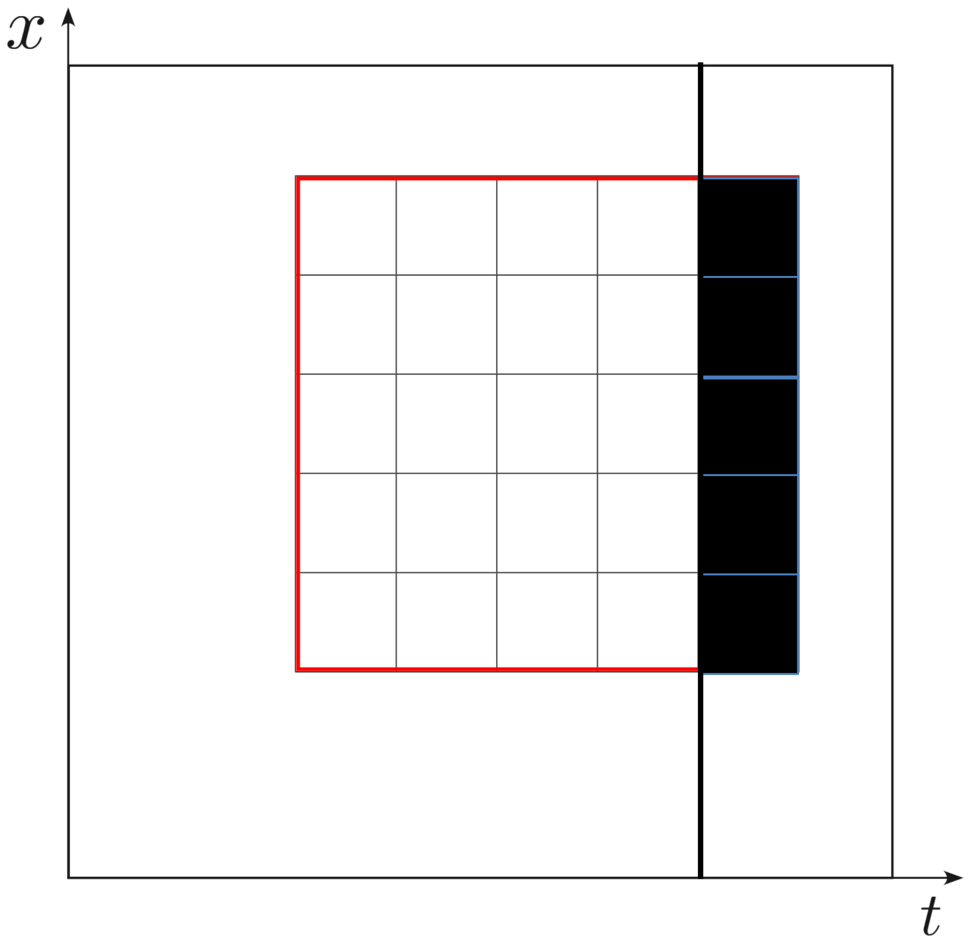
Zero-shot capabilities



Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks

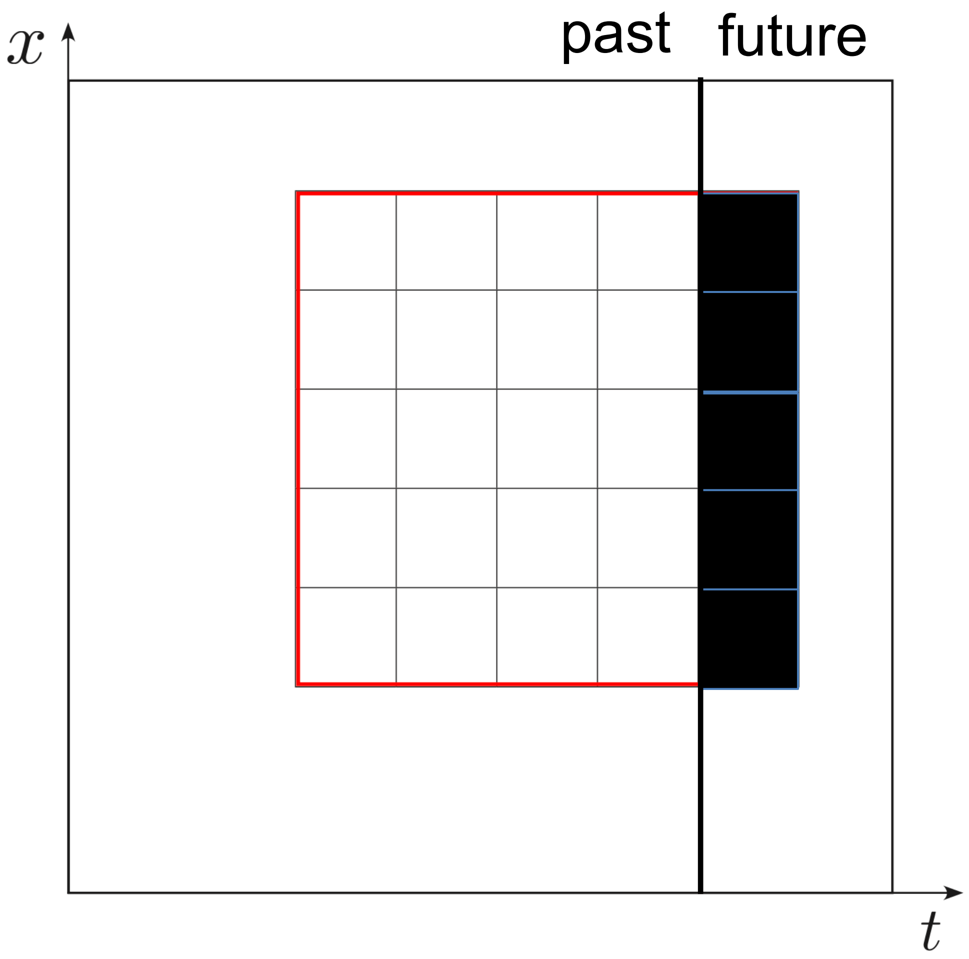
Zero-shot capabilities



Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks

Zero-shot capabilities

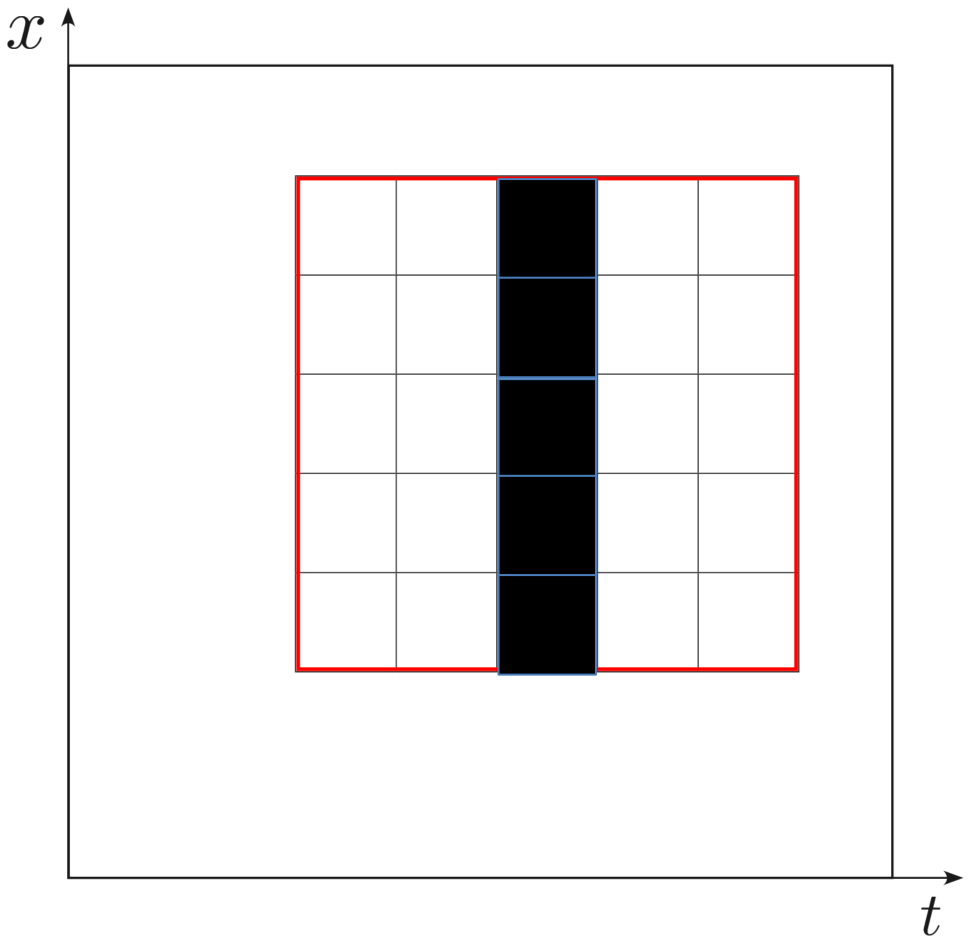


forecasting

Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks

Zero-shot capabilities

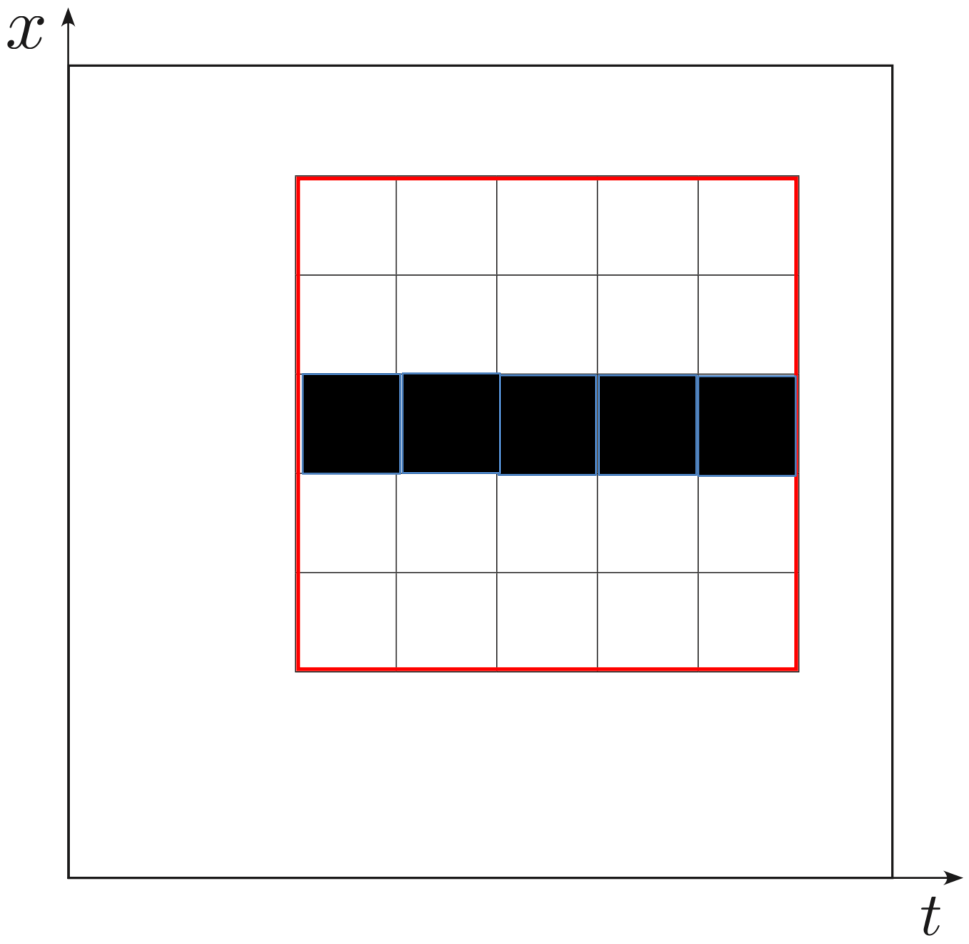


temporal interpolation

Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks

Zero-shot capabilities



temporal interpolation

Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

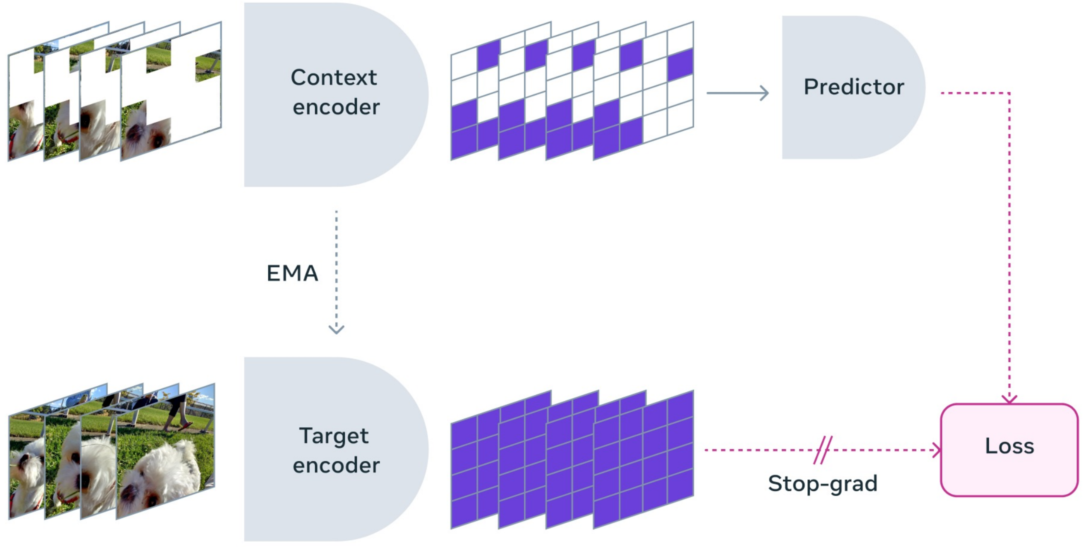
Self-supervised learning tasks

- Joined Embedding Predictive Architecture (JEPA)
  - Mask but compute loss in hidden/latent space instead of by reconstructing
  - Learn more abstract and robust representations

Bardes et al., Revisiting Feature Prediction for Learning Visual Representations from Video, 2024, https://scontent-cdg4-2.xx.fbcdn.net/v/t39.2365-6/427986745_768441298640104_1604906292521363076_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=3c67a6&_nc_ohc=Lpq5IeF5ftUAX9EN6b7&_nc_ht=scontent-cdg4-2.xx&oh=00_AfCFIyd8GMJnqQsG90WY-ccXwWEooa0XgiWXZm06nd1-pw&oe=65D69EB1

Self-supervised learning tasks

- Joined Embedding Predictive Architecture (JEPA)
  - Mask but compute loss in hidden/latent space instead of by reconstructing
  - Learn more abstract and robust representations

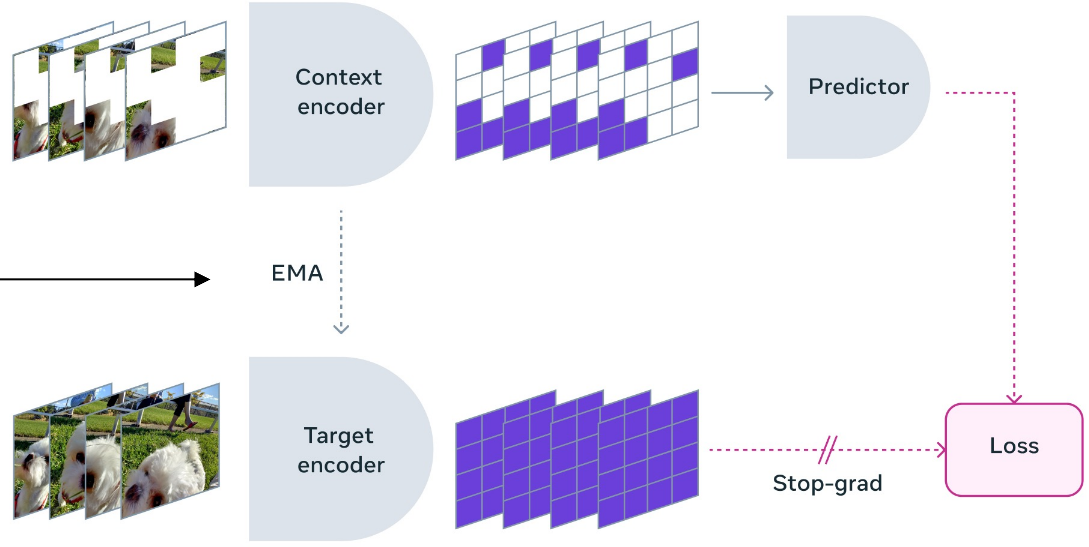


Bardes et al., Revisiting Feature Prediction for Learning Visual Representations from Video, 2024, https://scontent-cdg4-2.xx.fbcdn.net/v/t39.2365-6/427986745_768441298640104_1604906292521363076_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=3c67a6&_nc_ohc=Lpq5IeF5ftUAX9EN6b7&_nc_ht=scontent-cdg4-2.xx&oh=00_AfCFIyd8GMJnqQsG90WY-ccXwWEooa0XgiWXZm06nd1-pw&oe=65D69EB1

Self-supervised learning tasks

- Joined Embedding Predictive Architecture (JEPA)
  - Mask but compute loss in hidden/latent space instead of by reconstructing
  - Learn more abstract and robust representations

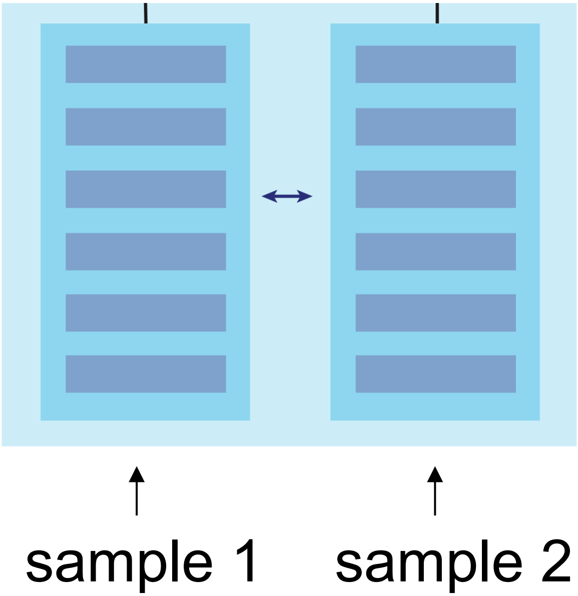
Exponential moving average



Bardes et al., Revisiting Feature Prediction for Learning Visual Representations from Video, 2024, https://scontent-cdg4-2.xx.fbcdn.net/v/t39.2365-6/427986745_768441298640104_1604906292521363076_n.pdf?_nc_cat=103&ccb=1-7&_nc_sid=3c67a6&_nc_ohc=Lpq5IeF5ftUAX9EN6b7&_nc_ht=scontent-cdg4-2.xx&oh=00_AfCFIyd8GMJnqQsG90WY-ccXwWEooa0XgiWXZm06nd1-pw&oe=65D69EB1

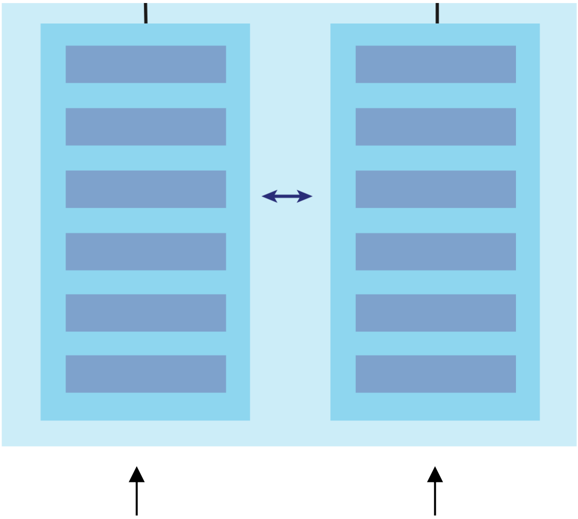
Self-supervised learning tasks

- Siamese networks



Self-supervised learning tasks

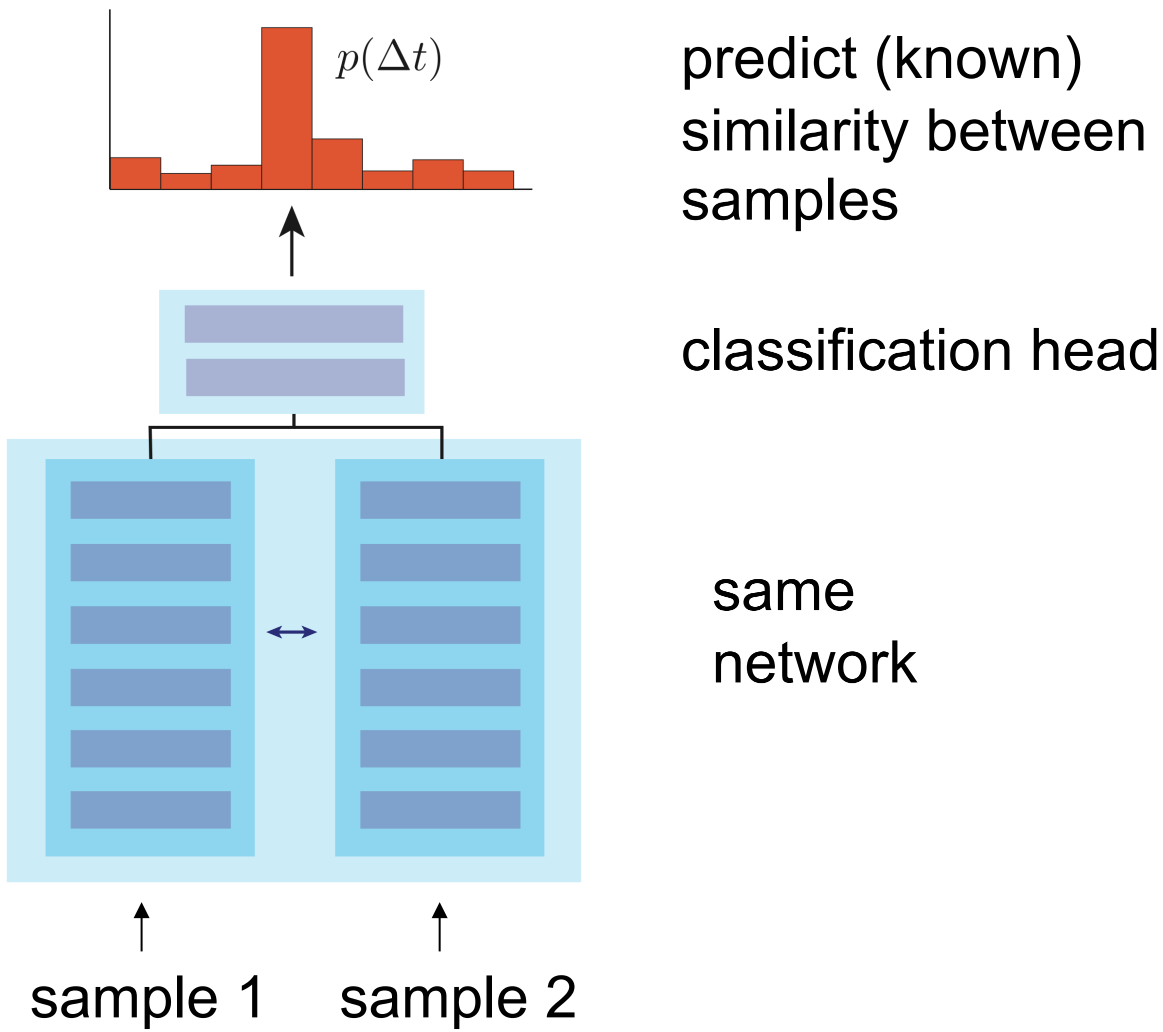
- Siamese networks



same network

Self-supervised learning tasks

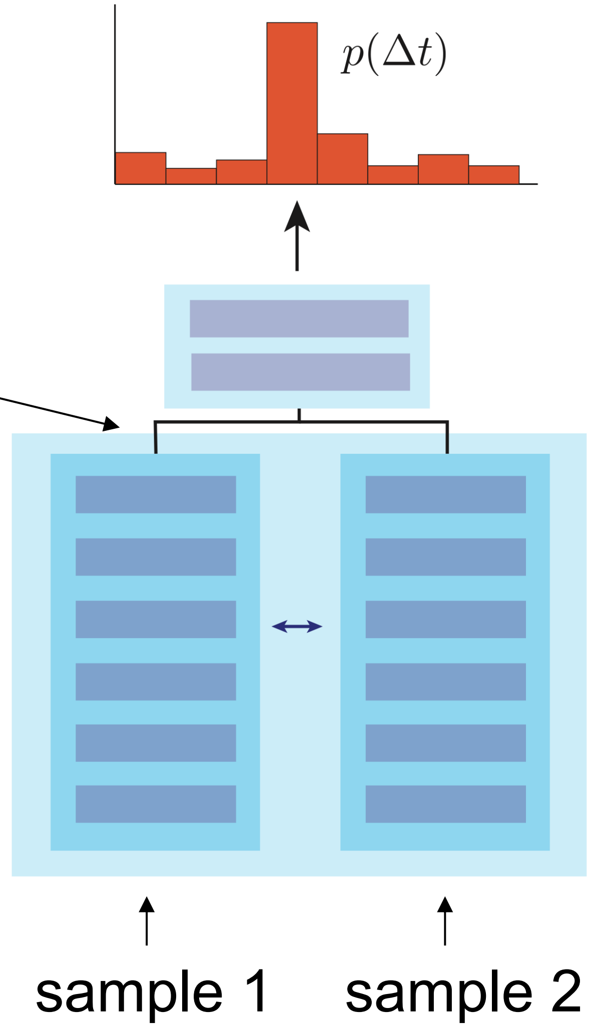
- Siamese networks



Self-supervised learning tasks

- Siamese networks

Latent representation is used for applications

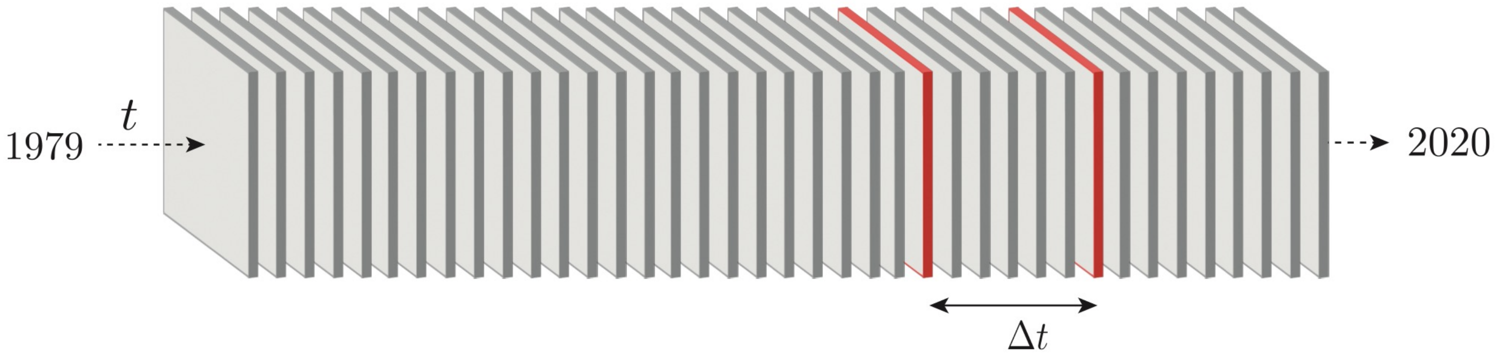


predict (known) similarity between samples

classification head

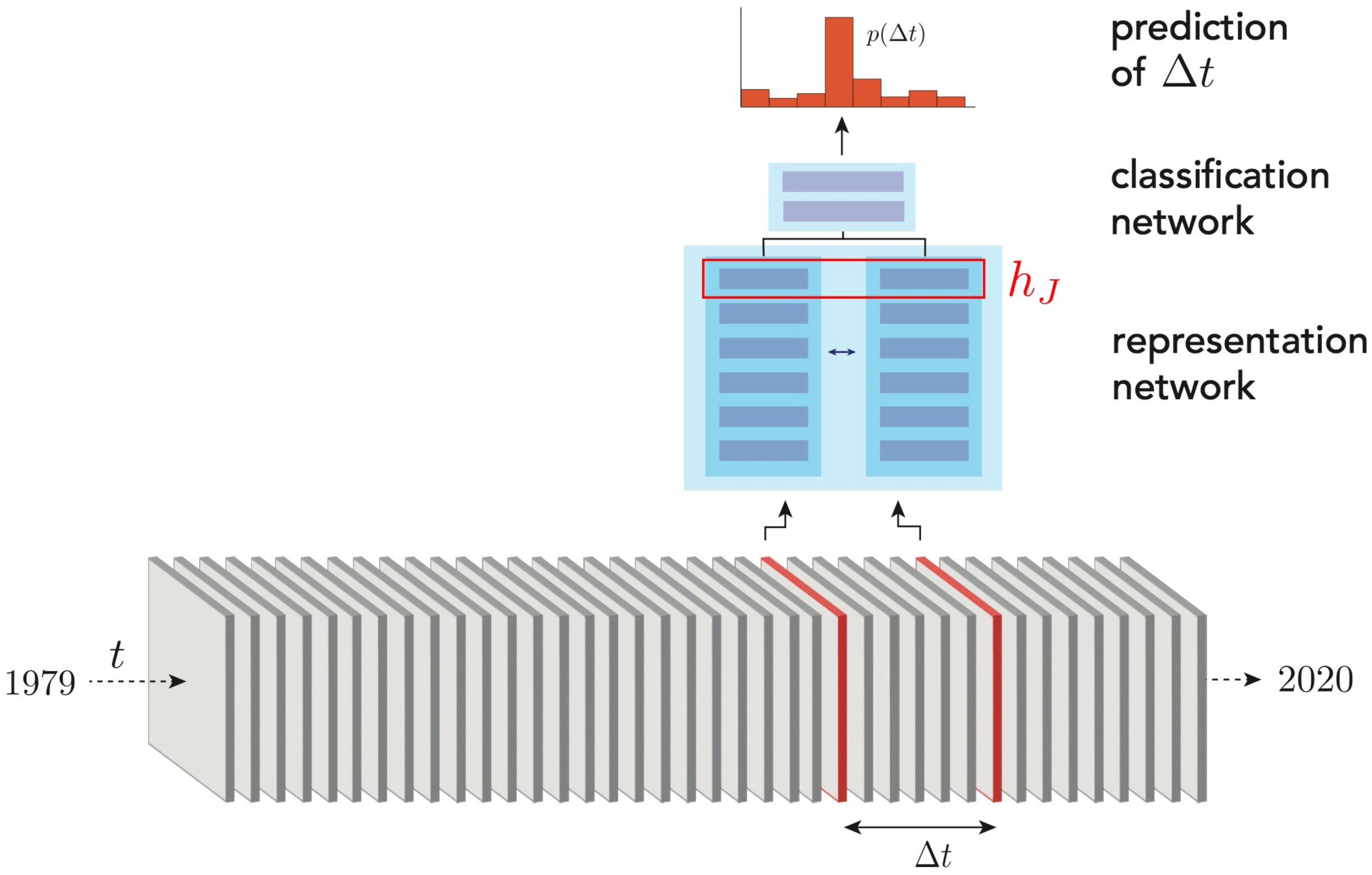
same network

Self-supervised learning tasks



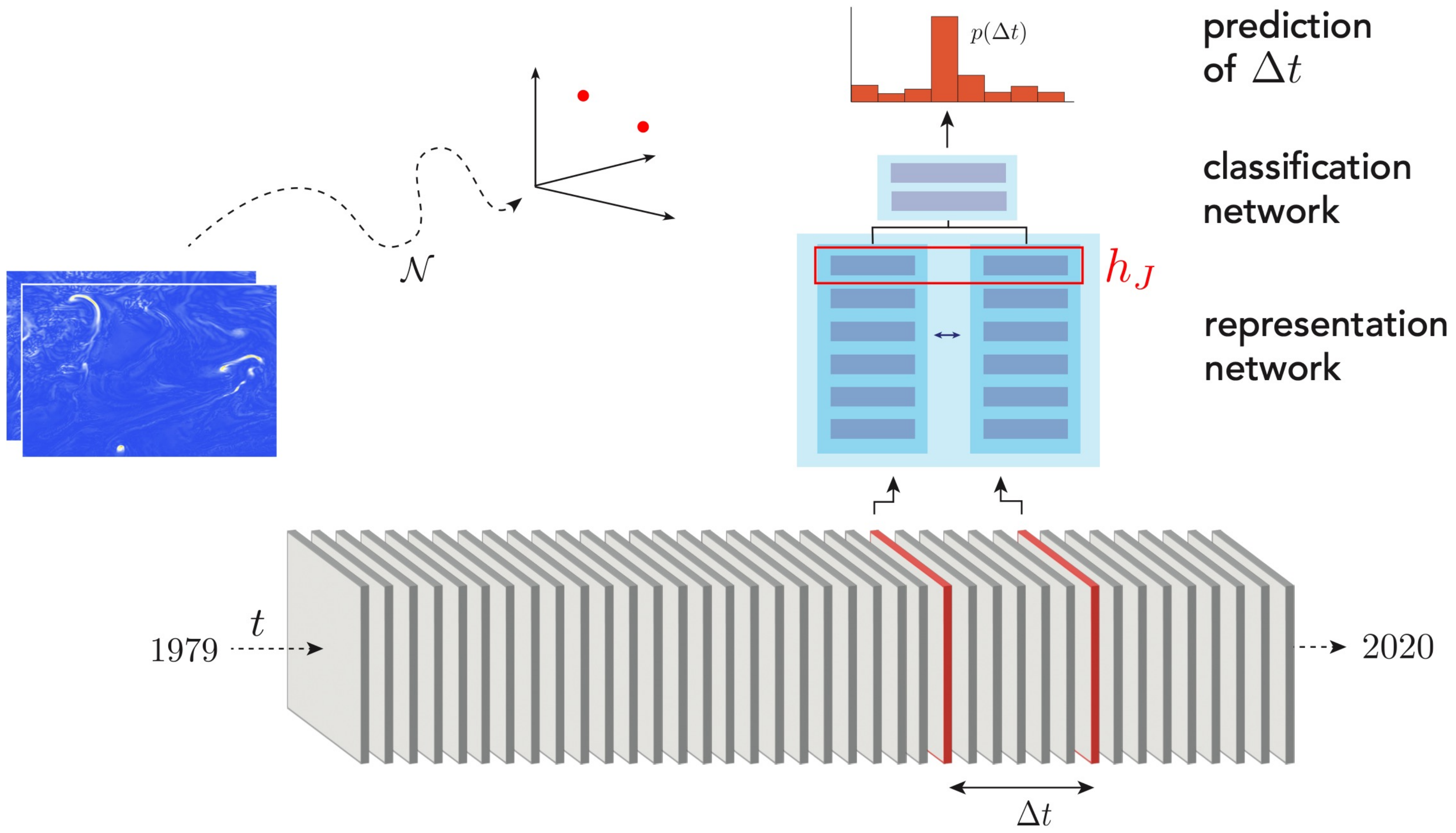
Hoffmann and Lessig, AtmoDist: Self-supervised Representation Learning for Atmospheric Dynamics, 2022, https://arxiv.org/abs/2202.01897

Self-supervised learning tasks



Hoffmann and Lessig, AtmoDist: Self-supervised Representation Learning for Atmospheric Dynamics, 2022, https://arxiv.org/abs/2202.01897

Self-supervised learning tasks



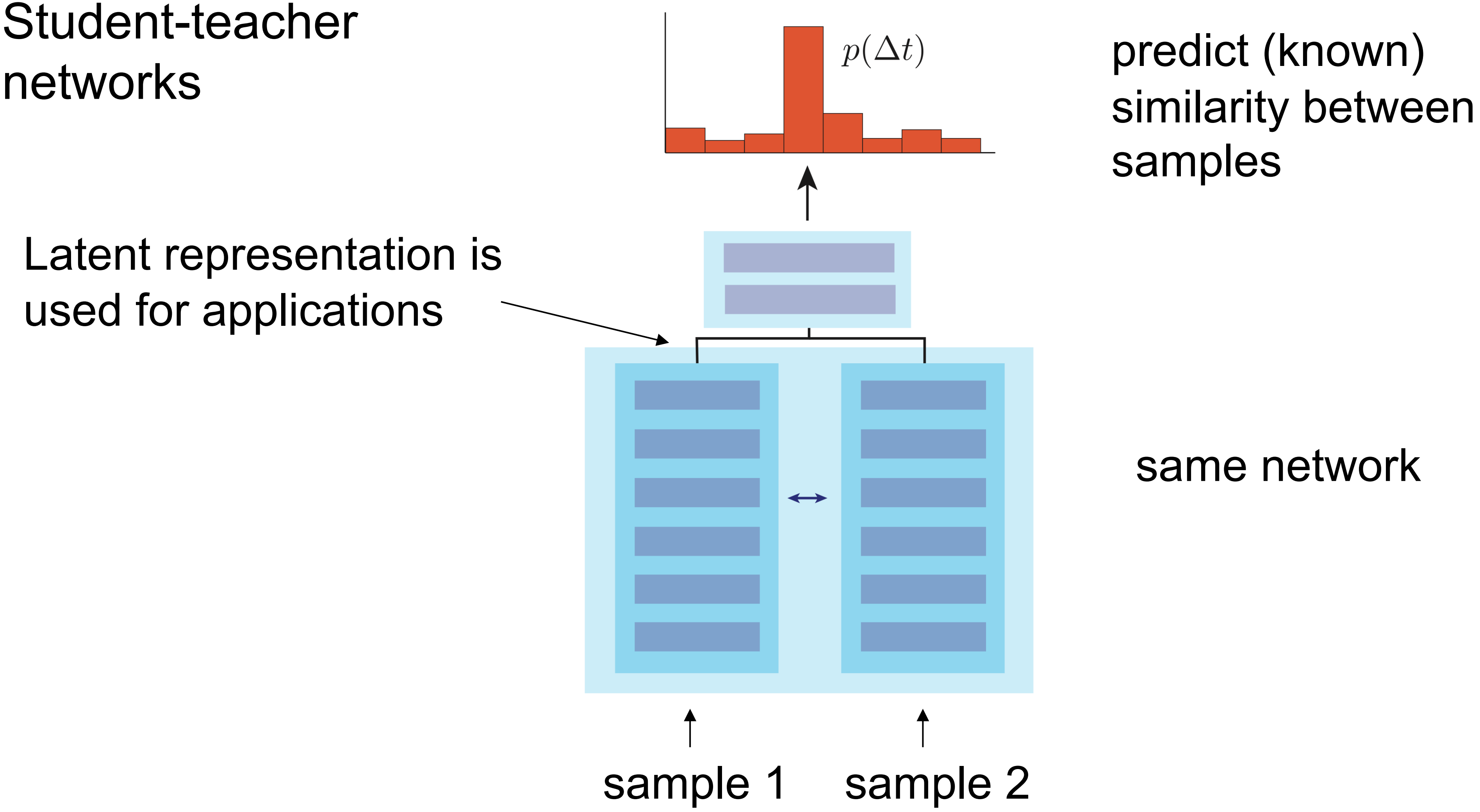
Hoffmann and Lessig, AtmoDist: Self-supervised Representation Learning for Atmospheric Dynamics, 2022, https://arxiv.org/abs/2202.01897

Self-supervised learning tasks

- Student-teacher networks

Self-supervised learning tasks

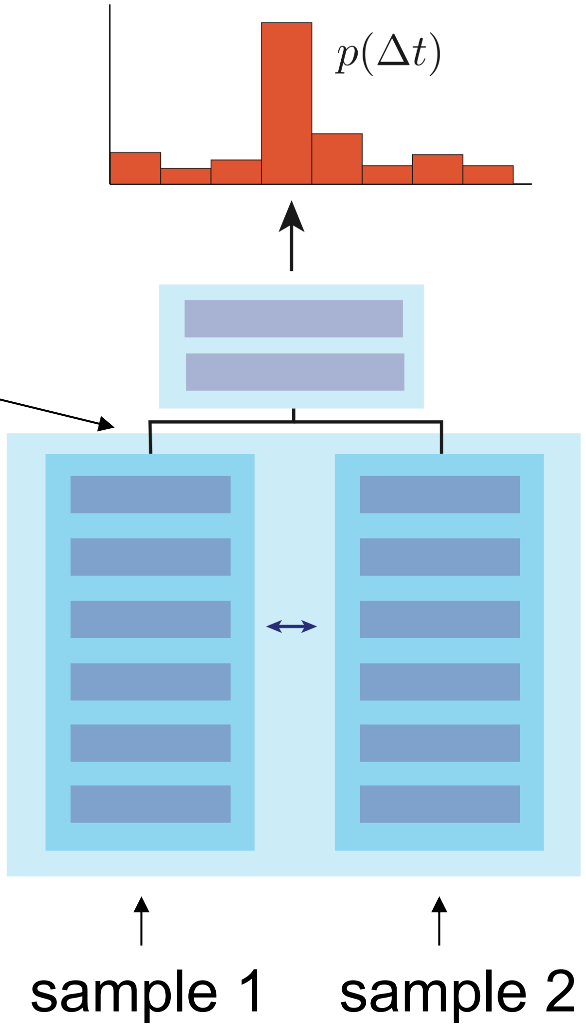
- Student-teacher networks



Self-supervised learning tasks

- Student-teacher networks

Latent representation is used for applications



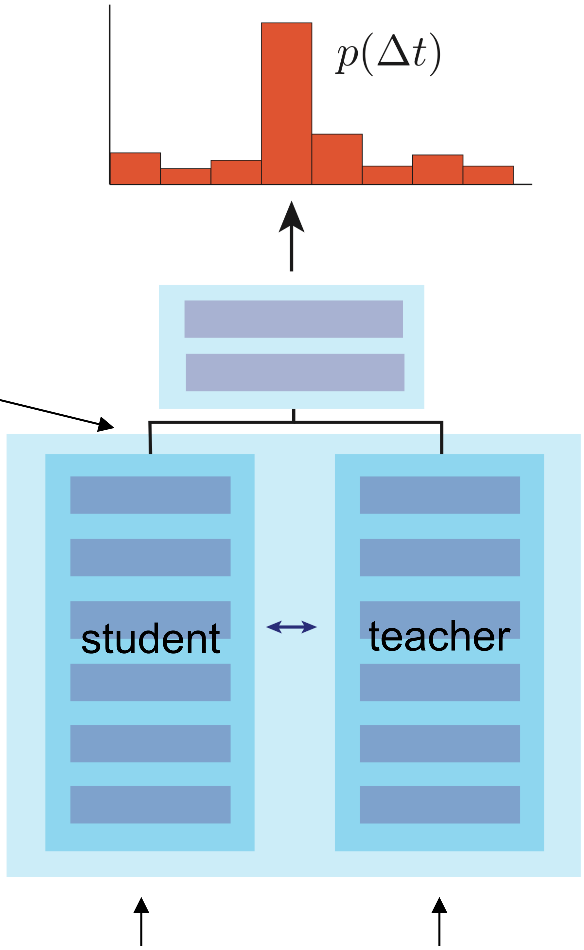
predict (known) similarity between samples

two networks so that weaker one can learn from the stronger one

Self-supervised learning tasks

- Student-teacher networks

Latent representation is used for applications



sample 1 sample 2

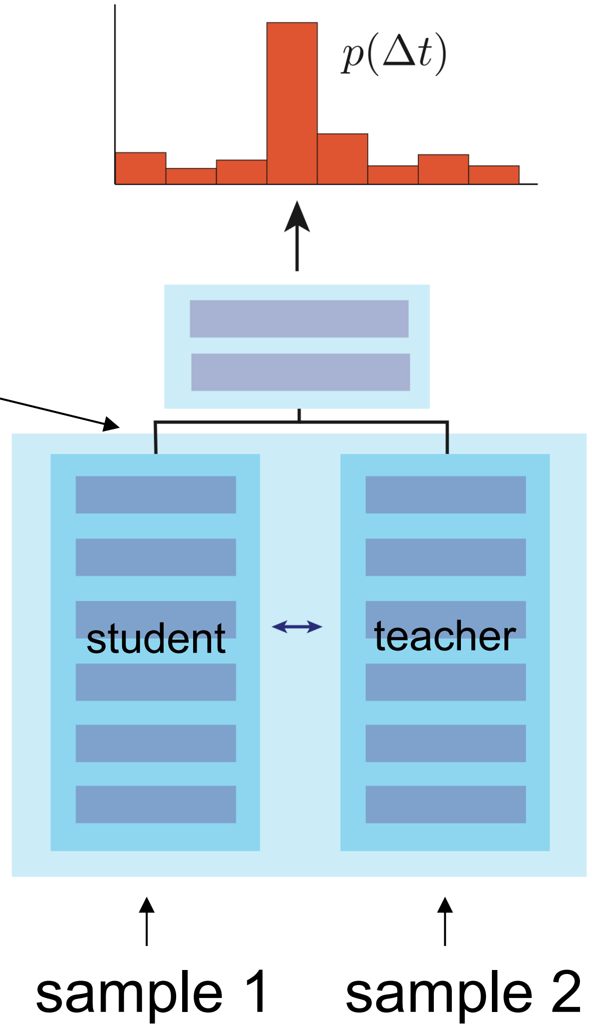
predict (known) similarity between samples

two networks so that weaker one can learn from the stronger one

Self-supervised learning tasks

- Student-teacher networks

Latent representation is used for applications



predict (known) similarity between samples

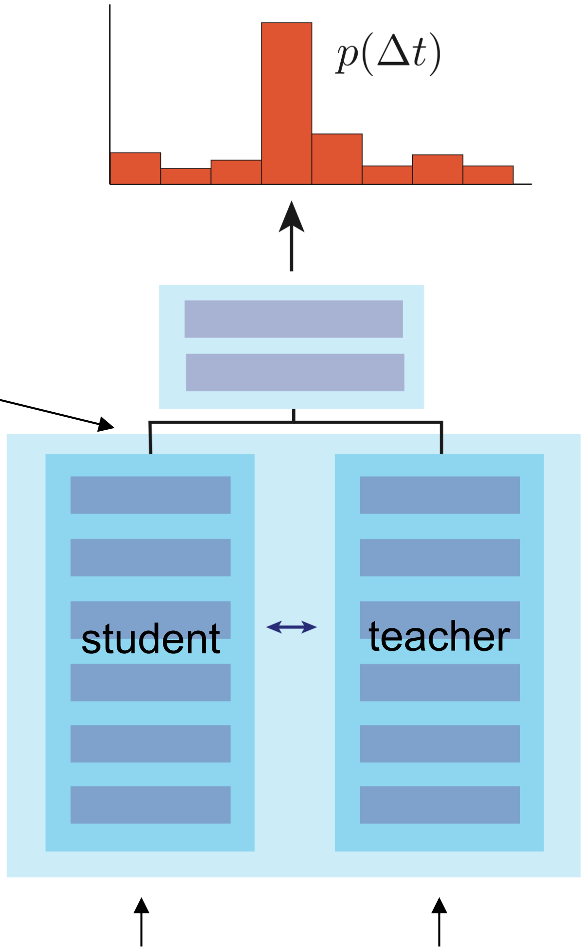
two networks so that weaker one can learn from the stronger one

teacher: take weighted average of student

Self-supervised learning tasks

- Student-teacher networks

Latent representation is used for applications



sample 1 sample 2

predict (known) similarity between samples

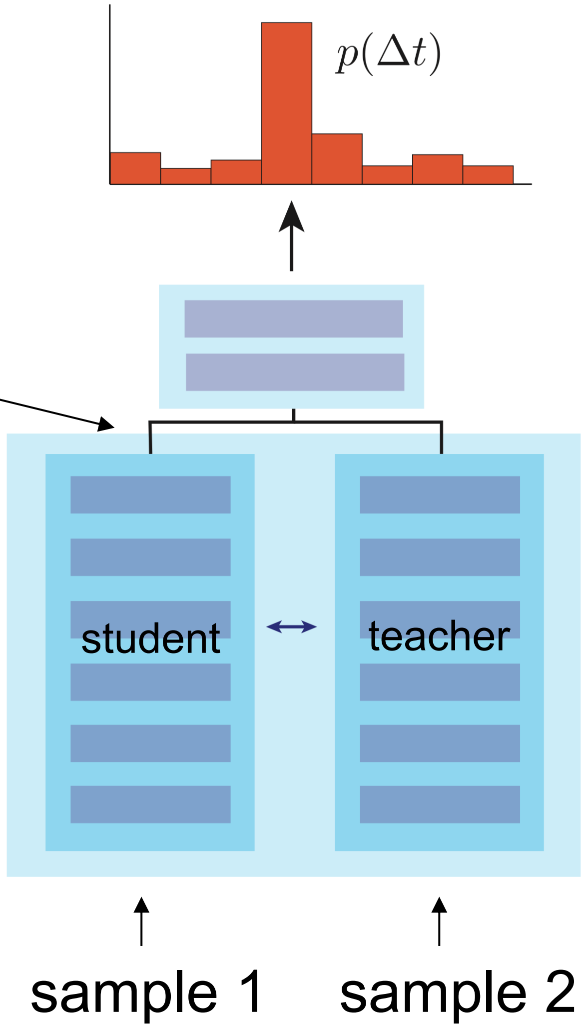
two networks so that weaker one can learn from the stronger one

teacher: take weighted average of student (exponential moving average (EMA))

Self-supervised learning tasks

- Student-teacher networks

Latent representation is used for applications



predict (known) similarity between samples

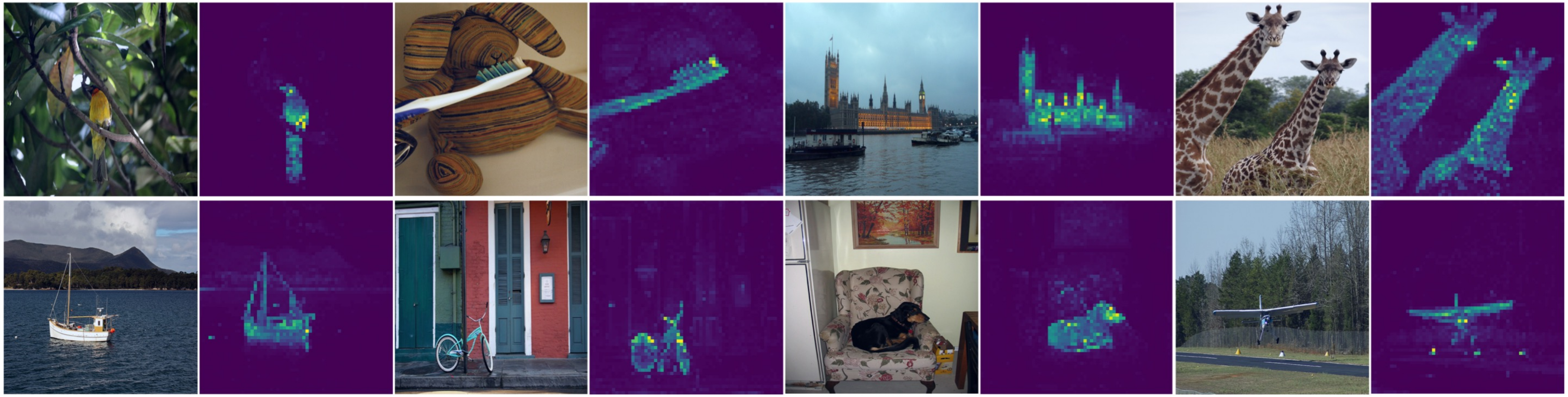
two networks so that weaker one can learn from the stronger one

teacher: take weighted average of student (exponential moving average (EMA))

https://www.slideshare.net/slideshow/enact-carrot-stick/53622951

Self-supervised learning tasks

- Student-teacher networks



Caron et al., Emerging Properties in Self-Supervised Vision Transformers, 2021, https://arxiv.org/abs/2104.14294

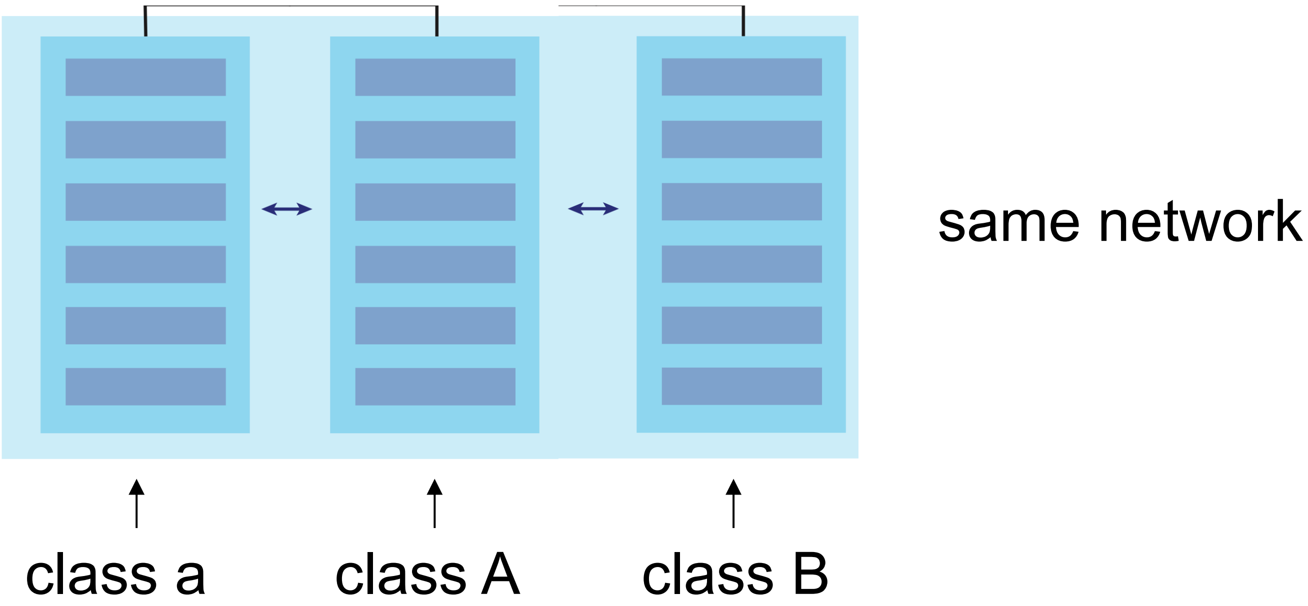
Self-supervised learning tasks

- Student-teacher networks

Caron et al., Emerging Properties in Self-Supervised Vision Transformers, 2021, https://arxiv.org/abs/2104.14294

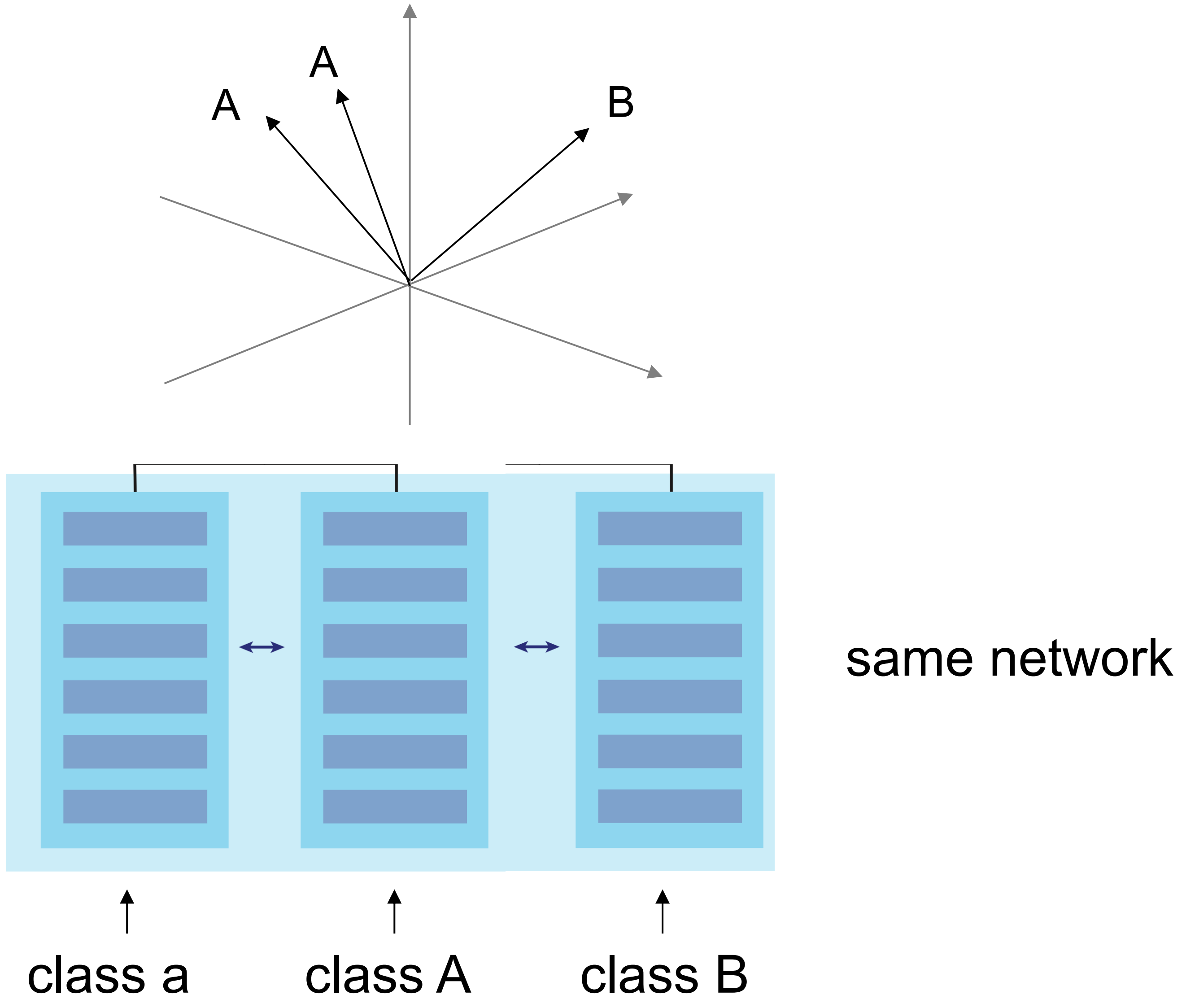
Self-supervised learning tasks

- Contrastive learning



Self-supervised learning tasks

- Contrastive learning



Summary

- Self-supervised learning
  - Overcome the limits imposed by requiring labeling of data
  - Learn task-agnostic neural networks
- Essentially all of the most powerful vision and language models use self-supervised training
  - Fine-tuning for specific applications
  - Increased robustness and flexibility

Literature

- Bengio et al., Representation Learning: A Review and New Perspectives, https://arxiv.org/abs/1206.5538
- https://ai.meta.com/blog/self-supervised-learning-the-dark-matter-of-intelligence/
- Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2018, https://arxiv.org/abs/1810.04805
- Radford et al., Improving Language Understanding by Generative Pre-Training, 2018, https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf
- Brown et al., Language Models are Few-Shot Learners, 2020, https://arxiv.org/abs/2005.14165.