Self-supervised learning

Christian Lessig
christian.lessig@ecmwf.int

ECMWF

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

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”.
  - Very limited generalization capabilities beyond training data

Why can this work at all?

- Small neural (e.g. a 10,000 parameter MLP) are interpolation “engines”.
  - Very limited generalization capabilities beyond training data

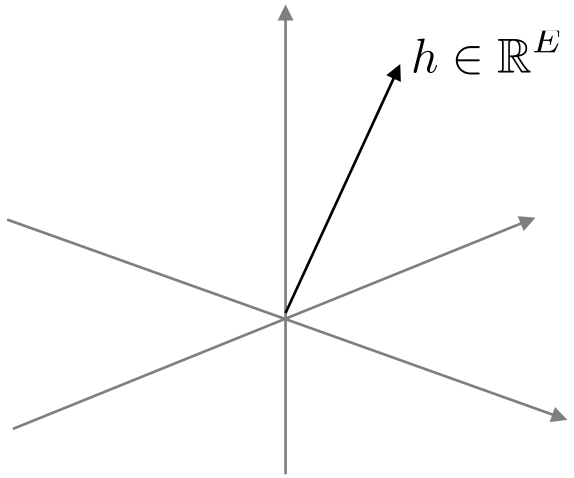
- Well-trained networks with 100s of millions or billions of parameters behave qualitatively differently and provide robust generalization
  - LLMs can answer a wide range of questions not seen during training
  - AIFS, Pangu-Weather, GraphCast provide skillful predictions multiple years past their training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

How can large neural networks generalize?

1. Feature spaces:

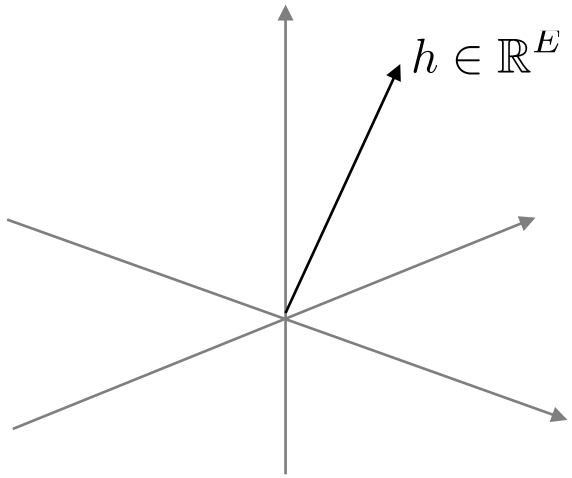
How can large neural networks generalize?

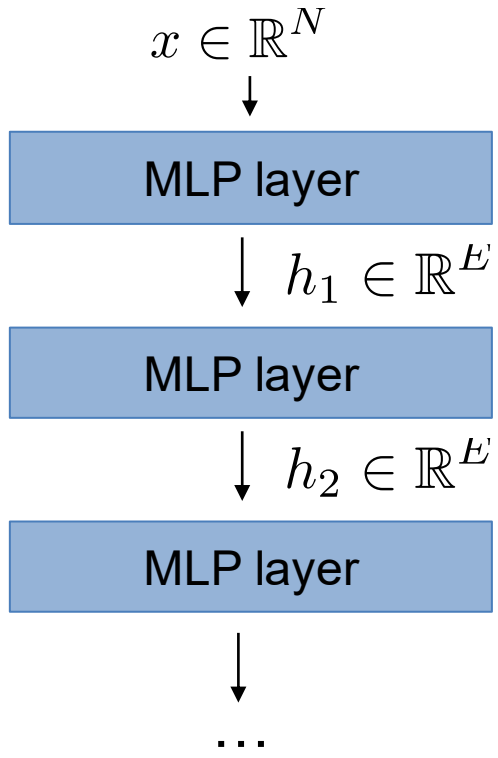
1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$



How can large neural networks generalize?

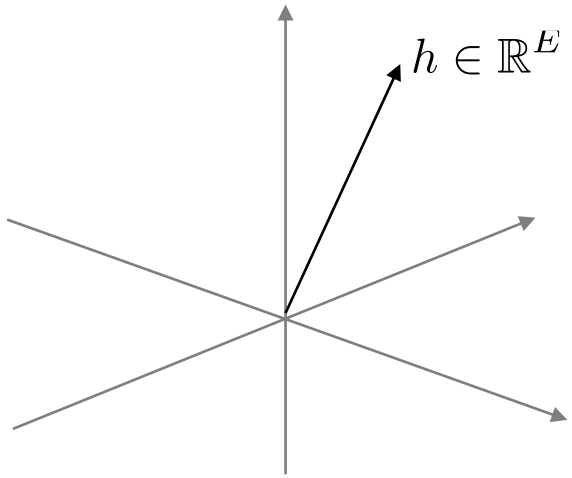
1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$

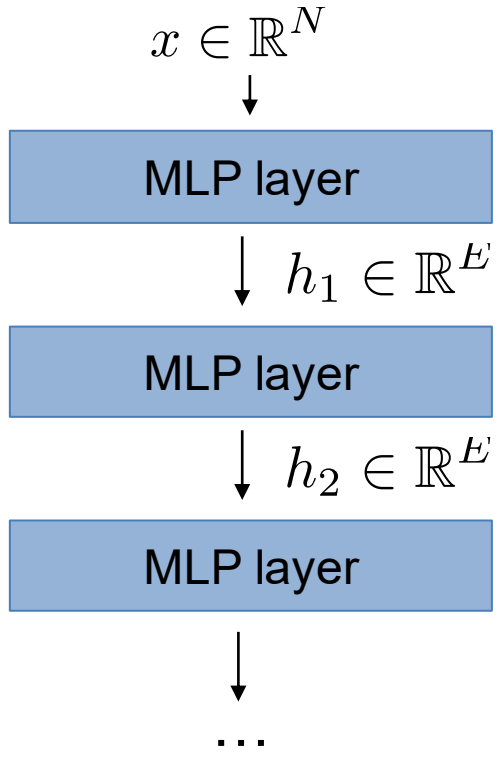




How can large neural networks generalize?

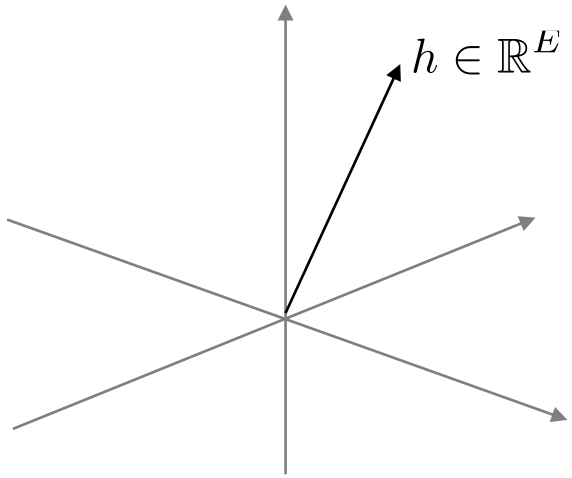
1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$
   - Feature spaces reveal important structures and remove noise

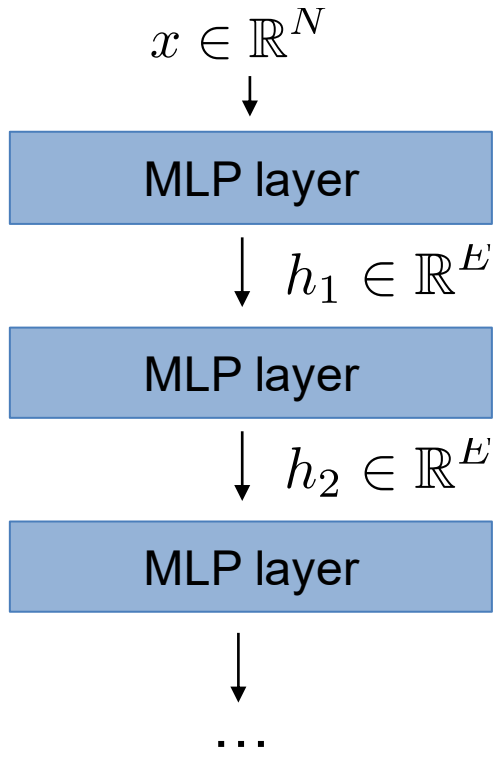




How can large neural networks generalize?

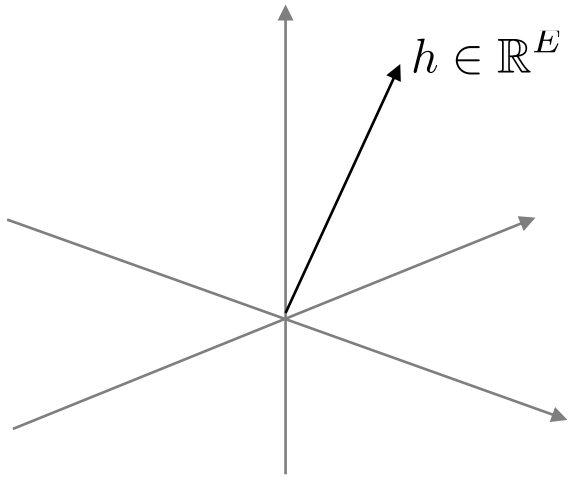
1. Feature spaces:
   - Hidden/latent state of a neural network is vector $h \in \mathbb{R}^E$
   - Feature spaces reveal important structures and remove noise
     - Analogous to Fourier domain, POD/PCA, …

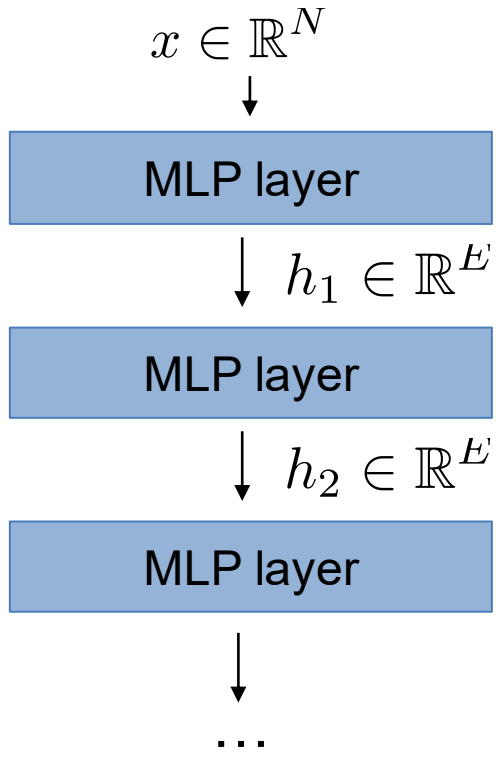




How can large neural networks generalize?

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





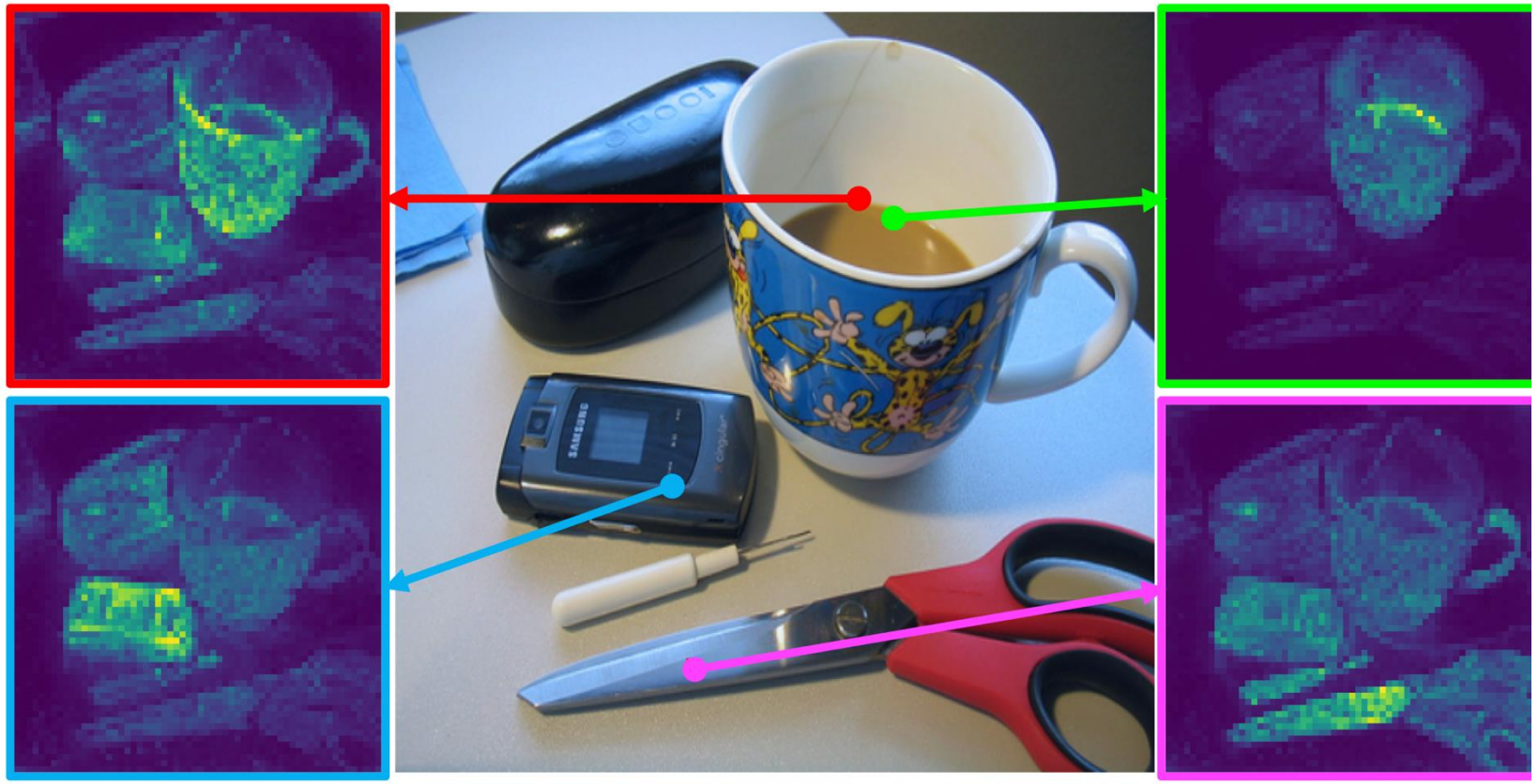
How can large neural networks generalize?

1. Feature spaces: convolutional features

https://arxiv.org/pdf/1506.06579

How can large neural networks generalize?

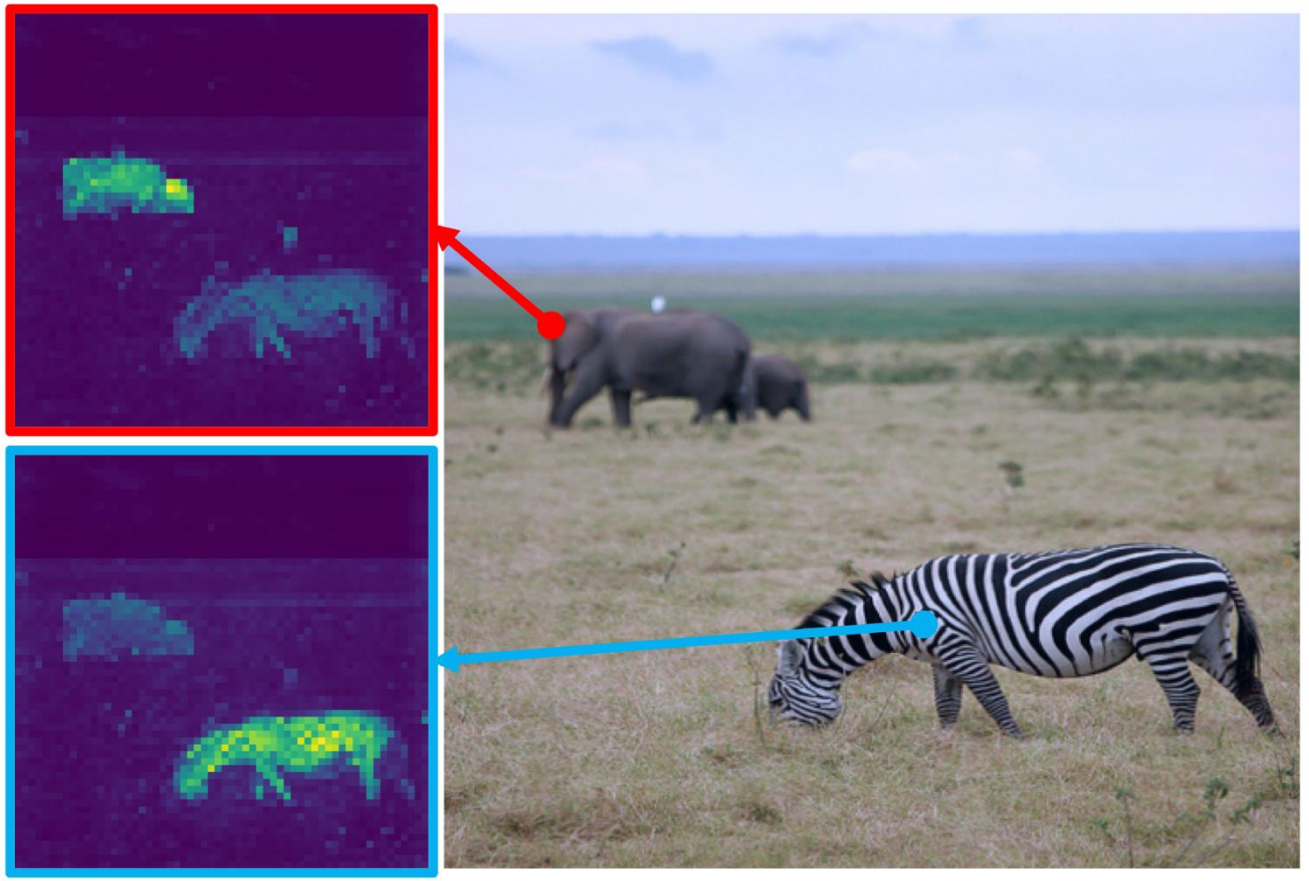
1. Feature spaces: attention maps



https://arxiv.org/pdf/2104.14294

How can large neural networks generalize?

1. Feature spaces: attention maps



https://arxiv.org/pdf/2104.14294

How can large neural networks generalize?

2. Learn general probabilistic model $p(y \mid x)$

How can large neural networks generalize?

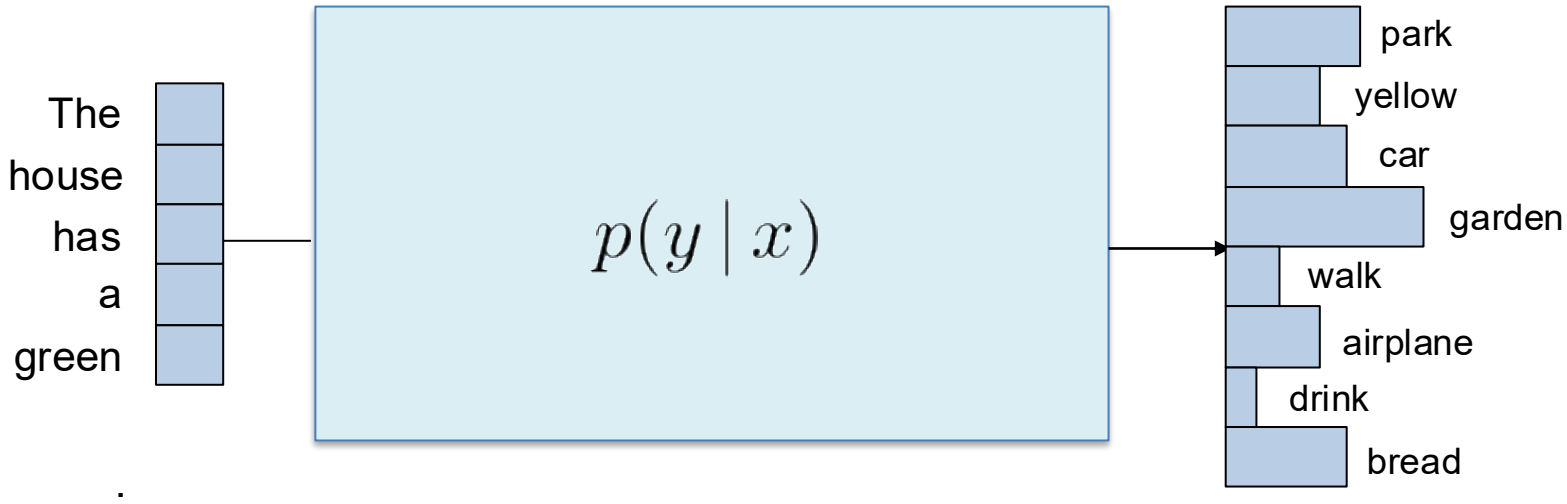
2. Learn general probabilistic model $p(y \mid x)$
   - When x, y are defined sufficiently general than this is task independent

How can large neural networks generalize?

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

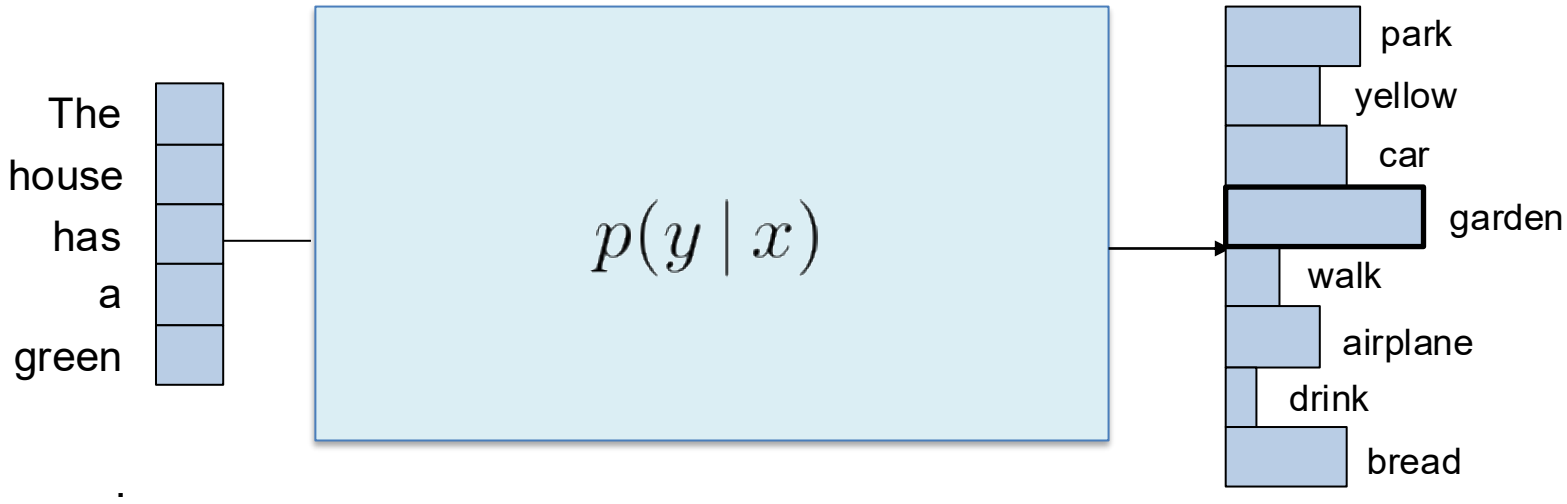
How can large neural networks generalize?

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



How can large neural networks generalize?

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



How can large neural networks generalize?

2. Learn general probabilistic model $p(y \mid x)$
   - When x, y are defined sufficiently general than this is task independent
   - E.g. $p(y \mid x)$ the joint distribution over natural language
     - Chat bot: x = question, y = answer

How can large neural networks generalize?

2. Learn general probabilistic model $p(y \mid x)$
   - When x, y are defined sufficiently general than this is task independent
   - E.g. $p(y \mid x)$ the joint distribution over natural language
     - Chat bot: x = question, y = answer
     - Translation: x = language A, y = language B

How can large neural networks generalize?

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

How can large neural networks generalize?

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

How can large neural networks generalize?

Zero-shot

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

```
1  Translate English to French:        ← task description
2  cheese =>                           ← prompt
```

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

How can large neural networks generalize?

One-shot

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

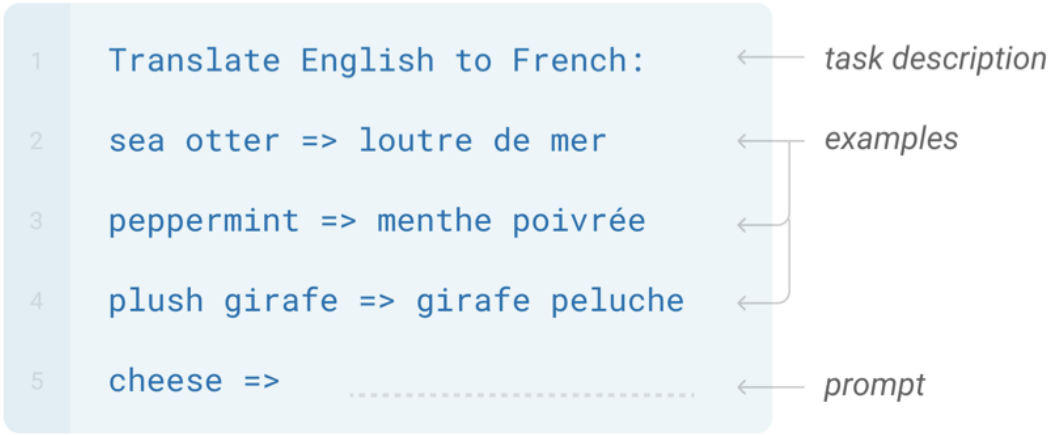
```
1  Translate English to French:        ← task description
2  sea otter => loutre de mer          ← example
3  cheese =>                           ← prompt
```

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

How can large neural networks generalize?

Few-shot

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



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

How can large neural networks generalize?

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

How can large neural networks generalize?

2. Learn general probabilistic model $p(y \mid x)$
   - When x, y are defined sufficiently general than this is task independent
   - E.g. $p(y \mid x)$ 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
   - Fine-tuning to calibrate the output distribution
     - E.g. make certain undesirable words/phrases less likely

How can large neural networks generalize?

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

How can large neural networks generalize?

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

How can large neural networks generalize?

2. Learn general probabilistic model $p(y \mid x)$
   - When x, y are defined sufficiently general than this is task independent
   - E.g. $p(y \mid x)$ is the joint distribution over atmospheric states
     - Forecasting: x = current state, y = future state

How can large neural networks generalize?

2. Learn general probabilistic model $p(y \mid x)$
   - When x, y are defined sufficiently general than this is task independent
   - E.g. $p(y \mid x)$ is the joint distribution over atmospheric states
     - Forecasting: x = current state, y = future state
     - Downscaling: x = coare res. state, y = fine res. state

How can large neural networks generalize?

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

How can large neural networks generalize?

Two perspectives:

1. Feature spaces
   - Learning yields representation of data that reveals intrinsic structure
2. Learn general probabilistic model $p(y \mid x)$
   - Training objective is a priori task-independent

How can large neural networks generalize?

Two perspectives:

1. Feature spaces
   - Learning yields representation of data that reveals intrinsic structure
2. Learn general probabilistic model $p(y \mid x)$
   - Training objective is a priori task-independent

=> Self-supervised learning to realize this

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 data when input to the network, 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

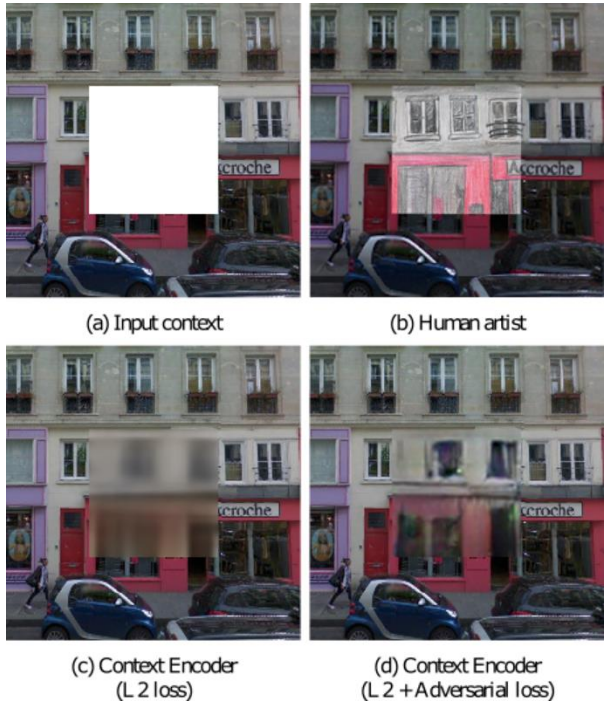
=> “hide” some information from the data when input to the network, and network predicts this information

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 data when input to the network, and network predicts this information

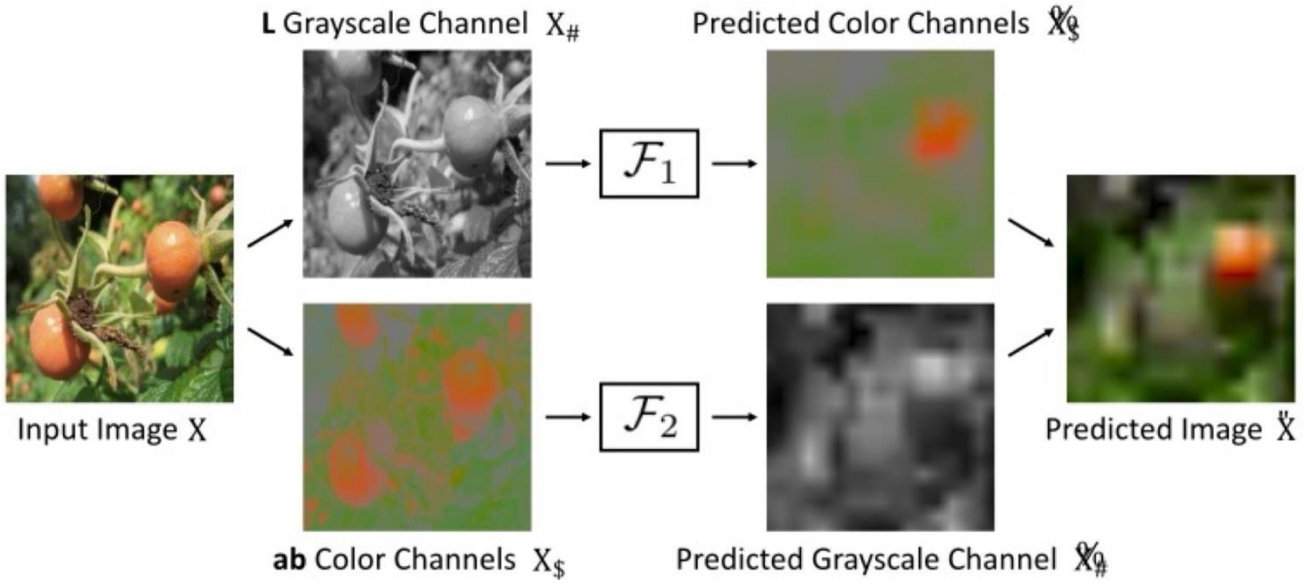


D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

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 data when input to the network, 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 data when input to the network, and network predicts this information

Transformer takes sequence of data chunks (tokens) as input

- ((sub-)words, image patches, local atmospheric states, …)

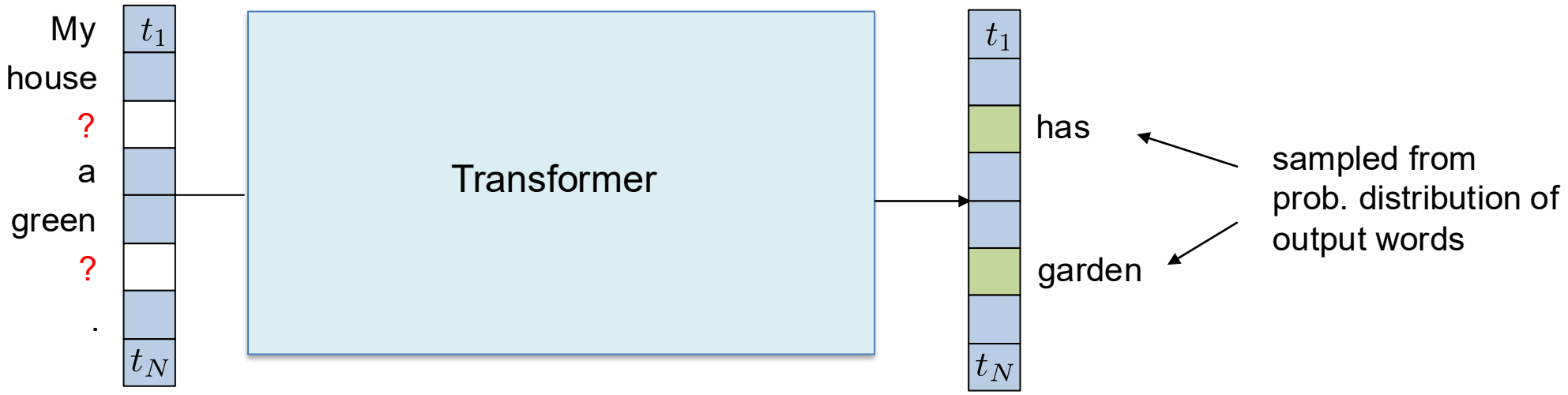
=> mask some of the patches from the network during input (or remove them entirely) and network predicts these

Self-supervised learning tasks

Self-supervised learning tasks

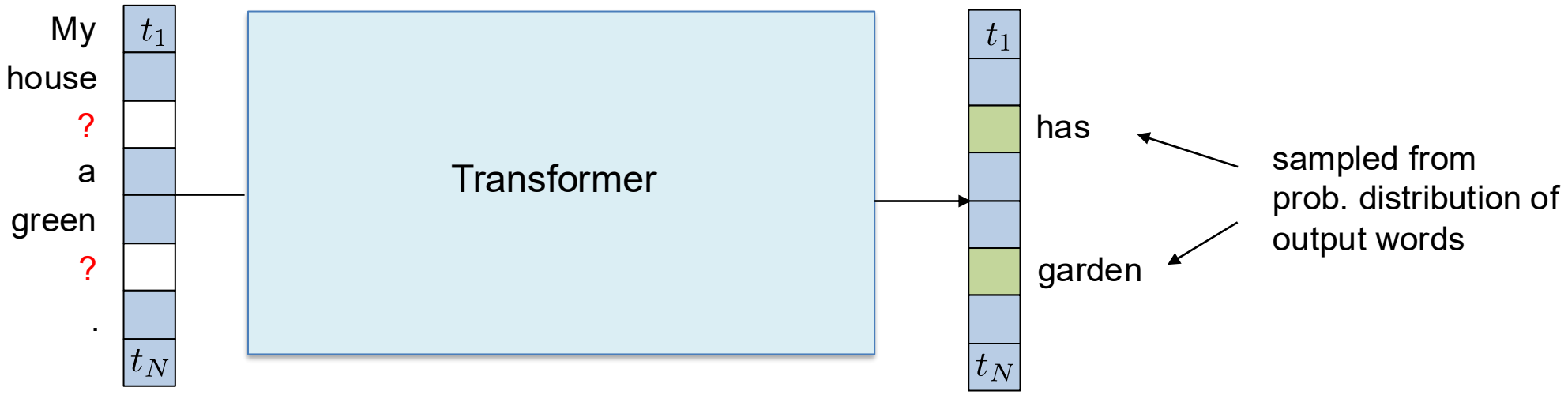
Self-supervised learning tasks

Self-supervised learning tasks



Self-supervised learning tasks

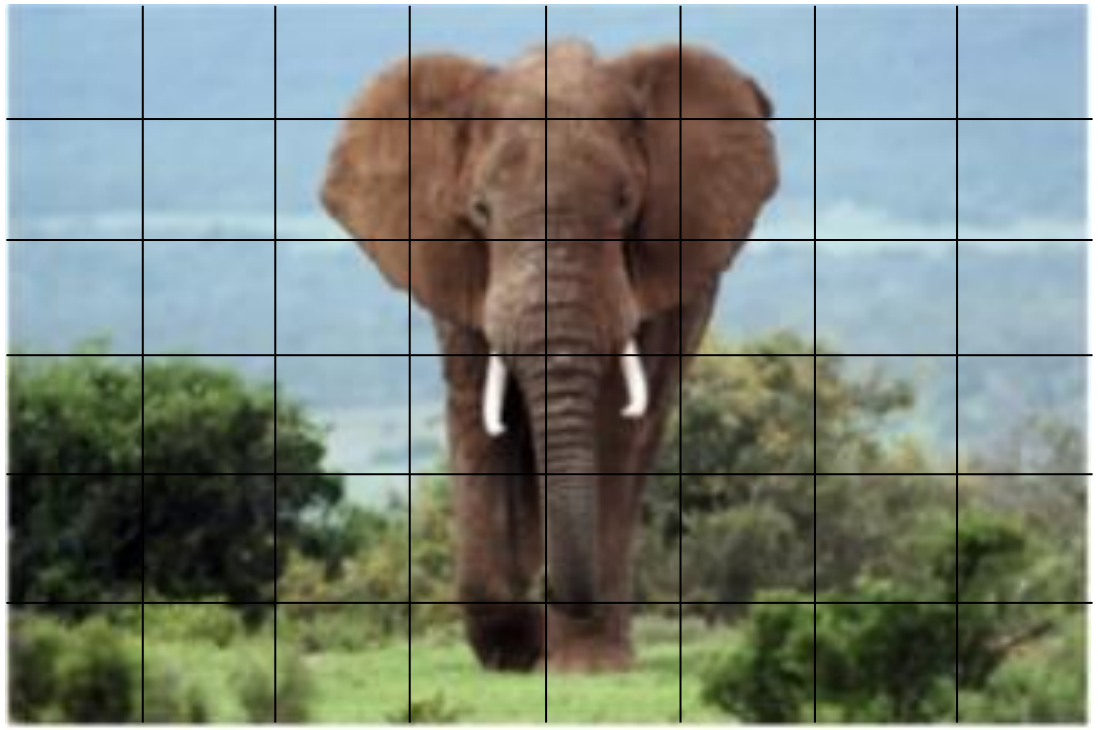
- BERT (Google):[1] randomly mask words from a sequence (and add some random distortions)
- Predictive masking (OpenAI):[2] always mask subsequent words



[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

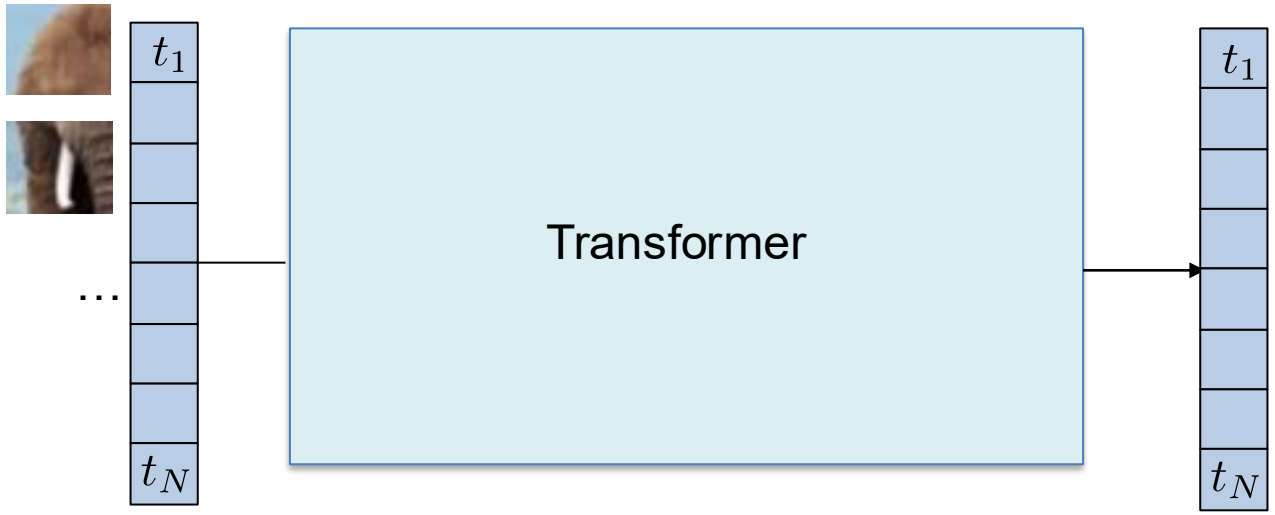
Vision transformer: image is a small patch



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

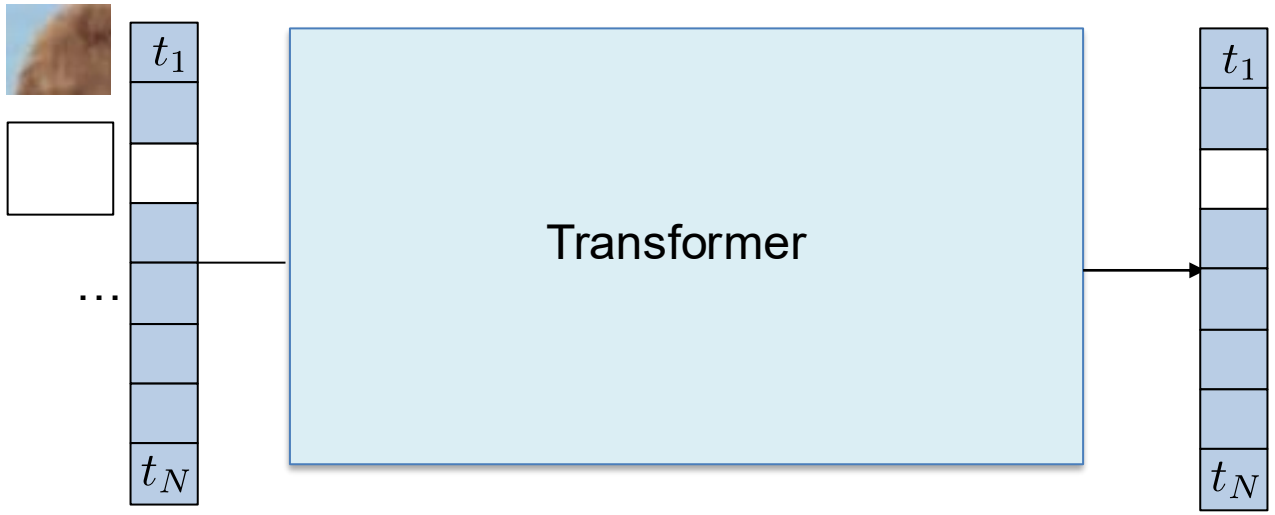
Masked token modeling for images



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

Self-supervised learning tasks

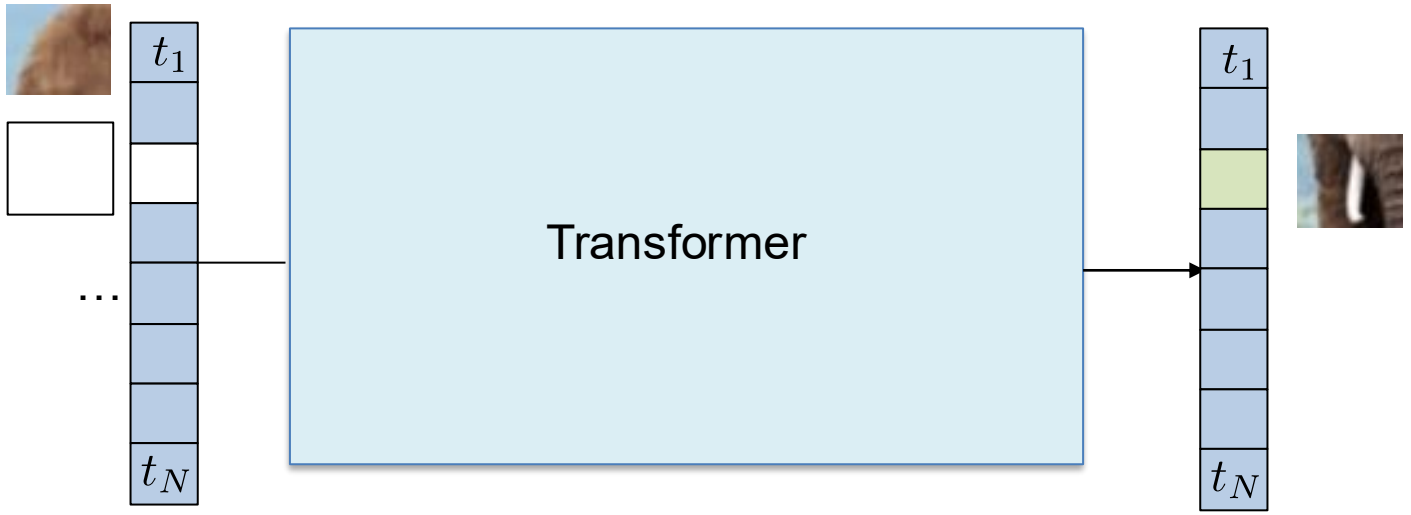
Masked token modeling for images



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

Self-supervised learning tasks

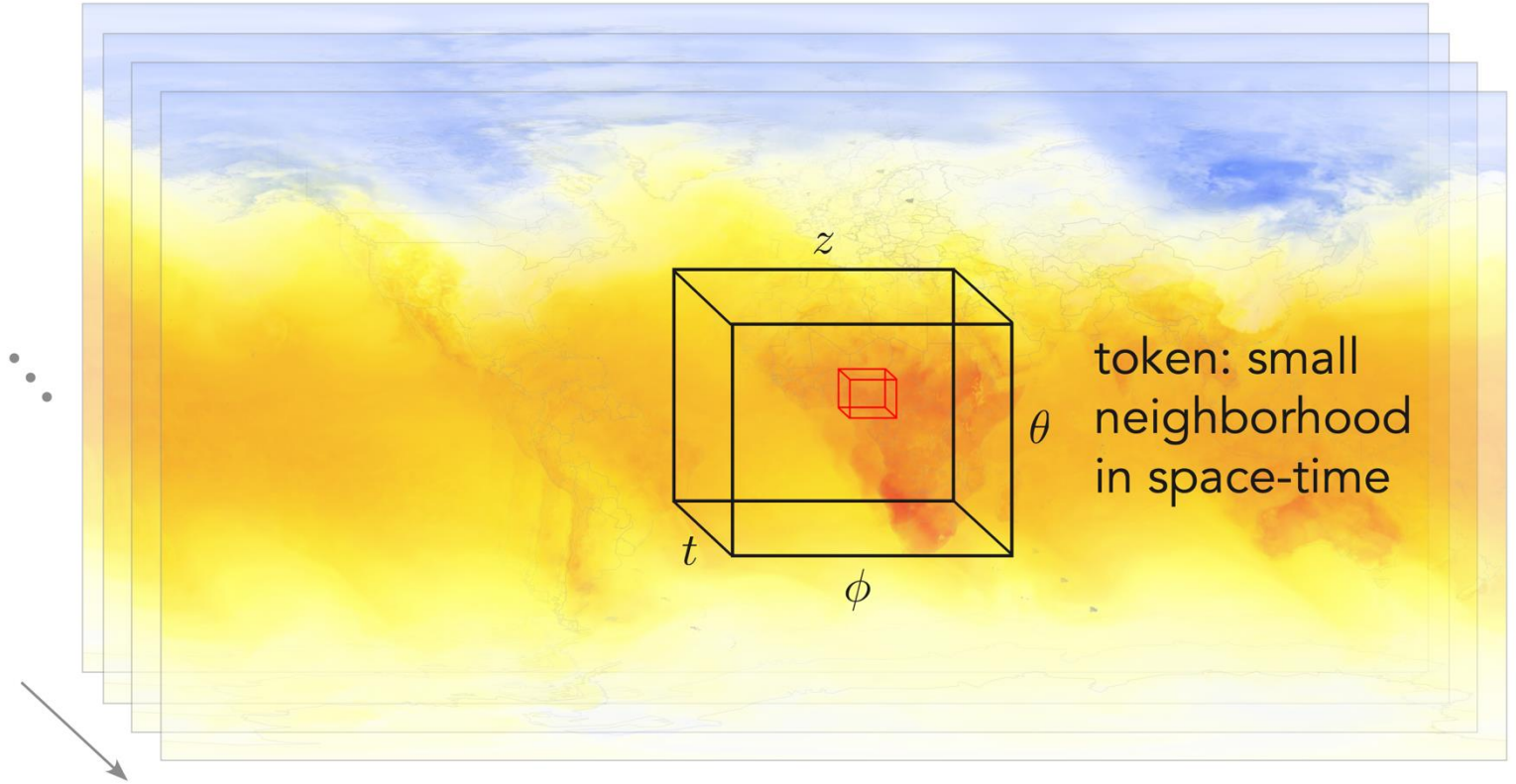
Masked token modeling for images



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

Self-supervised learning tasks

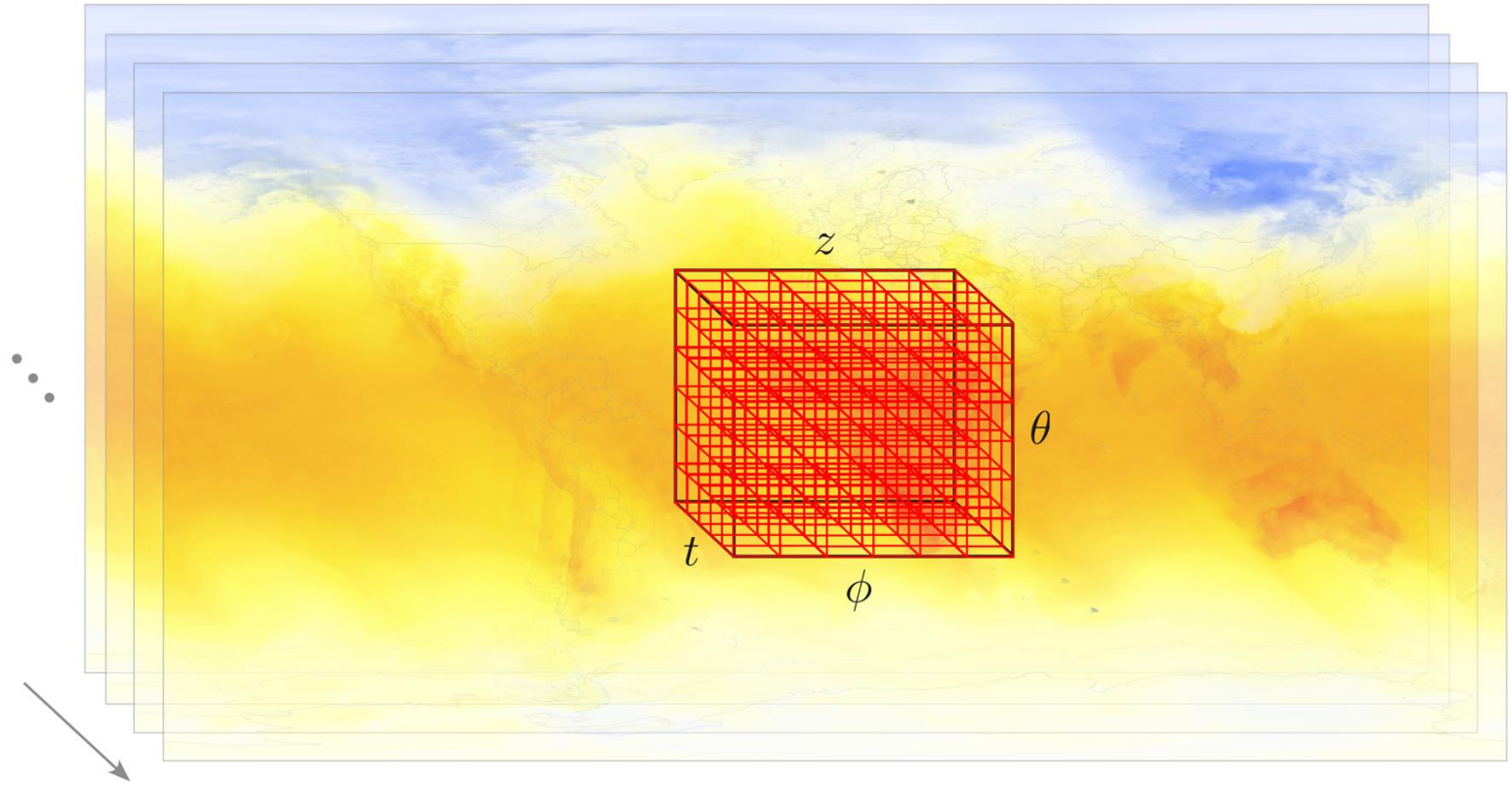
AtmoRep: masked token modeling for atmospheric dynamics



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

Self-supervised learning tasks

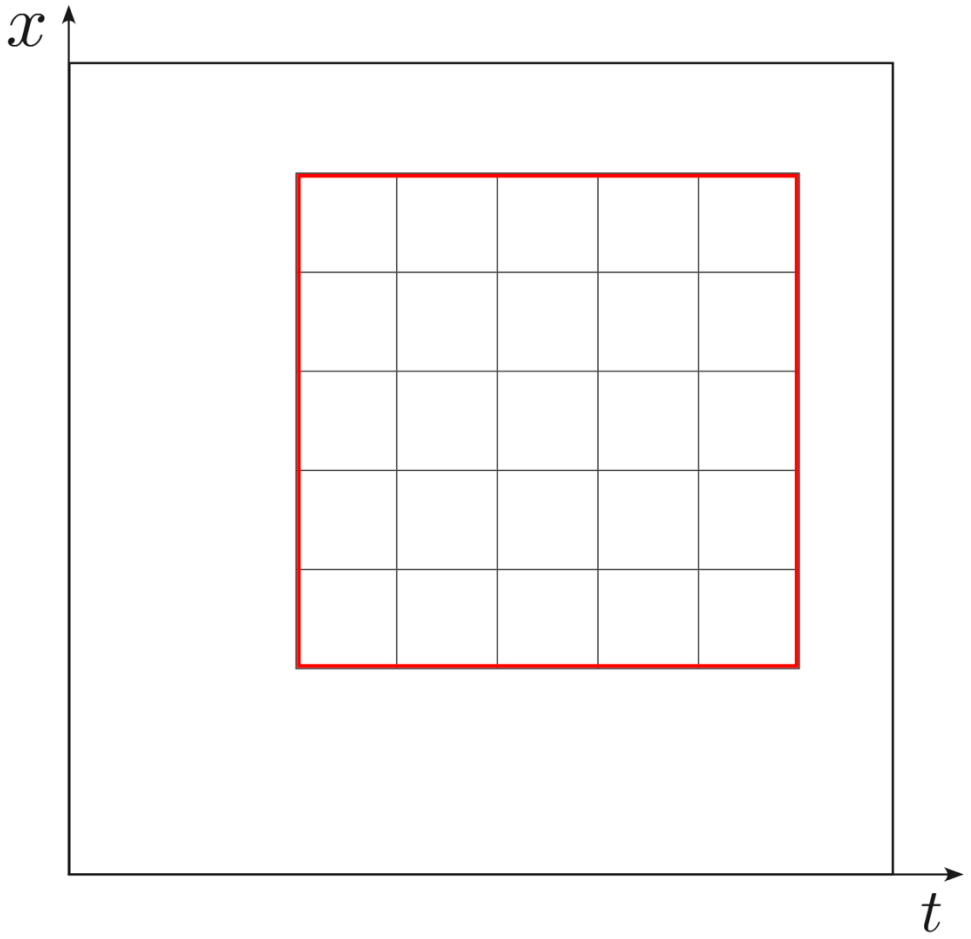
AtmoRep: masked token modeling for atmospheric dynamics



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

Self-supervised learning tasks

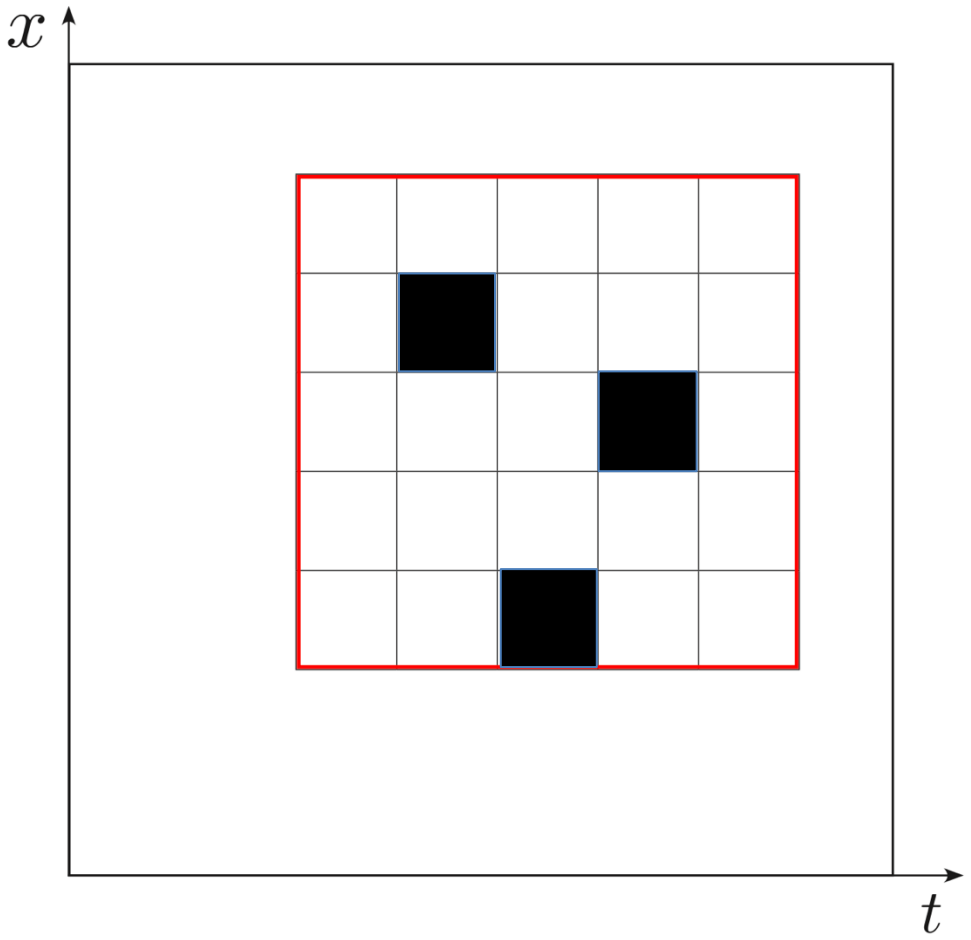
Flatland view



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

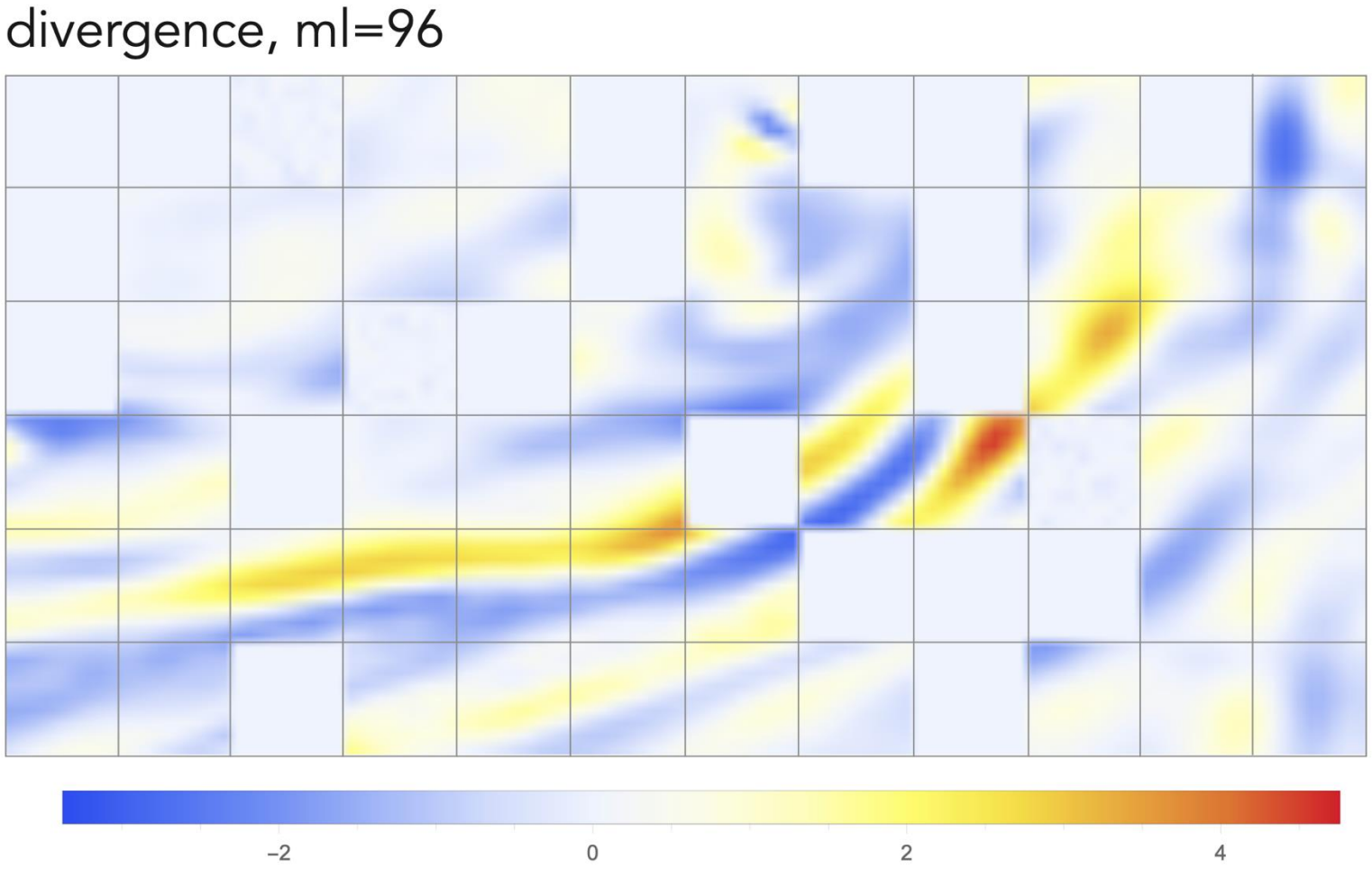
Self-supervised learning tasks

Flatland view



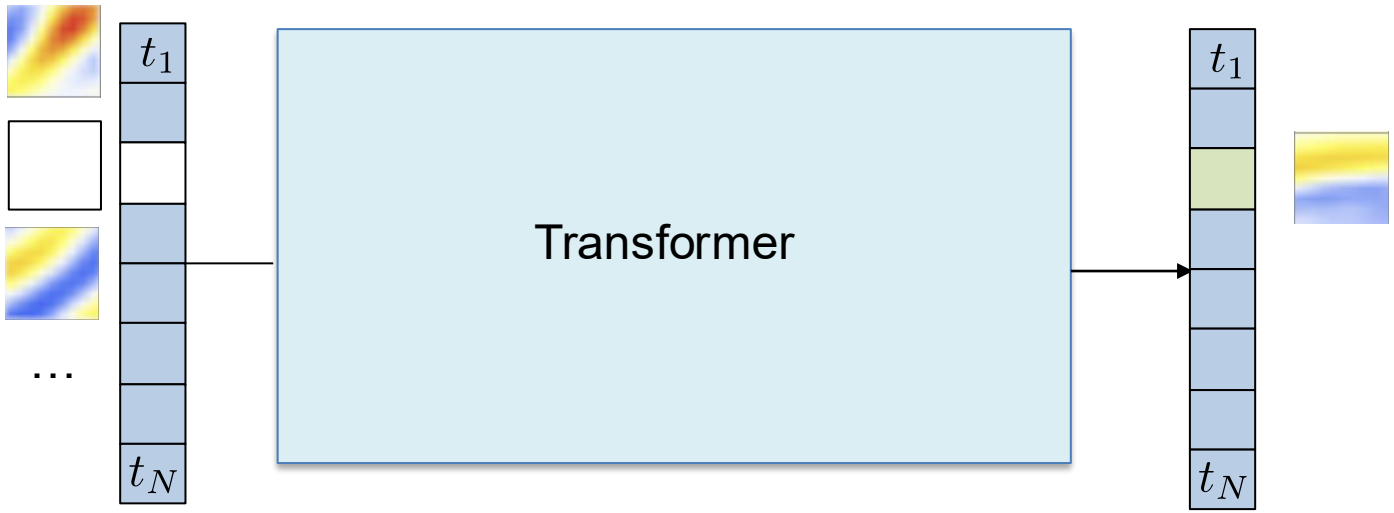
Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks



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

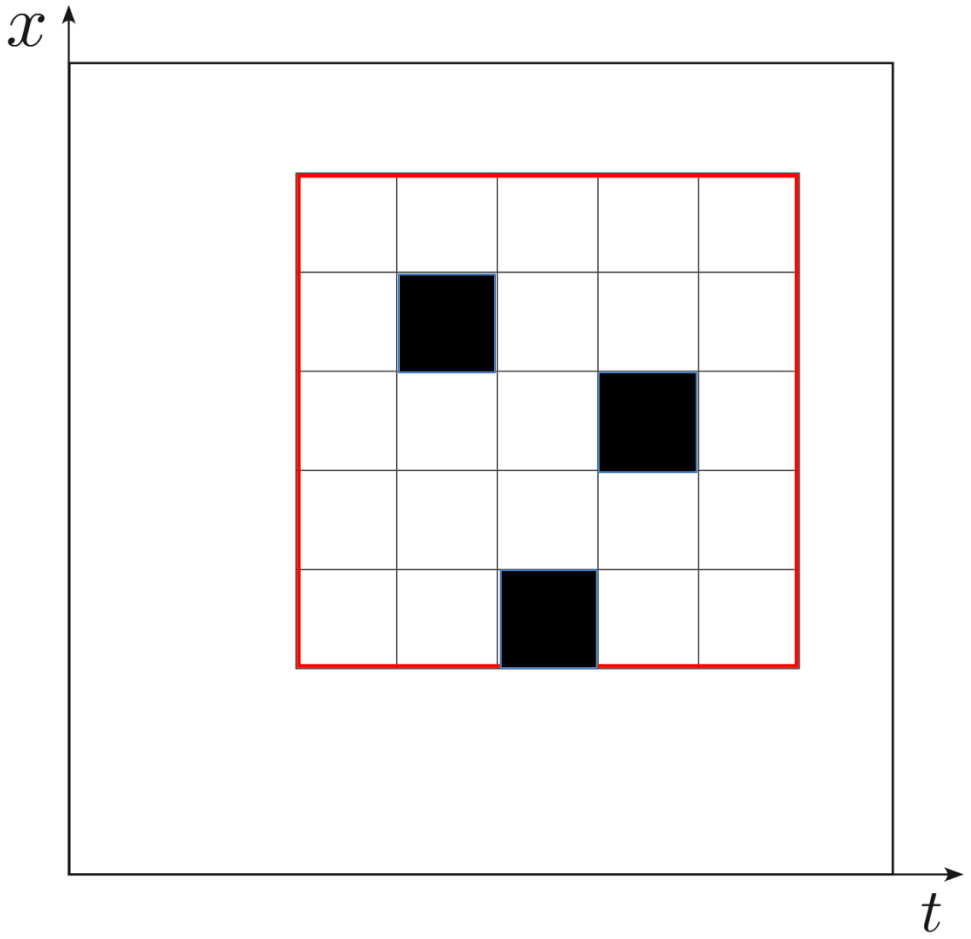
Self-supervised learning tasks



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

Self-supervised learning tasks

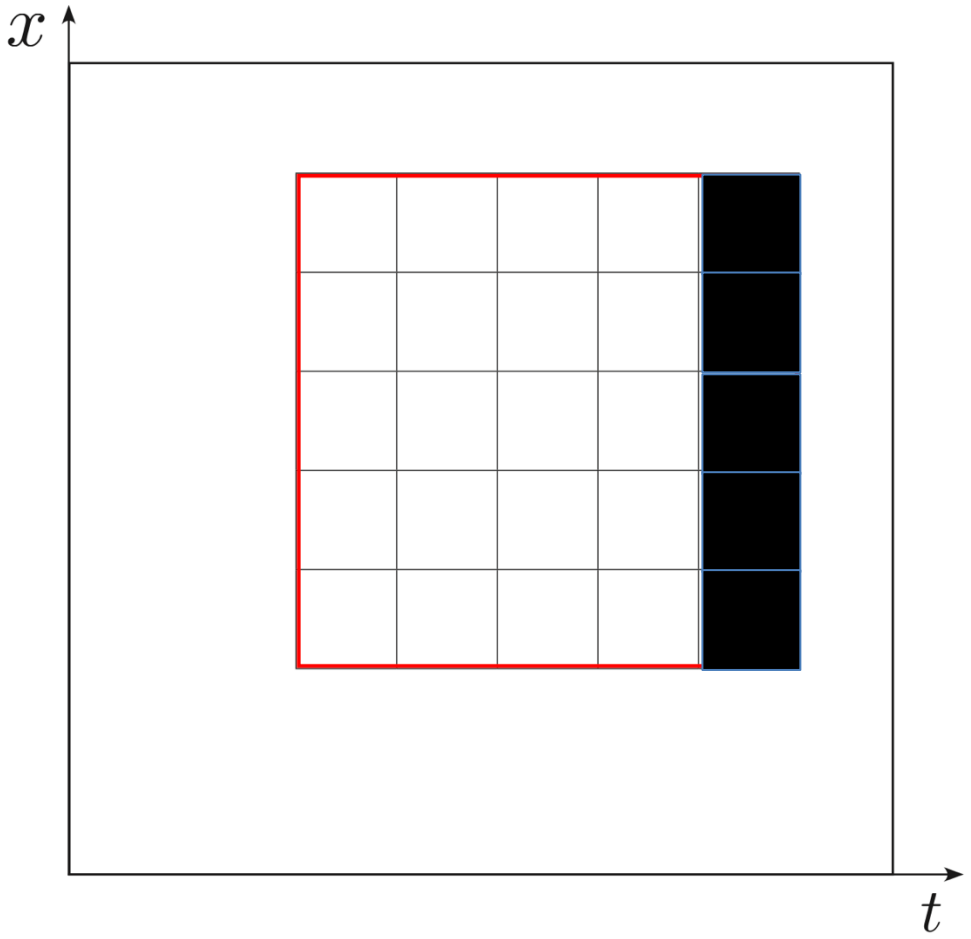
Zero-shot capabilities



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

Self-supervised learning tasks

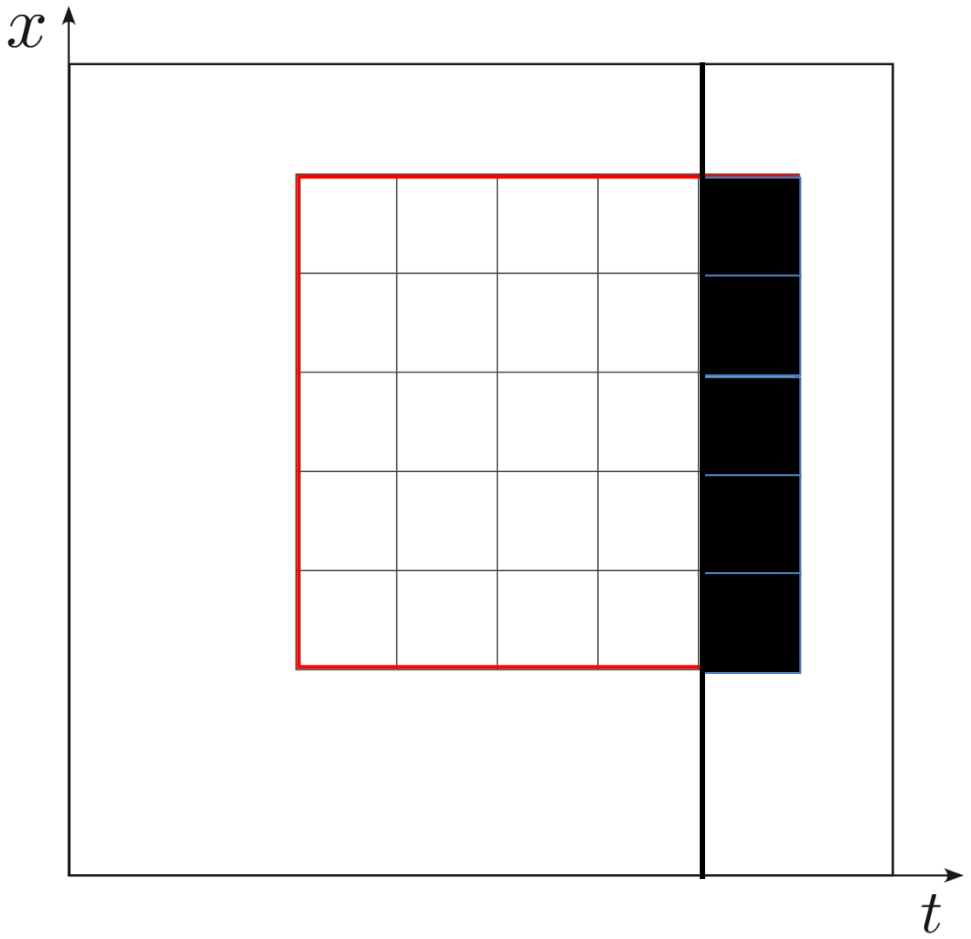
Zero-shot capabilities



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

Self-supervised learning tasks

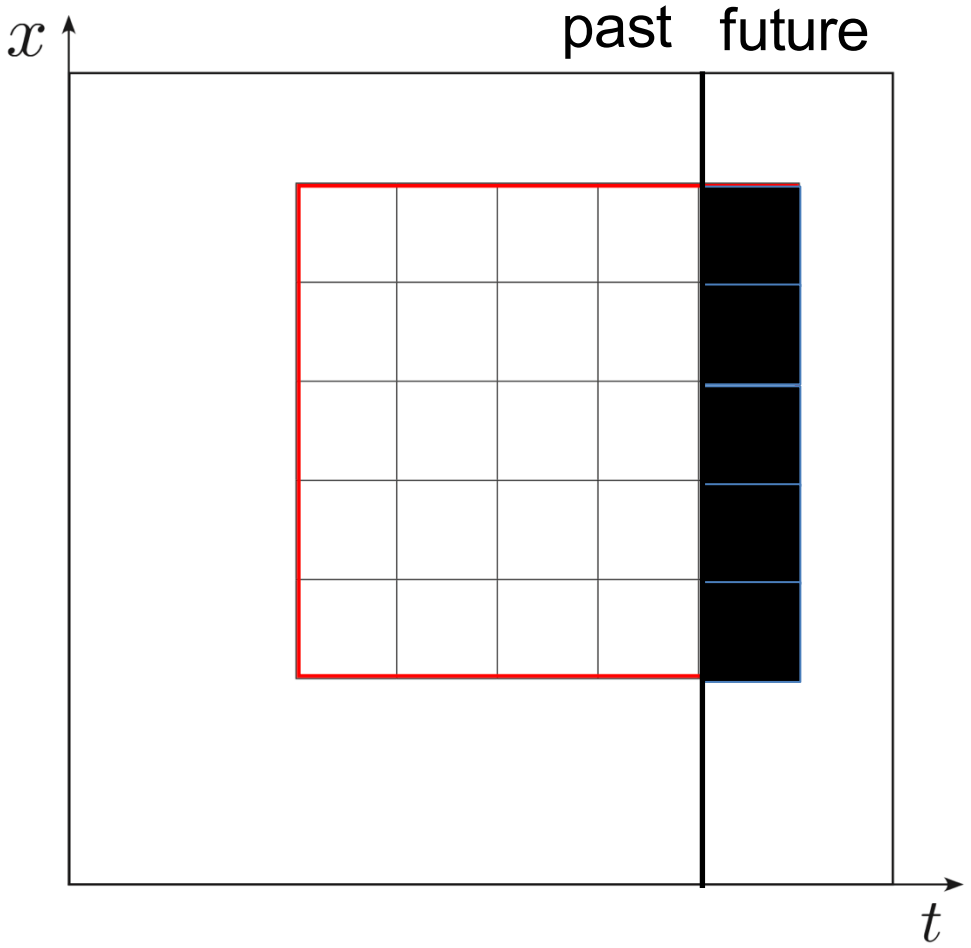
Zero-shot capabilities



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

Self-supervised learning tasks

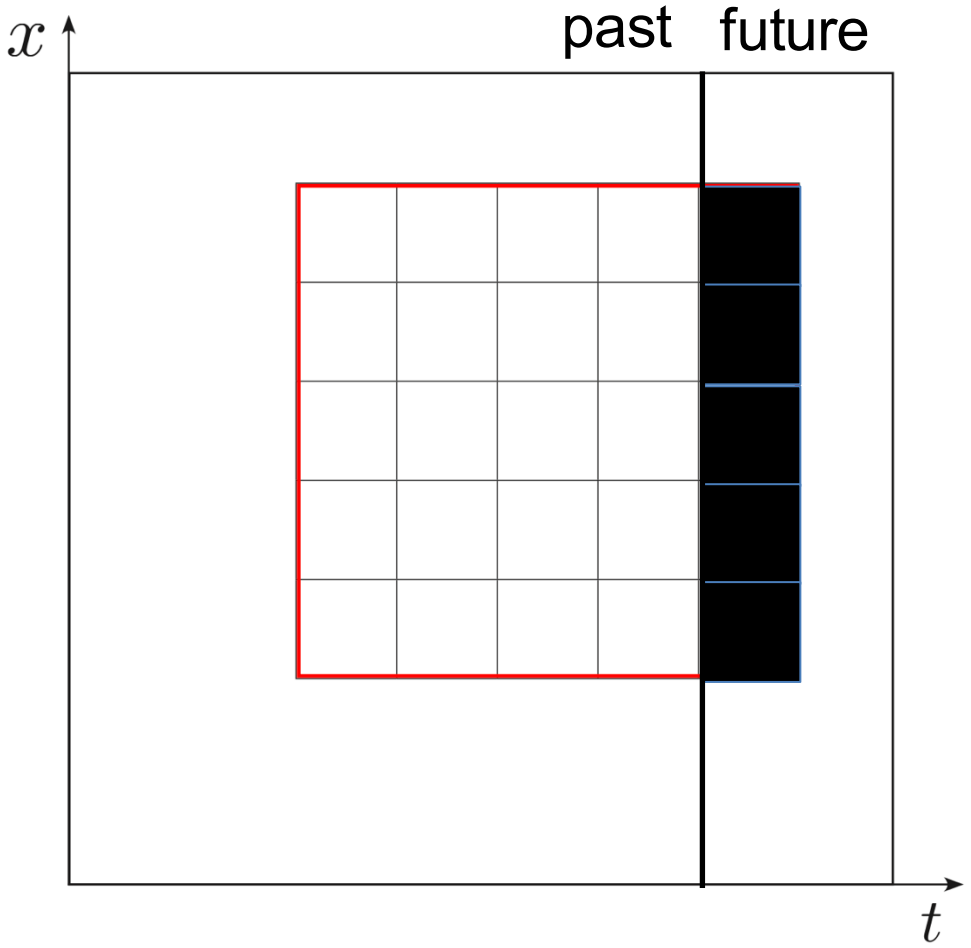
Zero-shot capabilities



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

Self-supervised learning tasks

Zero-shot capabilities

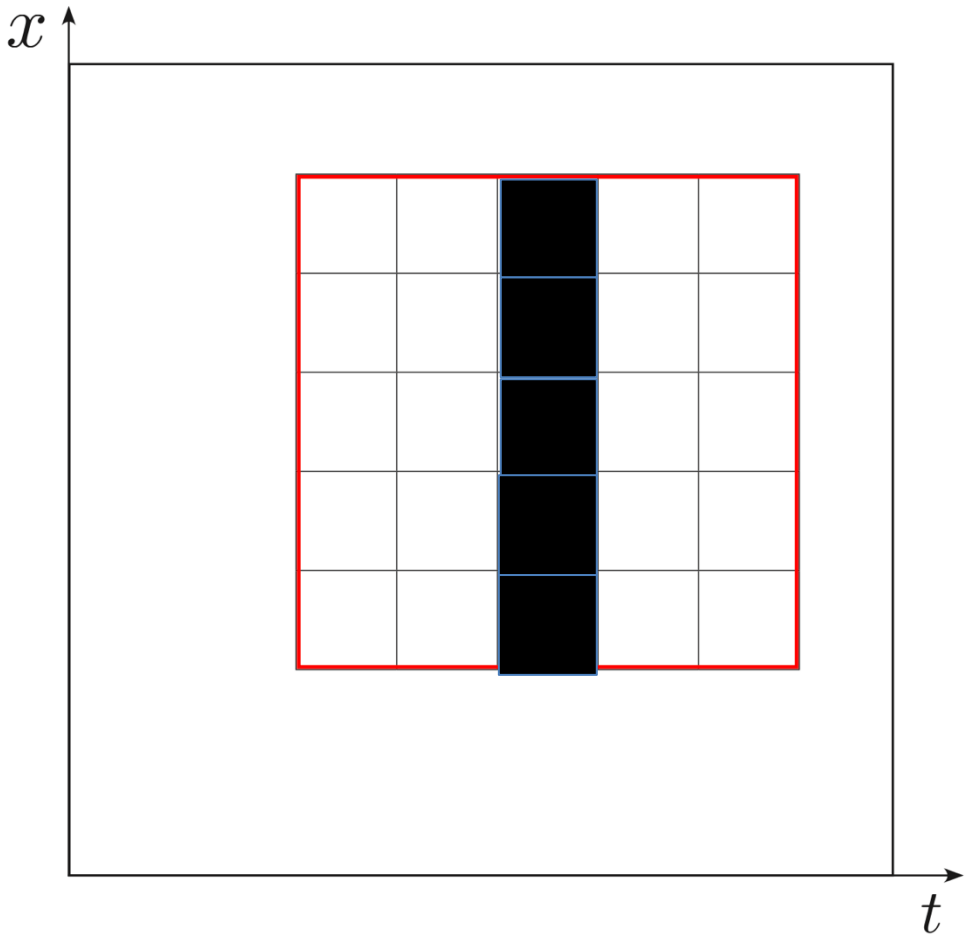


forecasting

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

Self-supervised learning tasks

Zero-shot capabilities

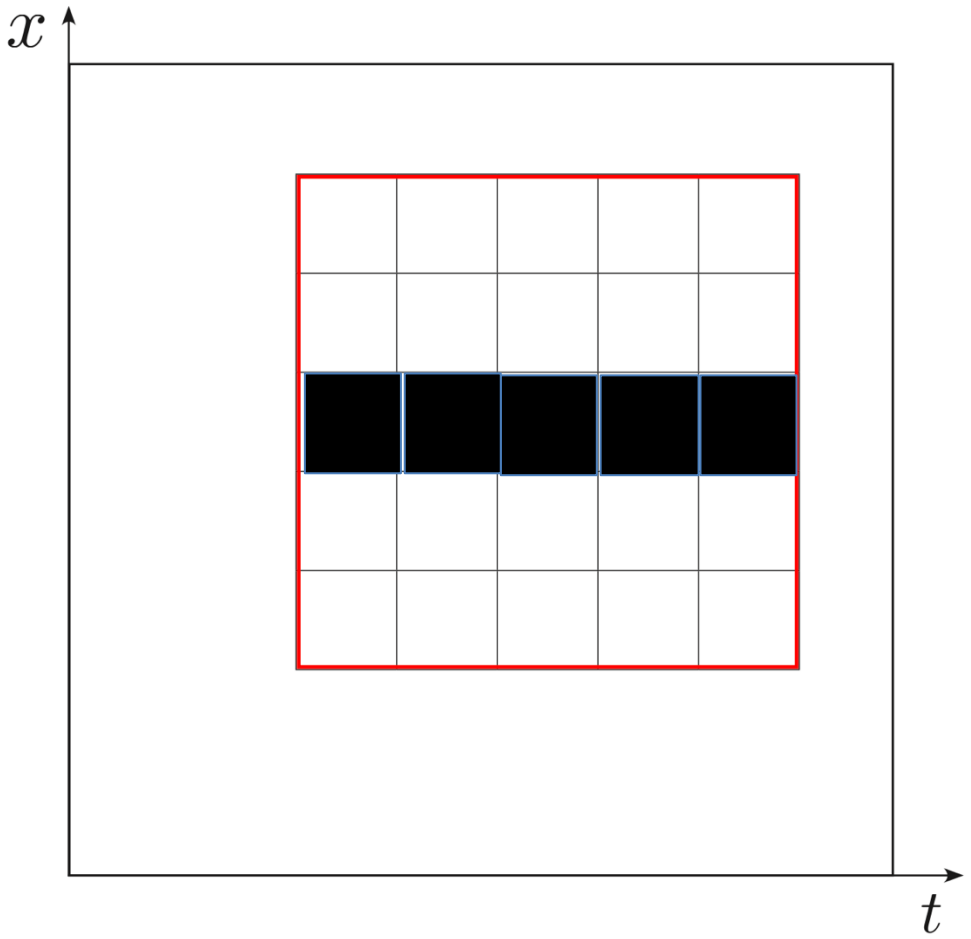


temporal interpolation

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

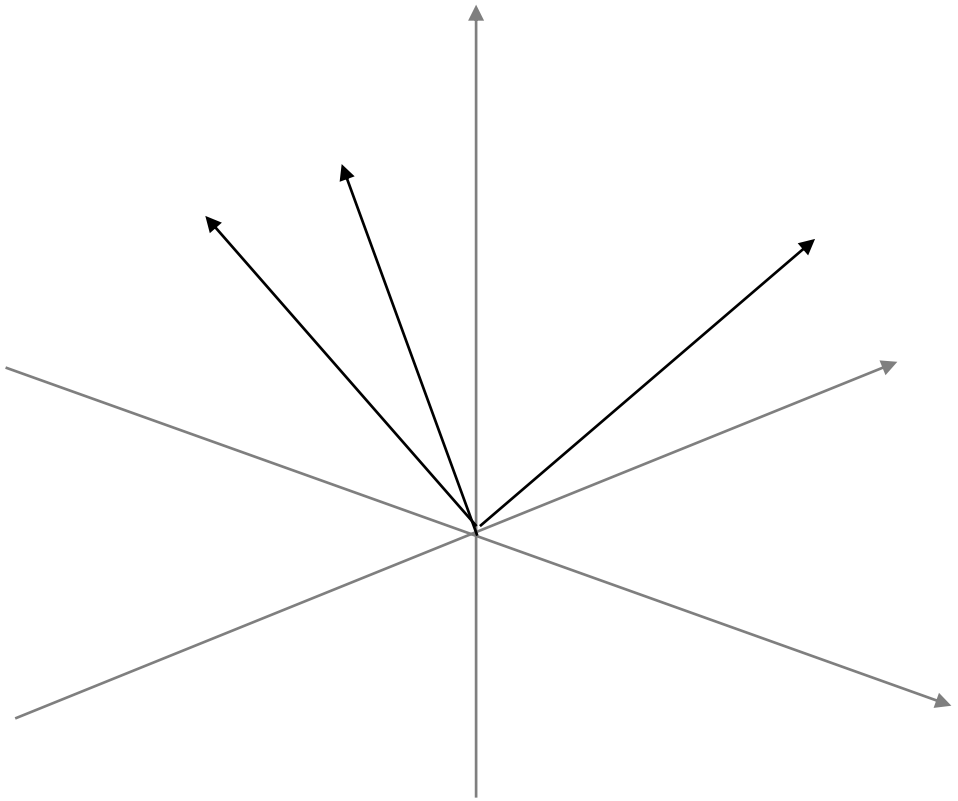
Lessig et al., AtmoRep: A Stochastic Model of Atmospheric Dynamics, 2023, https://arxiv.org/abs/2308.13280

Self-supervised learning tasks

- Contrastive learning: make sure latent representation is semantic

Self-supervised learning tasks

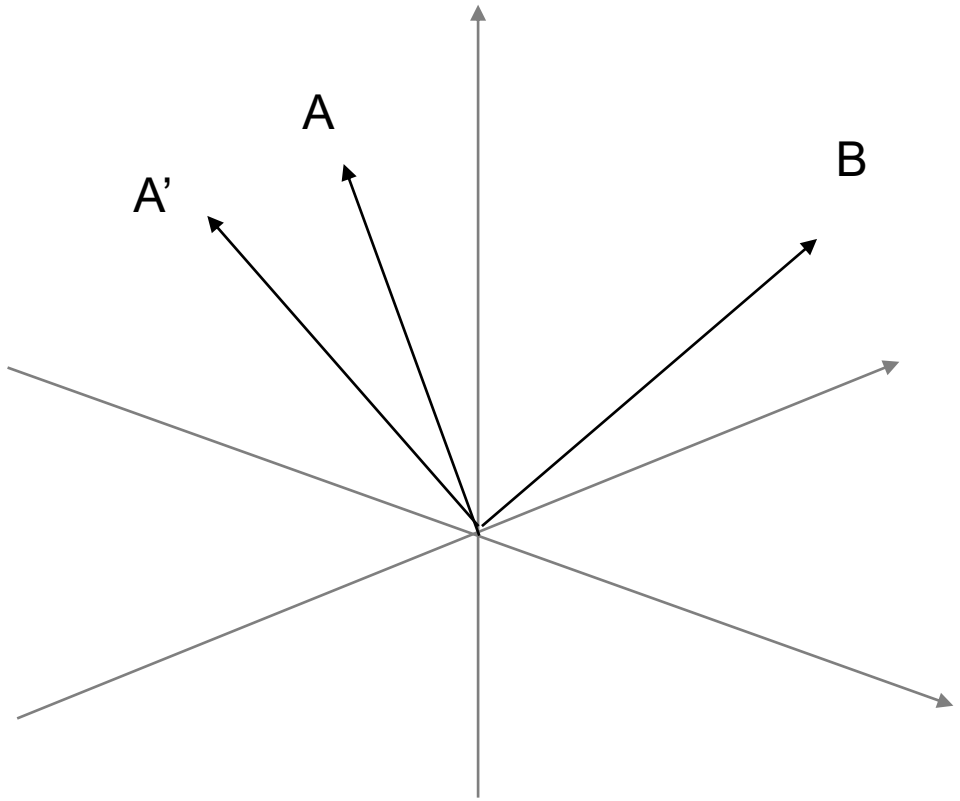
- Contrastive learning: make sure latent representation is semantic



From imagenet dataset

Self-supervised learning tasks

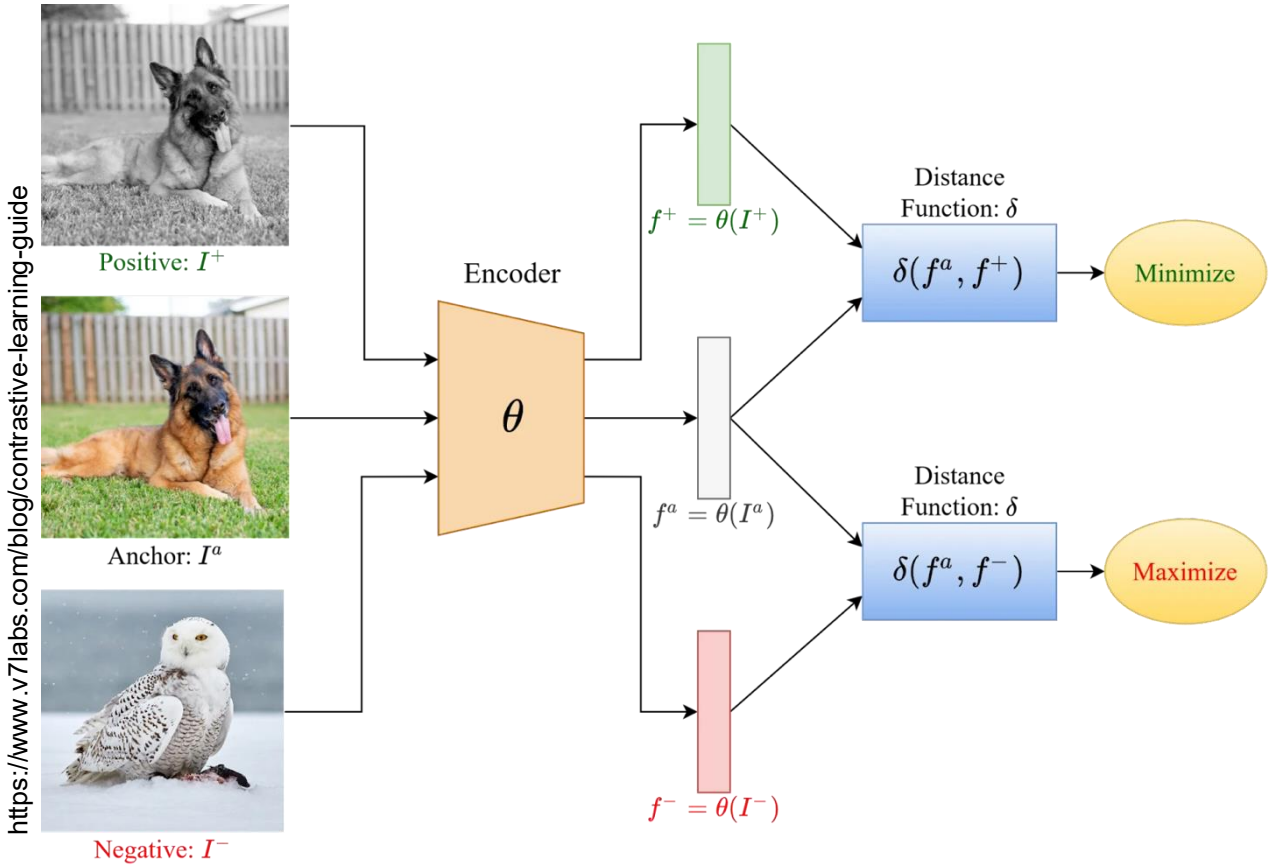
- Contrastive learning: make sure latent representation is semantic



From imagenet dataset

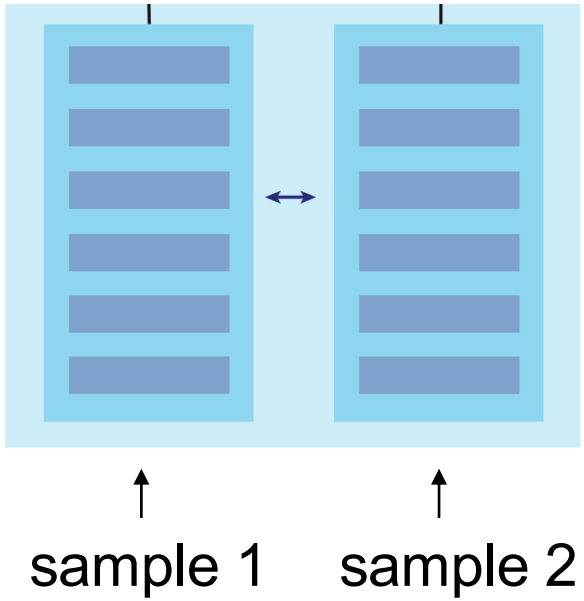
Self-supervised learning tasks

- Contrastive learning: make sure latent representation is semantic



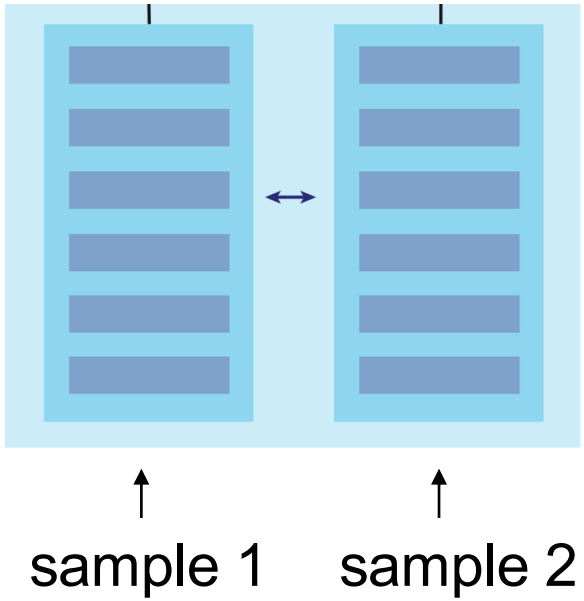
Self-supervised learning tasks

- Siamese networks



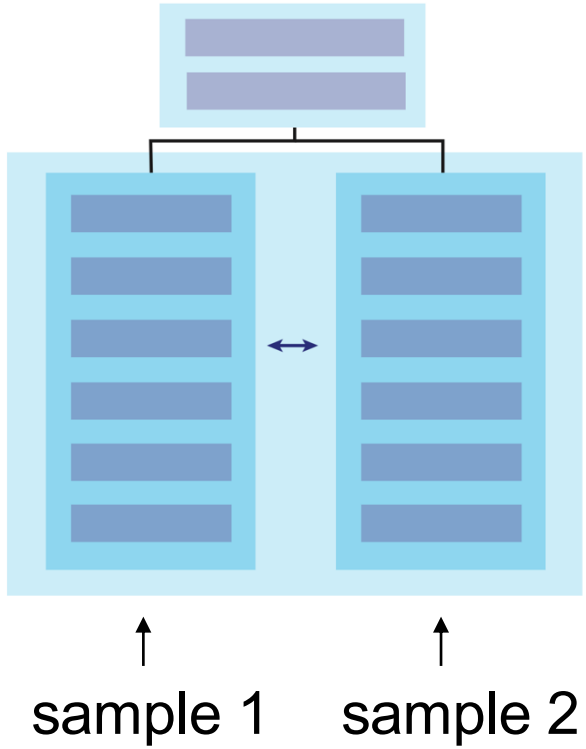
Self-supervised learning tasks

- Siamese networks



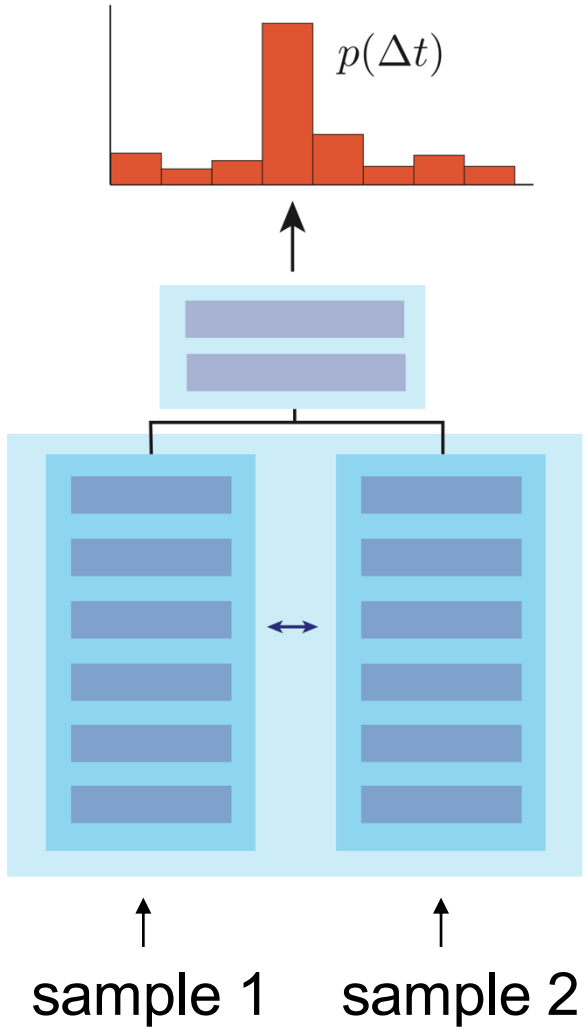
Self-supervised learning tasks

- Siamese networks



Self-supervised learning tasks

- Siamese networks



predict (known) similarity between samples

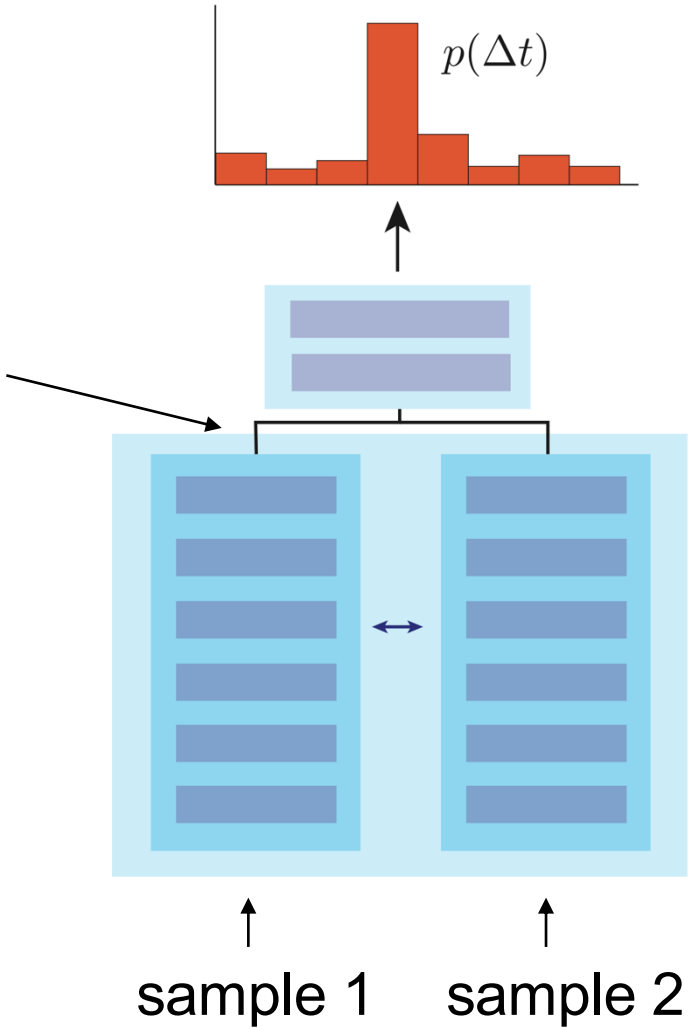
classification head

same network for both samples

Self-supervised learning tasks

- Siamese networks

Learn similar latent representation for similar samples

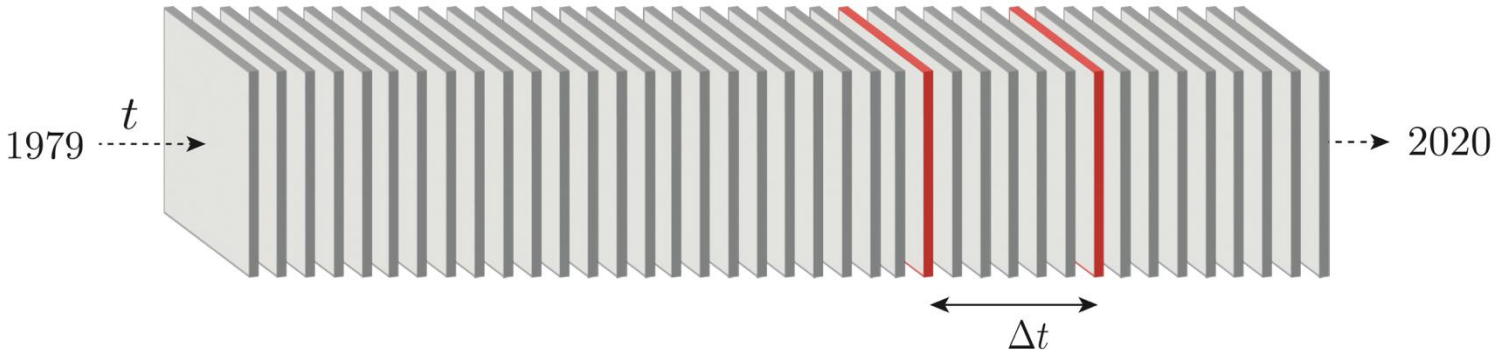


predict (known) similarity between samples

classification head

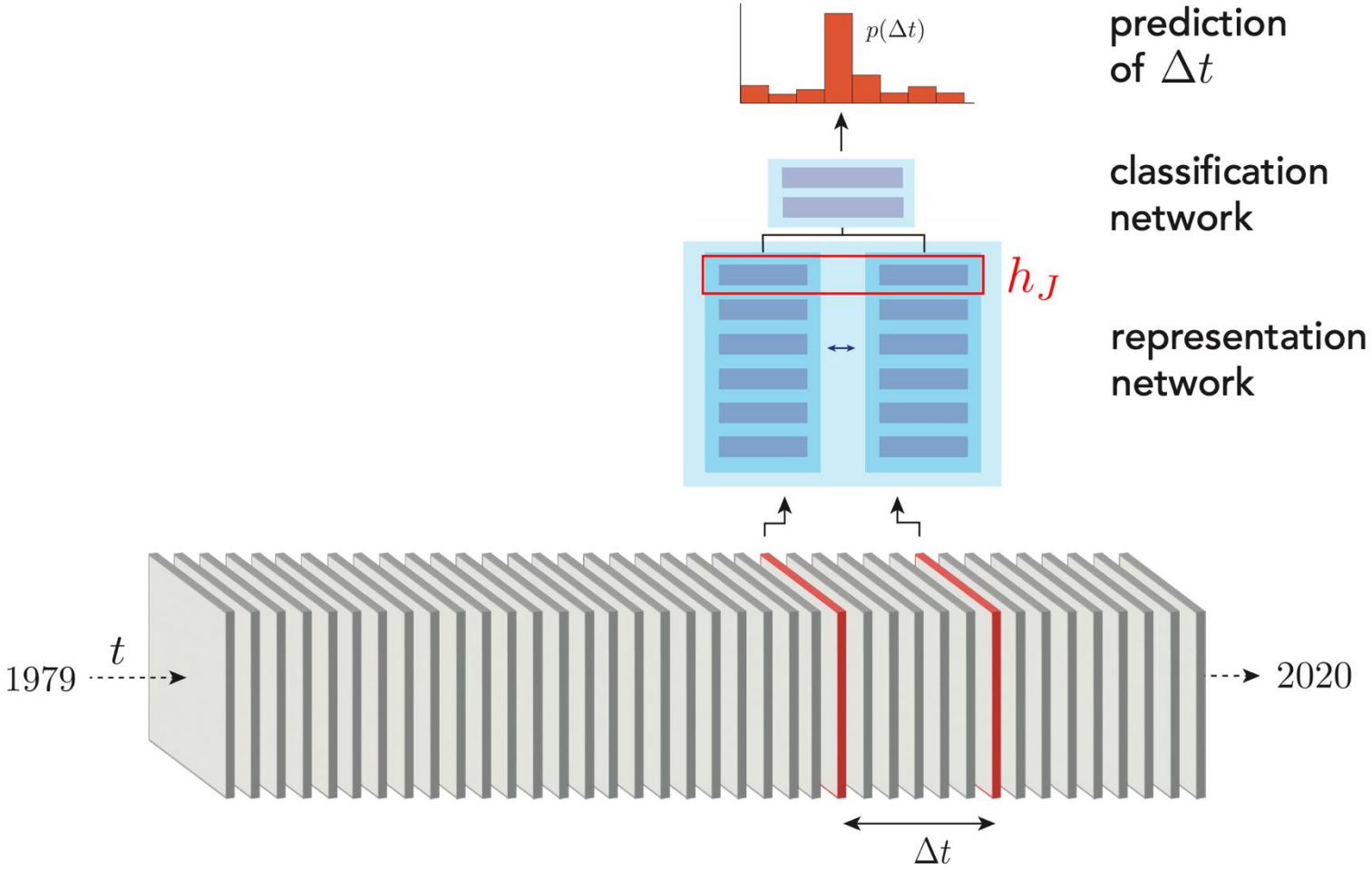
same network for both samples

Self-supervised learning tasks



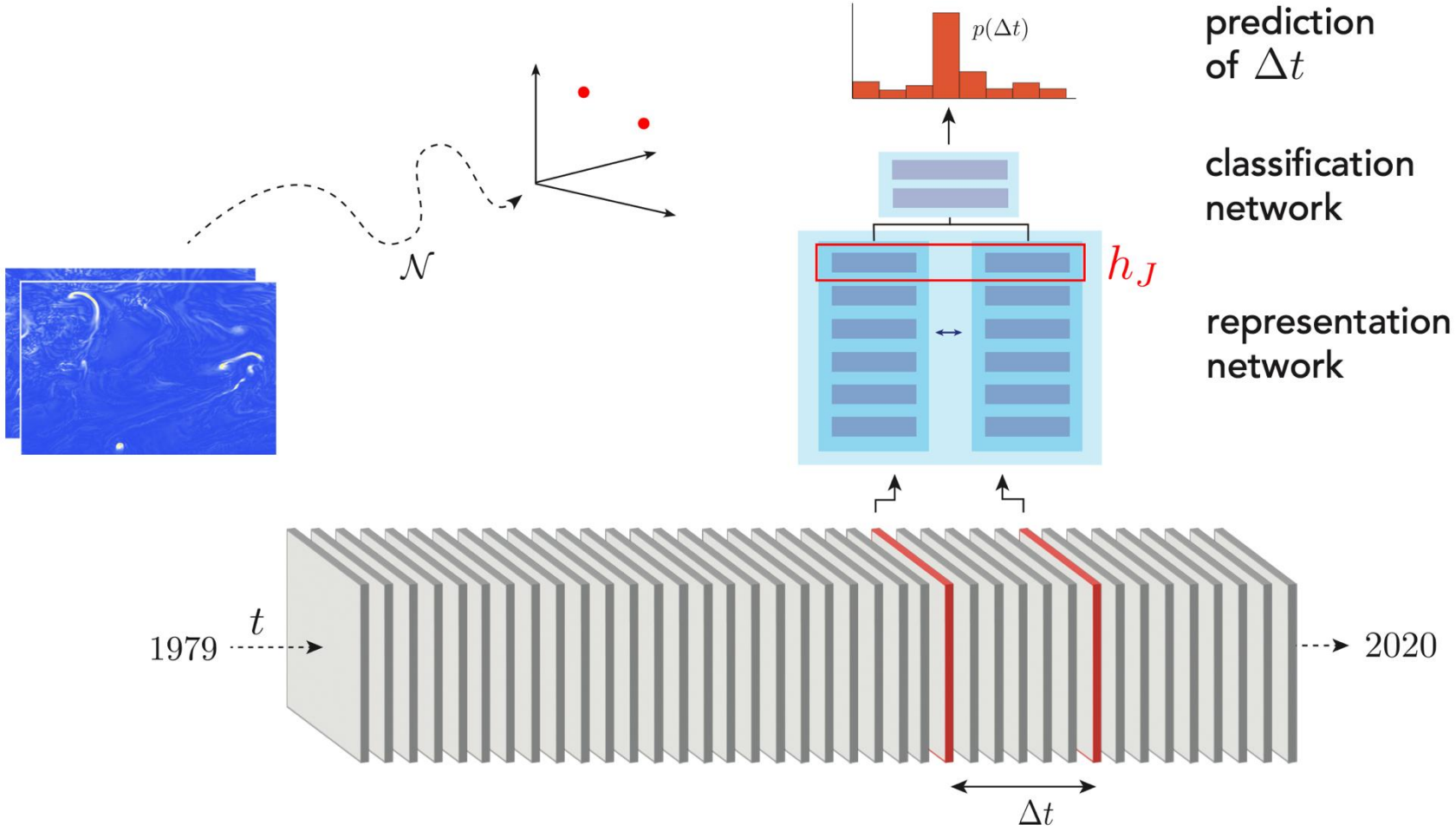
Hoffmann and Lessig, AtmoDist: Self-supervised Representation Learning for Atmospheric Dynamics, 2022, https://arxiv.org/abs/2202.01897

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

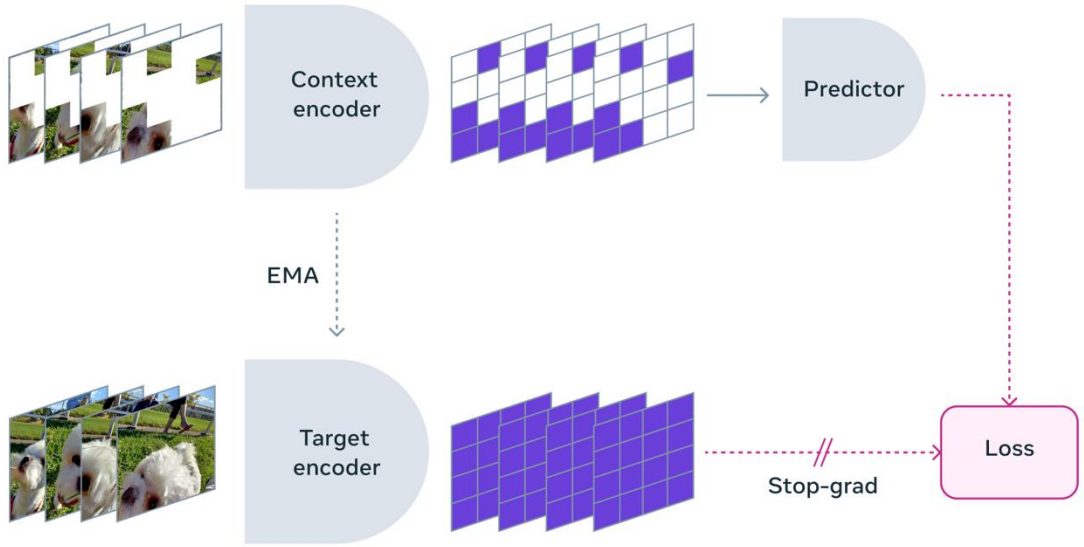
Self-supervised learning tasks

- 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

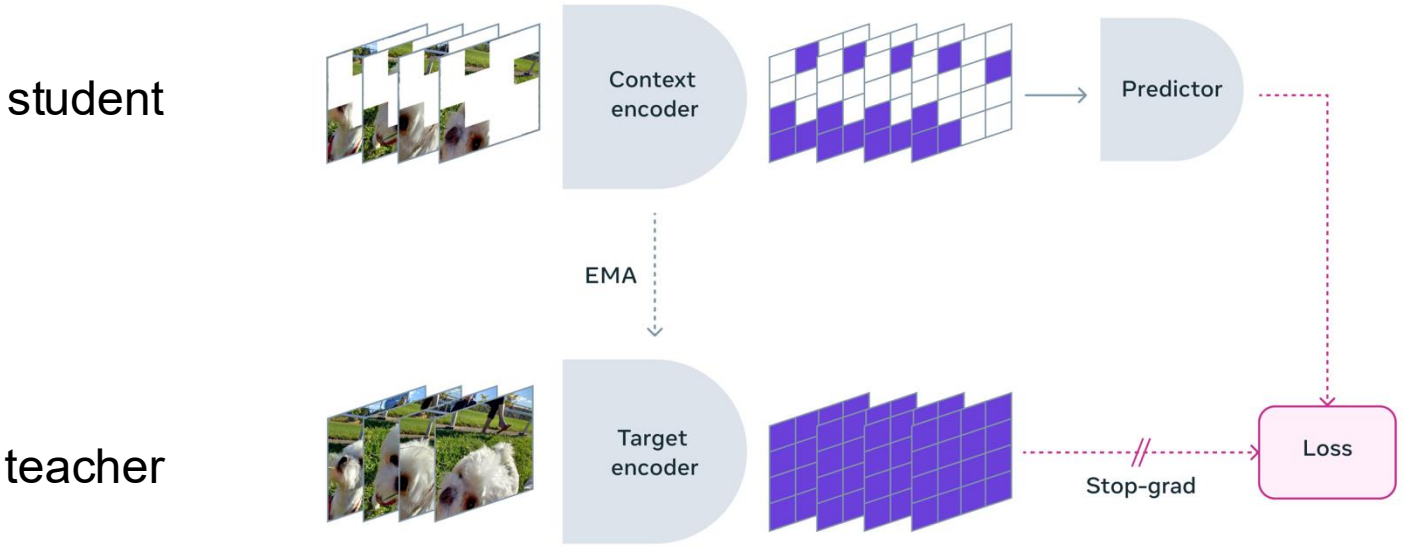
- Joined Embedding Predictive Architecture (JEPA)
  - Mask but compute loss in hidden/latent space instead of by reconstructing
  - Learn more abstract and robust representations



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

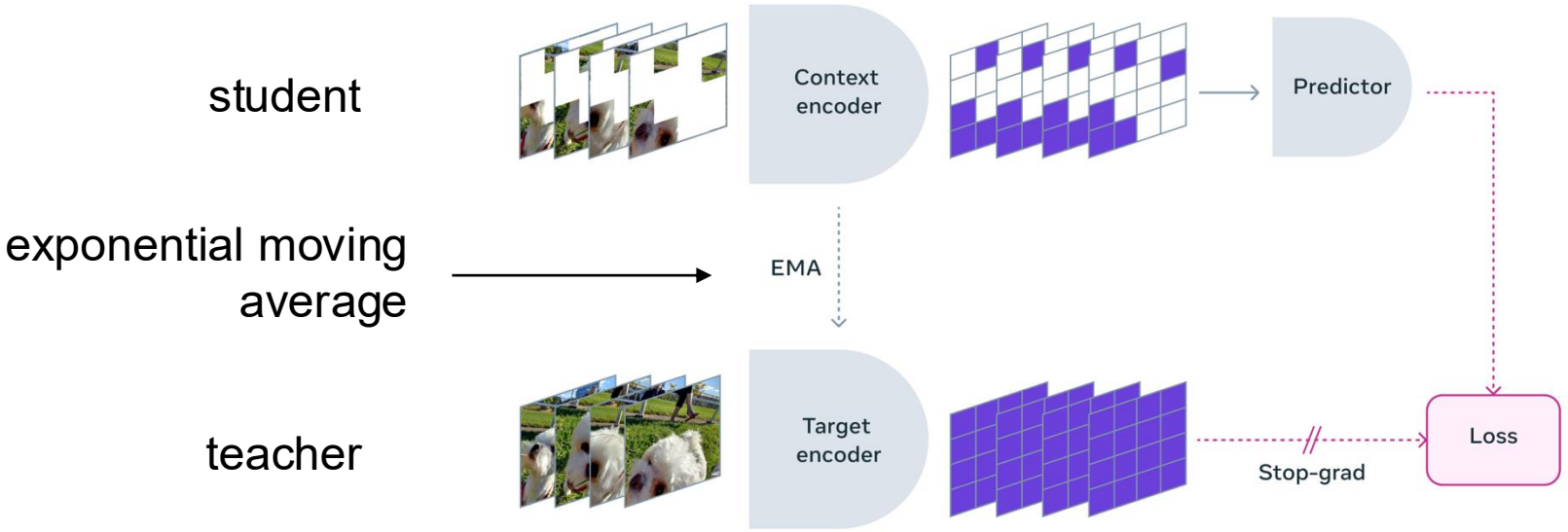
- Joined Embedding Predictive Architecture (JEPA)
  - Mask but compute loss in hidden/latent space instead of by reconstructing
  - Learn more abstract and robust representations



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

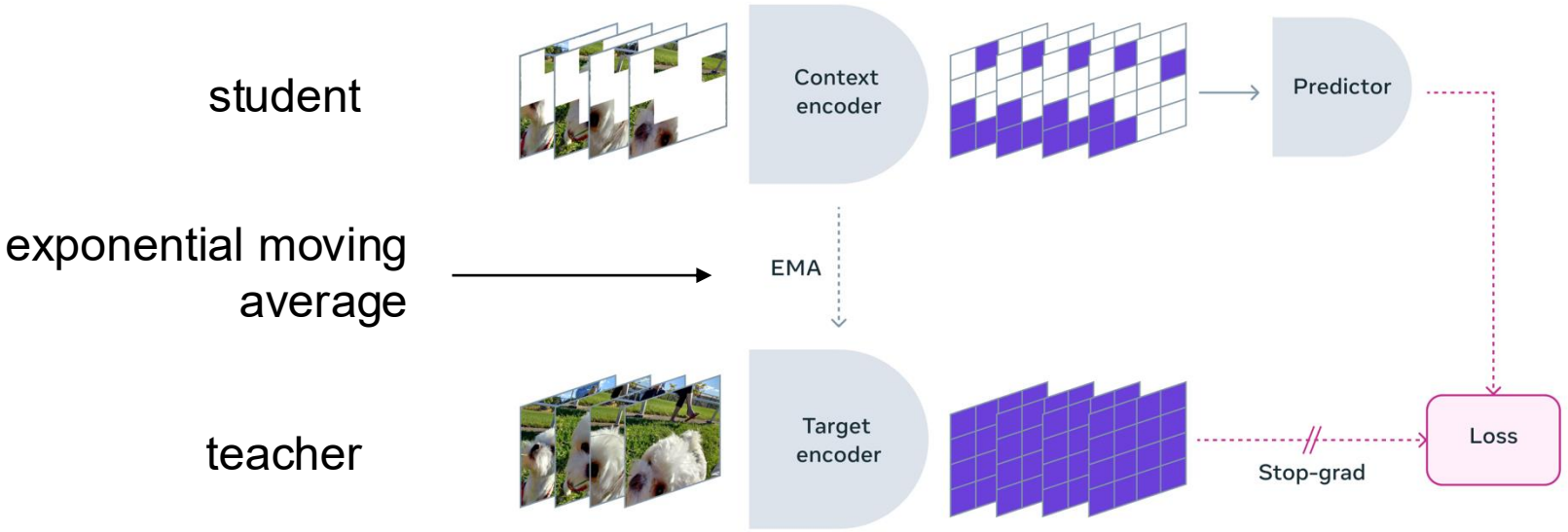
- Joined Embedding Predictive Architecture (JEPA)
  - Mask but compute loss in hidden/latent space instead of by reconstructing
  - Learn more abstract and robust representations



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

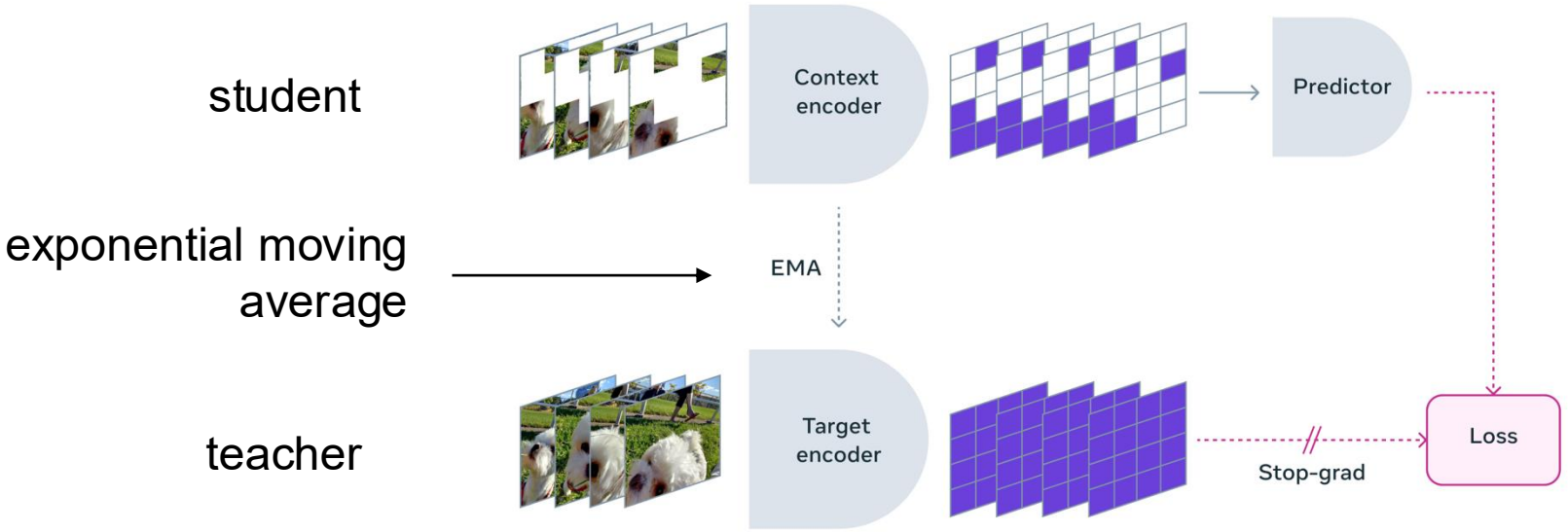


teacher is identical to student except that weights are averaged from last training steps

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_AfCFlyd8GMJnqQsG90WY-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

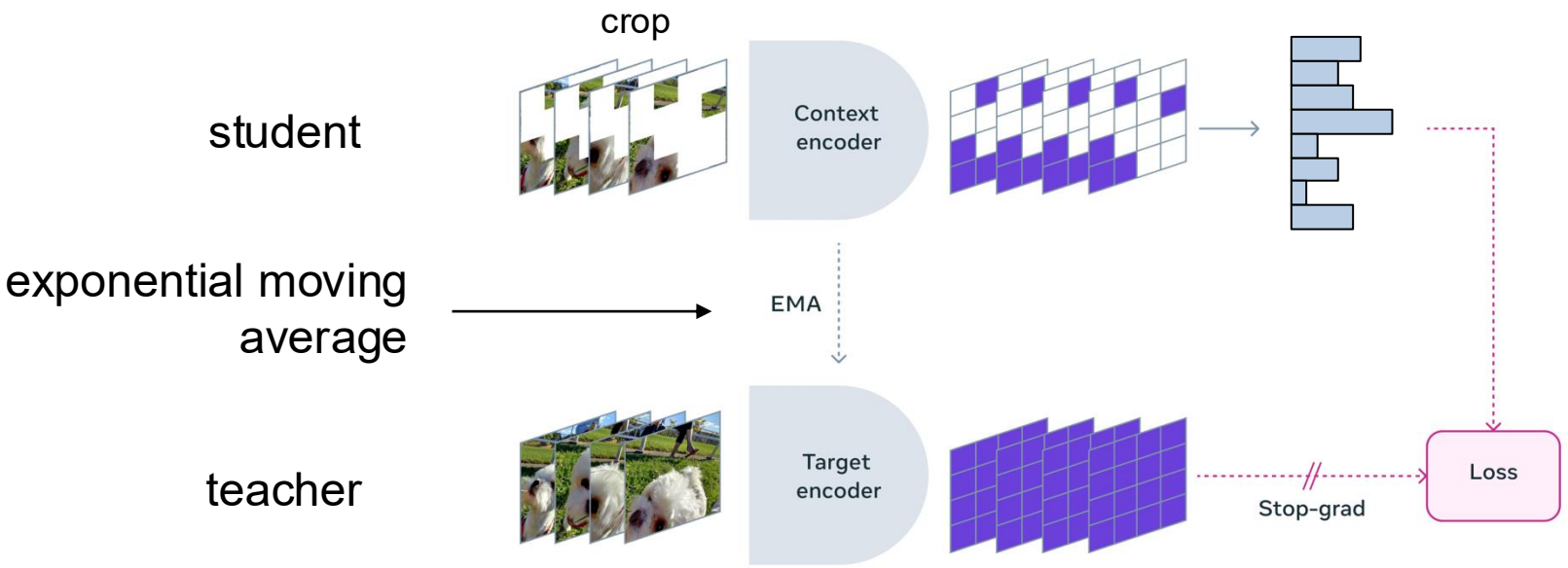


https://www.slideshare.net/slideshow/enact-carrot-stick/53622951

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

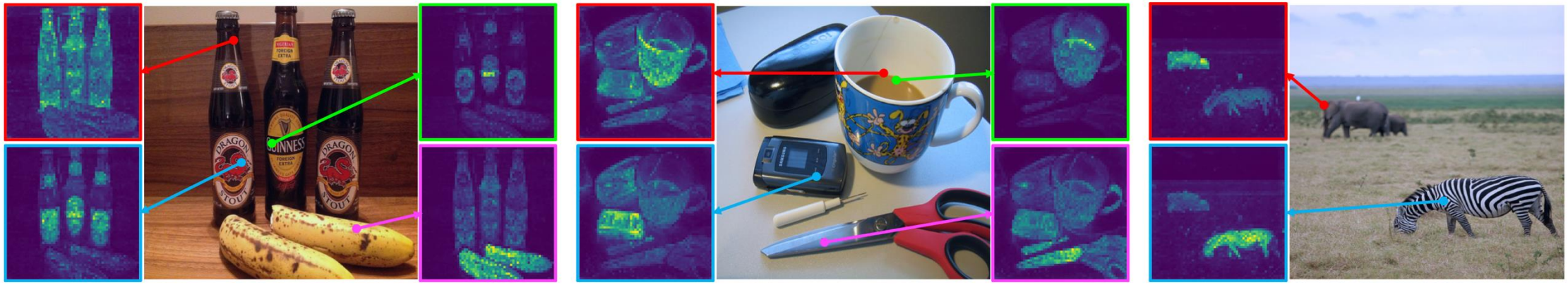
- DINO
  - Mask but compute loss in hidden/latent space instead of by reconstructing
  - Learn more abstract and robust representations



M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, Emerging properties in self-supervised vision transformers, CoRR, abs/2104.14294, 2021, arXiv: 2104.14294

Self-supervised learning tasks

- DINO



M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, Emerging properties in self-supervised vision transformers, CoRR, abs/2104.14294, 2021, arXiv: 2104.14294

Self-supervised learning tasks

- DINO

M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, Emerging properties in self-supervised vision transformers, CoRR, abs/2104.14294, 2021, arXiv: 2104.14294

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.