



NEW YORK UNIVERSITY

Differentiable Associative Memories, Attention, & Transformers

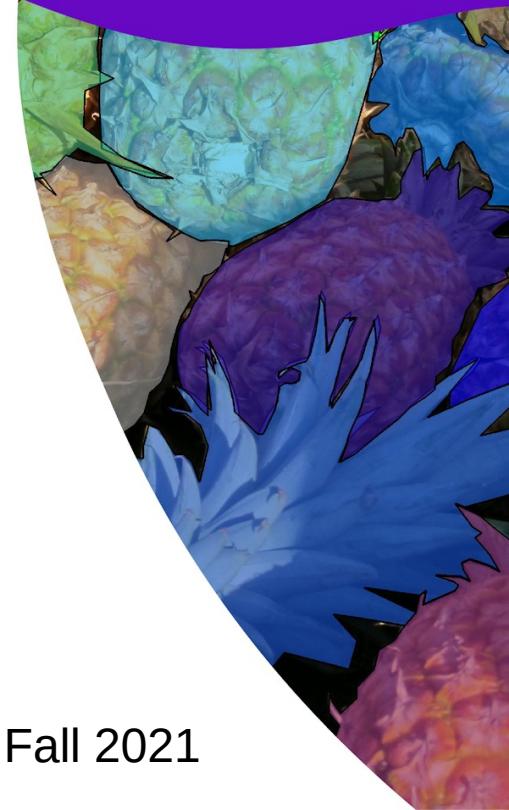
Yann LeCun

NYU - Courant Institute & Center for Data Science

Facebook AI Research

<http://yann.lecun.com>

Deep Learning, NYU, Fall 2021

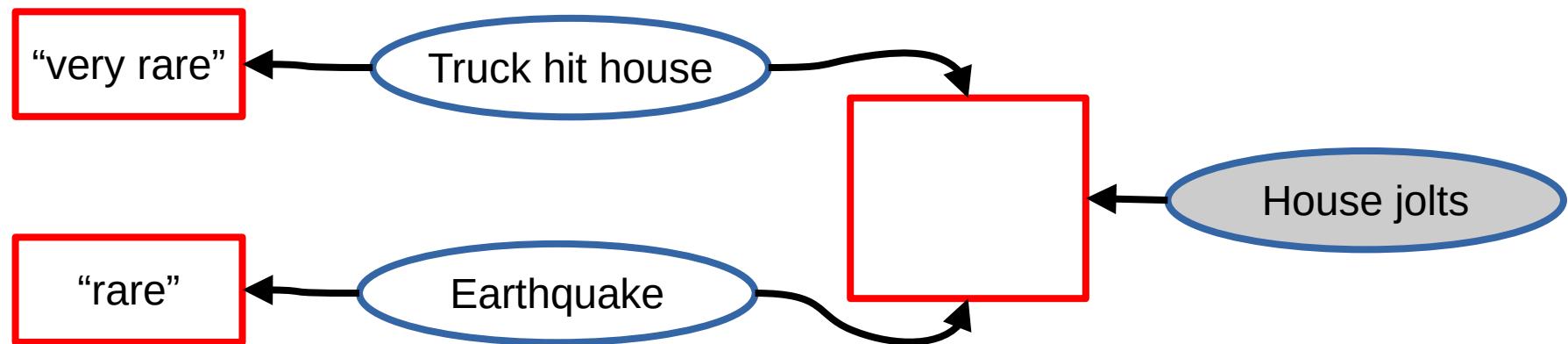


Getting DL systems to reason & use working memory

- ▶ **DL systems are good for:**
 - ▶ Perception: vision, audition...
 - ▶ Reactive action: acting reactively, without planning
 - ▶ Learning hierarchical representation of the perceptual world
- ▶ **How do we get DL systems to:**
 - ▶ Use a working memory?
 - ▶ Perform long chains of reasoning?
 - ▶ Remember massive amounts of factual knowledge?
 - ▶ Plan complex sequences of actions?
 - ▶ Learn hierarchical representations of action plans?

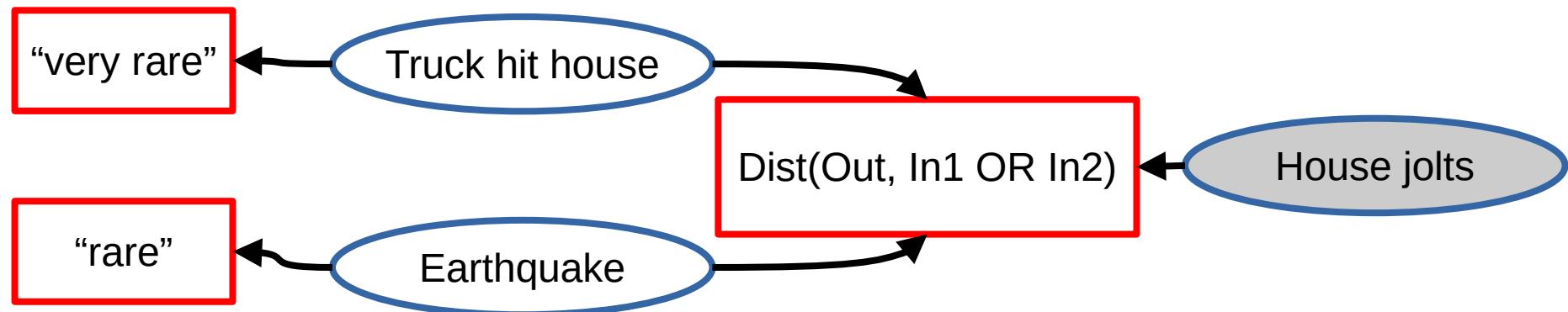
What is reasoning?

- ▶ Reasoning as constraint satisfaction / energy minimization
- ▶ Energy function represents constraints between observed and unobserved variables.
 - ▶ “house jolts” \leftarrow “truck hit house” OR “earthquake”
 - ▶ Minimizing the energy = finding consistent explanations
- ▶ Most forms of reasoning can be reduced to energy minimization



What is reasoning?

- ▶ **Reasoning as constraint satisfaction / energy minimization**
- ▶ Energy function represents constraints between observed and unobserved variables.
 - ▶ “house jolts” \leftarrow “truck hit house” OR “earthquake”
 - ▶ Minimizing the energy = finding consistent explanations
- ▶ **Most forms of reasoning can be reduced to energy minimization**
- ▶ e.g. inference in graphical models / factor graphs

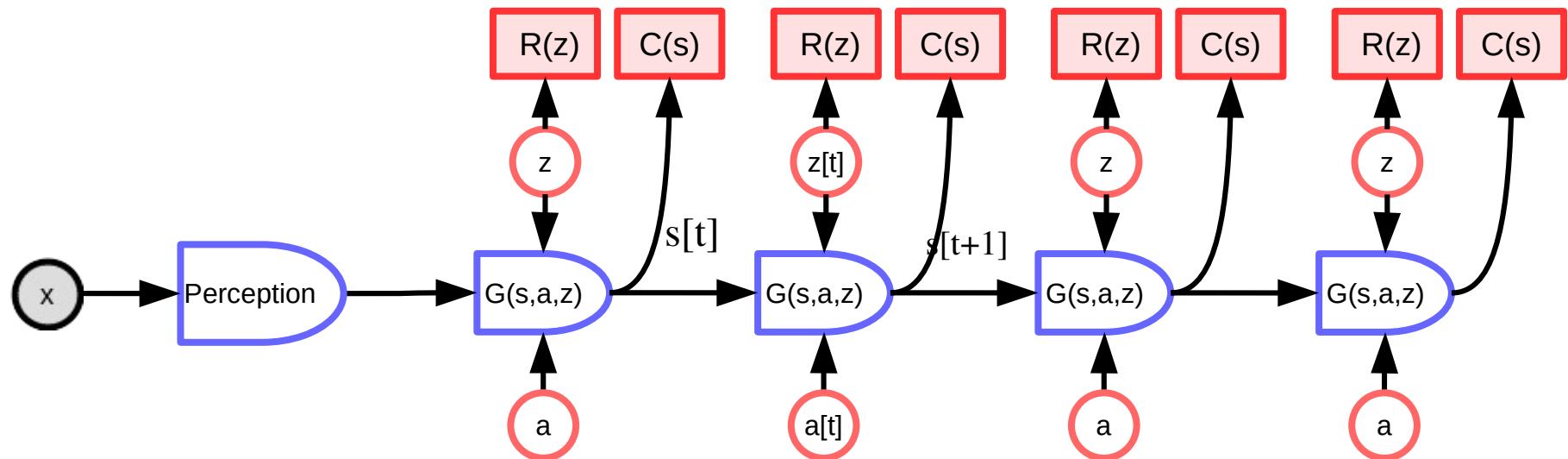


What is reasoning?

- ▶ Planning through energy minimization

- ▶ Optimal control

- ▶ Model Predictive Control:



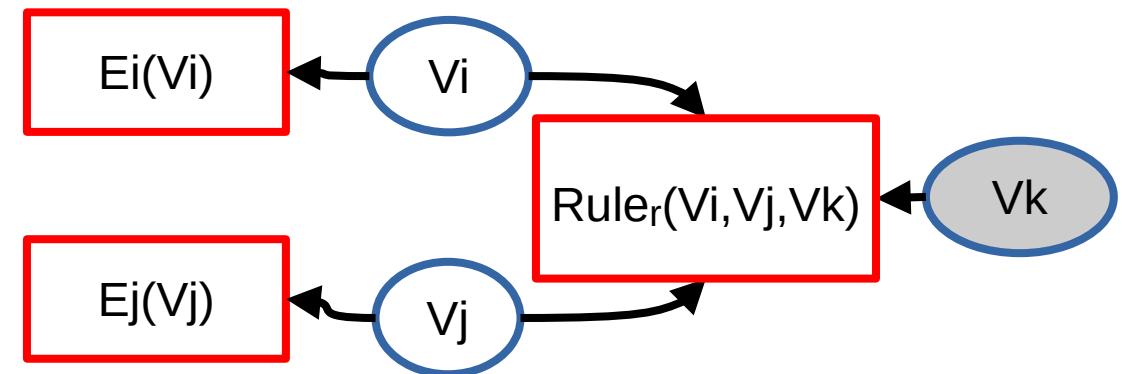
Reasoning with a working memory

- ▶ Long chains of reasoning require a short-term memory of facts deduced by applying rules (energy terms) to sets of variables (some of them observed, other inferred)
- ▶ Knowledge base of variables (facts) and energy terms (rules)

Variable values

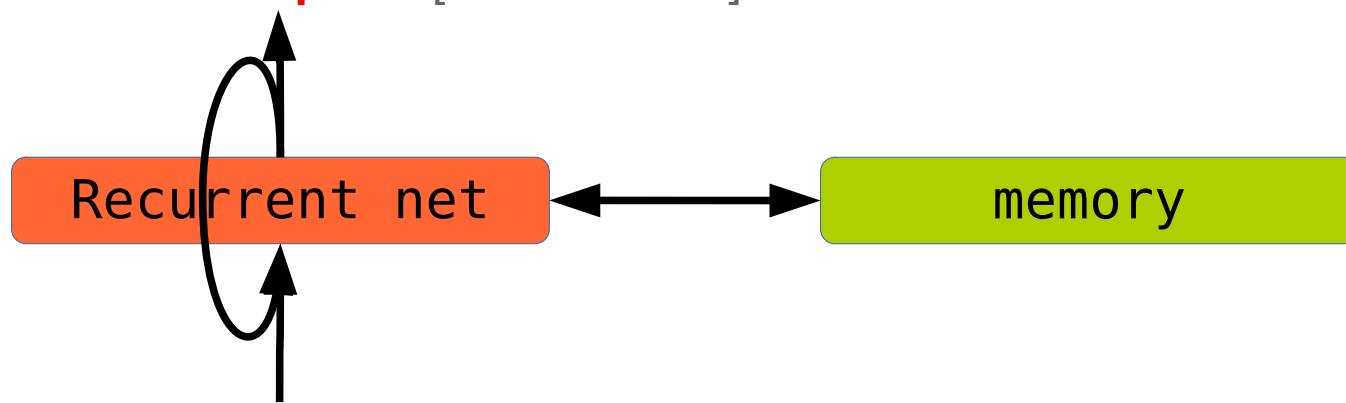


Rules



Memory-Augmented Networks

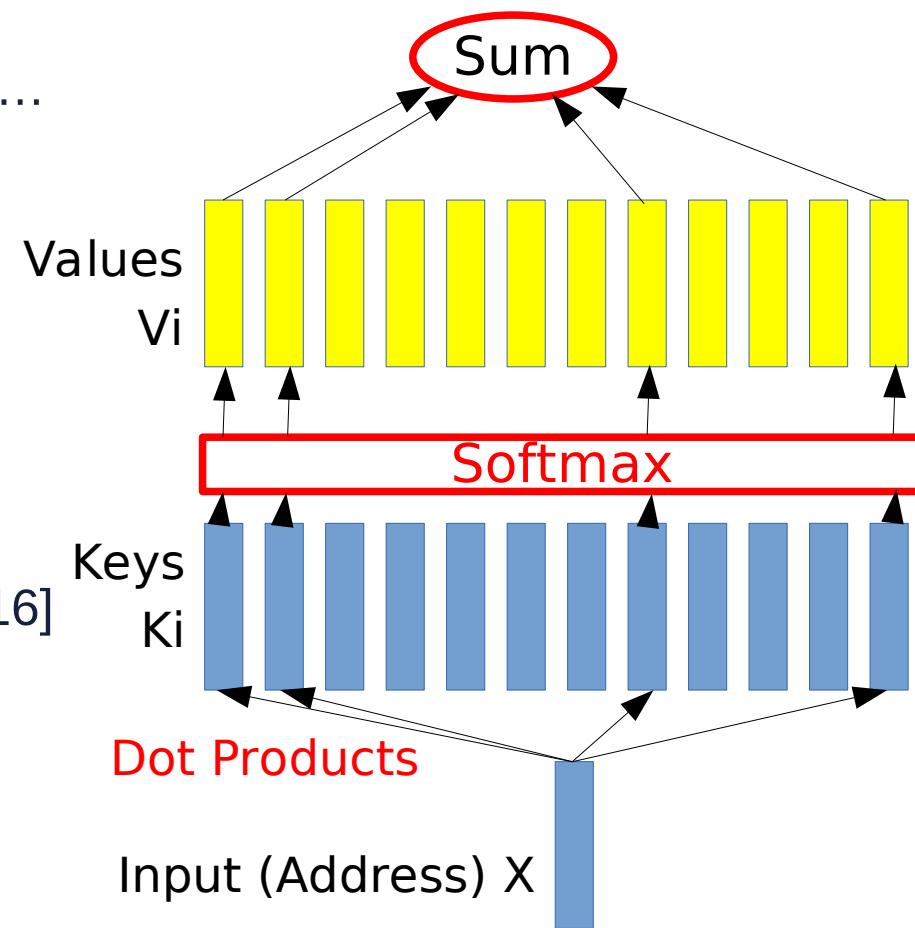
- Recurrent networks cannot remember things for very long
 - ▶ The cortex only remember things for 20 seconds
- We need a “hippocampus” (a separate memory module)
 - ▶ LSTM [Hochreiter 1997], registers
 - ▶ **Memory networks** [Weston et 2014] (FAIR), associative memory
 - ▶ **Stacked-Augmented Recurrent Neural Net** [Joulin & Mikolov 2014] (FAIR)
 - ▶ **Neural Turing Machine** [Graves 2014],
 - ▶ **Differentiable Neural Computer** [Graves 2016]



Differentiable Associative Memory == “soft RAM”

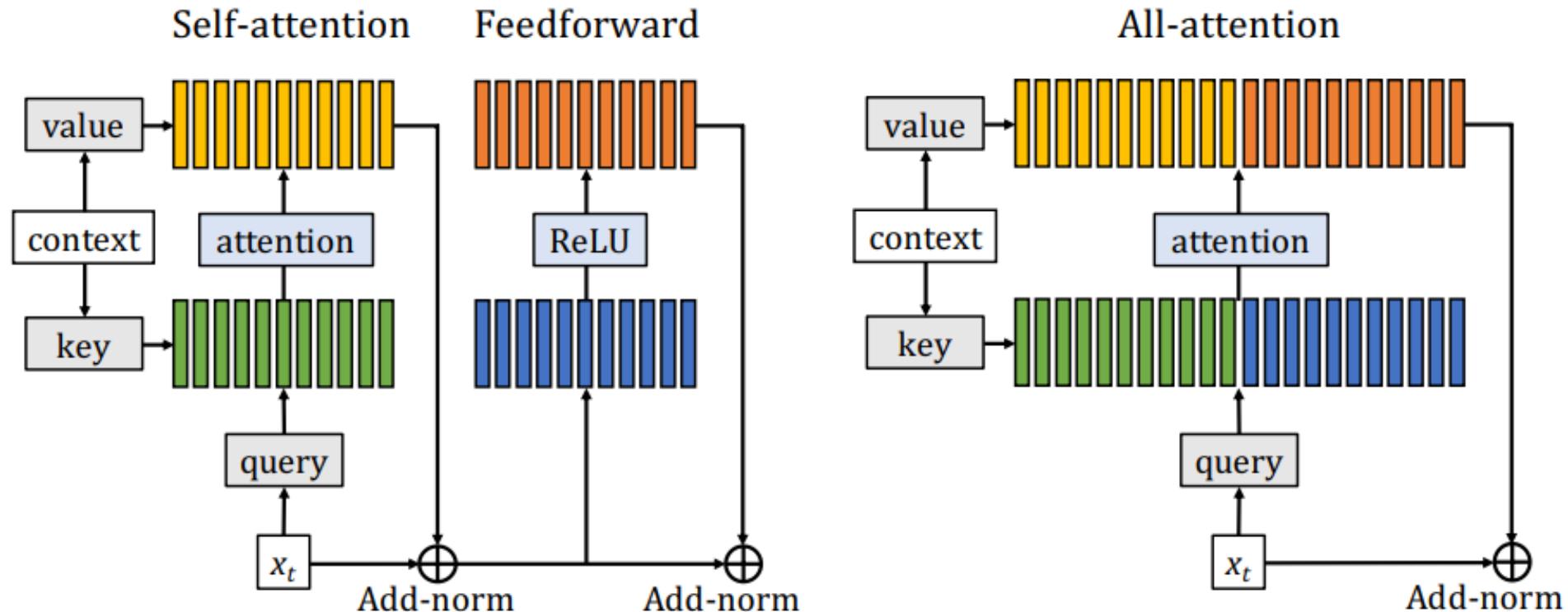
- ▶ Memory Networks, Transformer Network,
- ▶ ELMO, GPT, BERT, GPT2, RoBERTa, XLM-R...
- ▶ Used very widely in NLP
- ▶ Essentially a “soft” RAM or hash table
- ▶ **Memory networks** [Weston et 2014] (FAIR)
- ▶ **Stacked-Augmented Recurrent Neural Net**
[Joulin & Mikolov 2014] (FAIR)
- ▶ **Neural Turing Machine** [Graves 2014],
- ▶ **Differentiable Neural Computer** [Graves 2016]

$$Y = \sum_i C_i V_i \quad C_i = \frac{e^{K_i^T X}}{\sum_j e^{K_j^T X}}$$



All-Attention Circuit with persistent memory

► [Sukhbaatar arXiv:1907.01470]

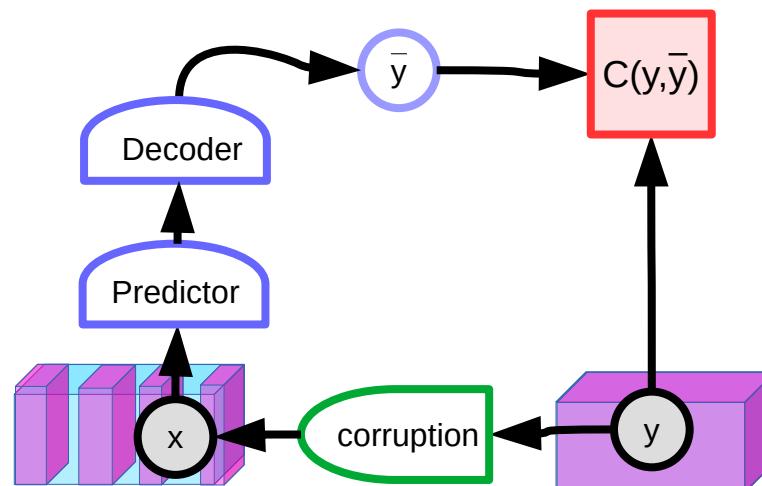


Contrastive Methods vs Regularized/Architectural Methods

- ▶ **Contrastive:** [they all are different ways to pick which points to push up]
 - ▶ C1: push down of the energy of data points, push up everywhere else: Max likelihood (needs tractable partition function or variational approximation)
 - ▶ C2: push down of the energy of data points, push up on chosen locations: max likelihood with MC/MMC/HMC, Contrastive divergence, Metric learning/Siamese nets, Ratio Matching, Noise Contrastive Estimation, Min Probability Flow, adversarial generator/GANs
 - ▶ C3: train a function that maps points off the data manifold to points on the data manifold: denoising auto-encoder, masked auto-encoder (e.g. BERT)
- ▶ **Regularized/Architectural:** [Different ways to limit the information capacity of the latent representation]
 - ▶ A1: build the machine so that the volume of low energy space is bounded: PCA, K-means, Gaussian Mixture Model, Square ICA, normalizing flows...
 - ▶ A2: use a regularization term that measures the volume of space that has low energy: Sparse coding, sparse auto-encoder, LISTA, Variational Auto-Encoders, discretization/VQ/VQVAE.
 - ▶ A3: $F(x,y) = C(y, G(x,y))$, make $G(x,y)$ as "constant" as possible with respect to y : Contracting auto-encoder, saturating auto-encoder
 - ▶ A4: minimize the gradient and maximize the curvature around data points: score matching

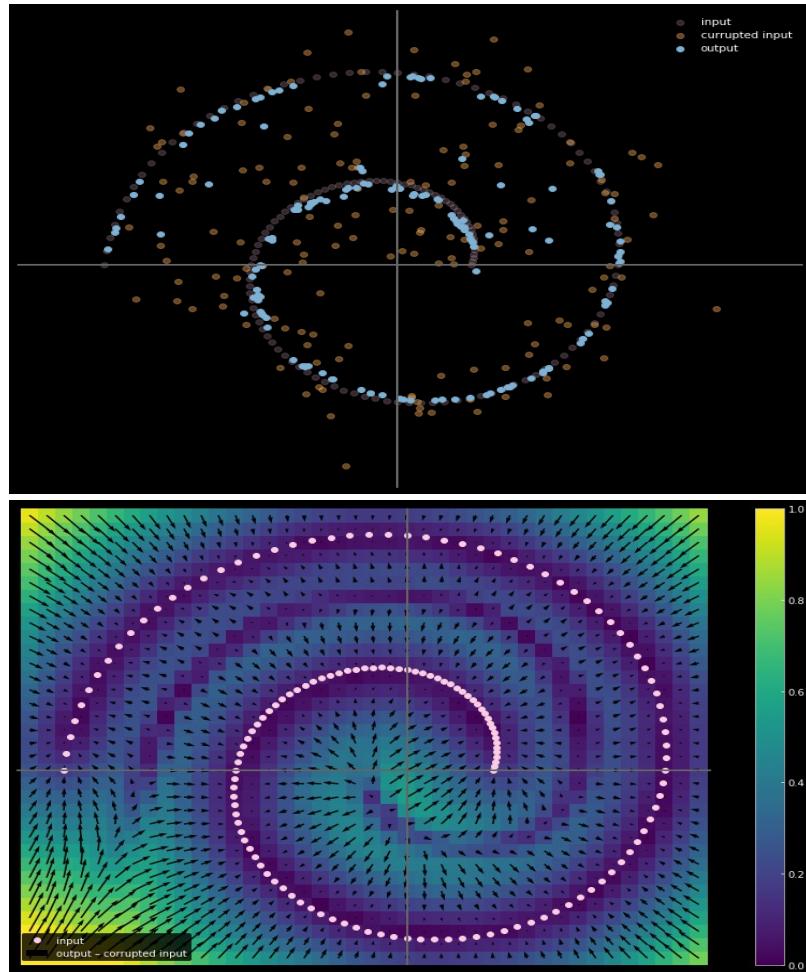
Contrastive Method: Denoising AE / Masked AE

- ▶ Denoising AE [Vincent 2008] /
- ▶ Masked AE: Learning text representations
- ▶ [Collobert-Weston 2011],
- ▶ BERT [Devlin 2018], RoBERTa [Ott 2019]



This is a [...] of text extracted
[...] a large set of [...] articles

This is a piece of text extracted
from a large set of news articles

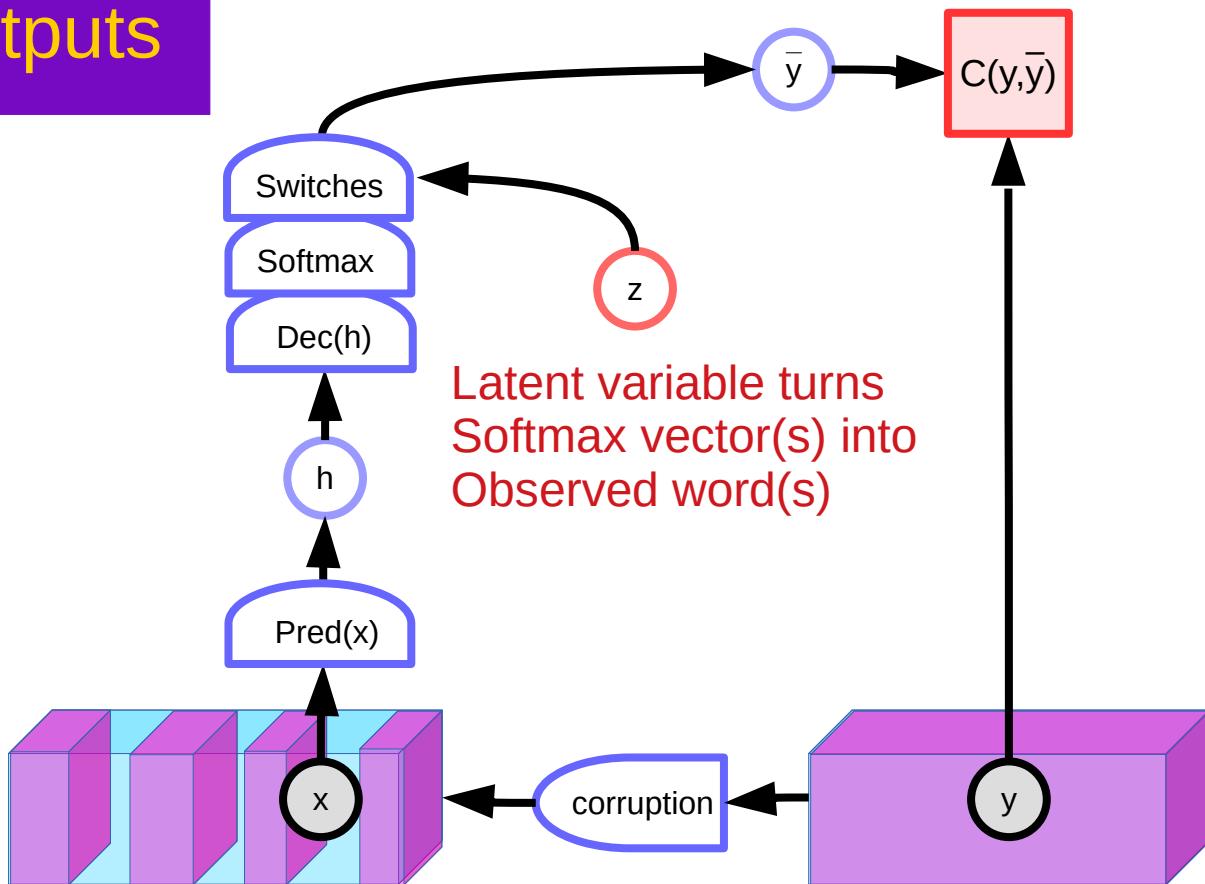


Figures: Alfredo Canziani

Denoising AE: discrete outputs

[Vincent et al. JMLR 2008]

- ▶ Masked Auto-Encoder
- ▶ BERT [Waswani 2018]
- ▶ RoBERTa [Ott 2019]
- ▶ Issues:
 - ▶ latent variables are in output space
 - ▶ No abstract LV to control the output
 - ▶ How to cover the space of corruptions?



This is a [...] of text extracted
[...] a large set of [...] articles

This is a piece of text extracted
from a large set of news articles

Transformer Architecture

- ▶ Multi-head attention (associative memory)
- ▶ [Waswani ArXiv:1706.03762]
- ▶ 10 to 60 stages
- ▶ BERT model
- ▶ [Devlin ArXiv:1810.04805]
- ▶ Trained to fill in missing words

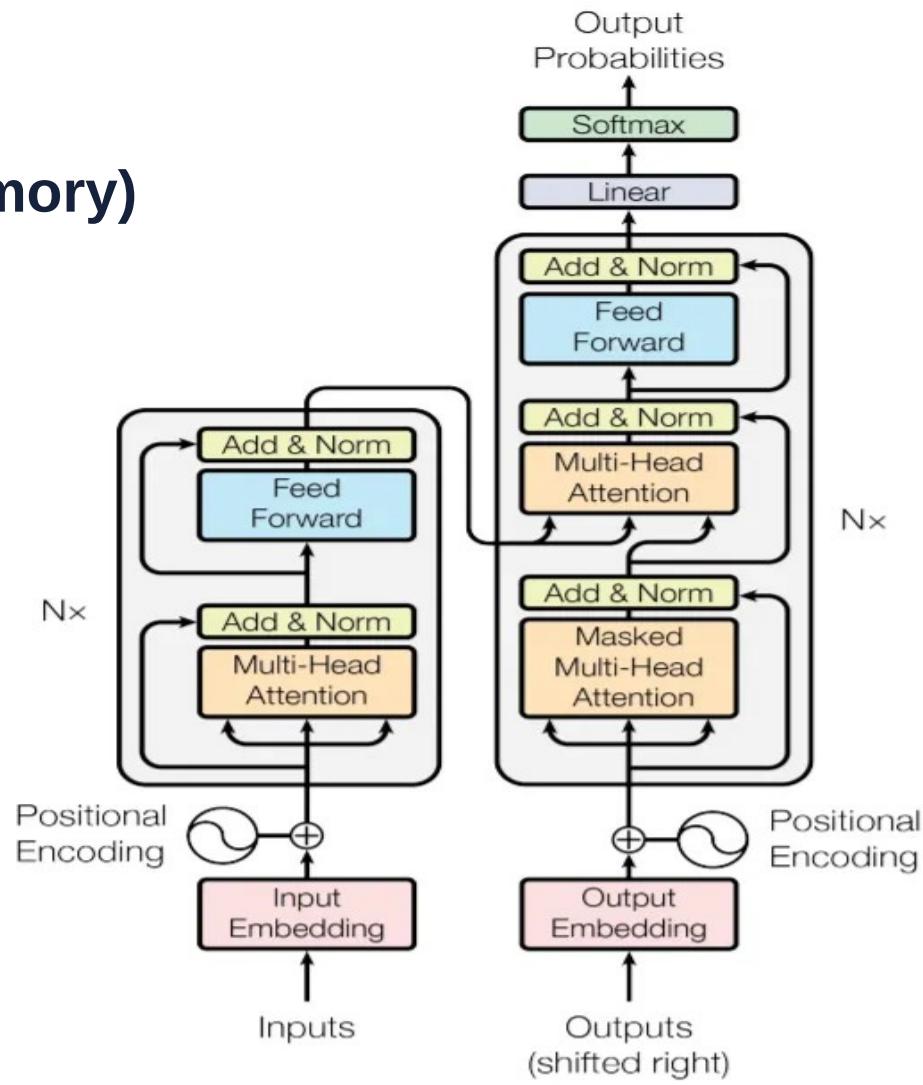
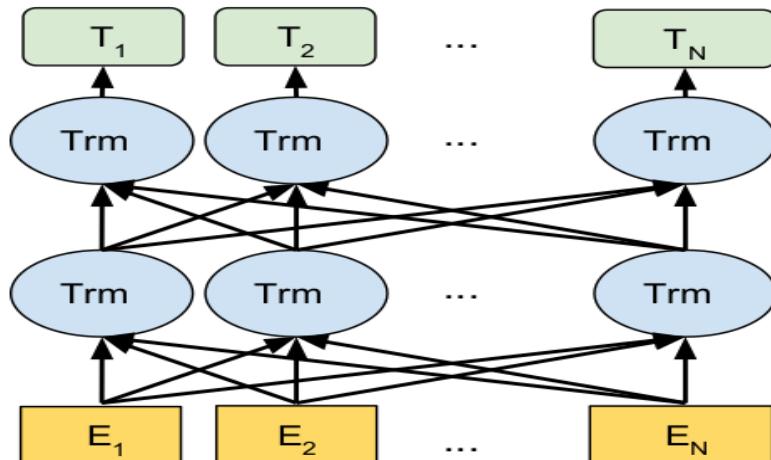
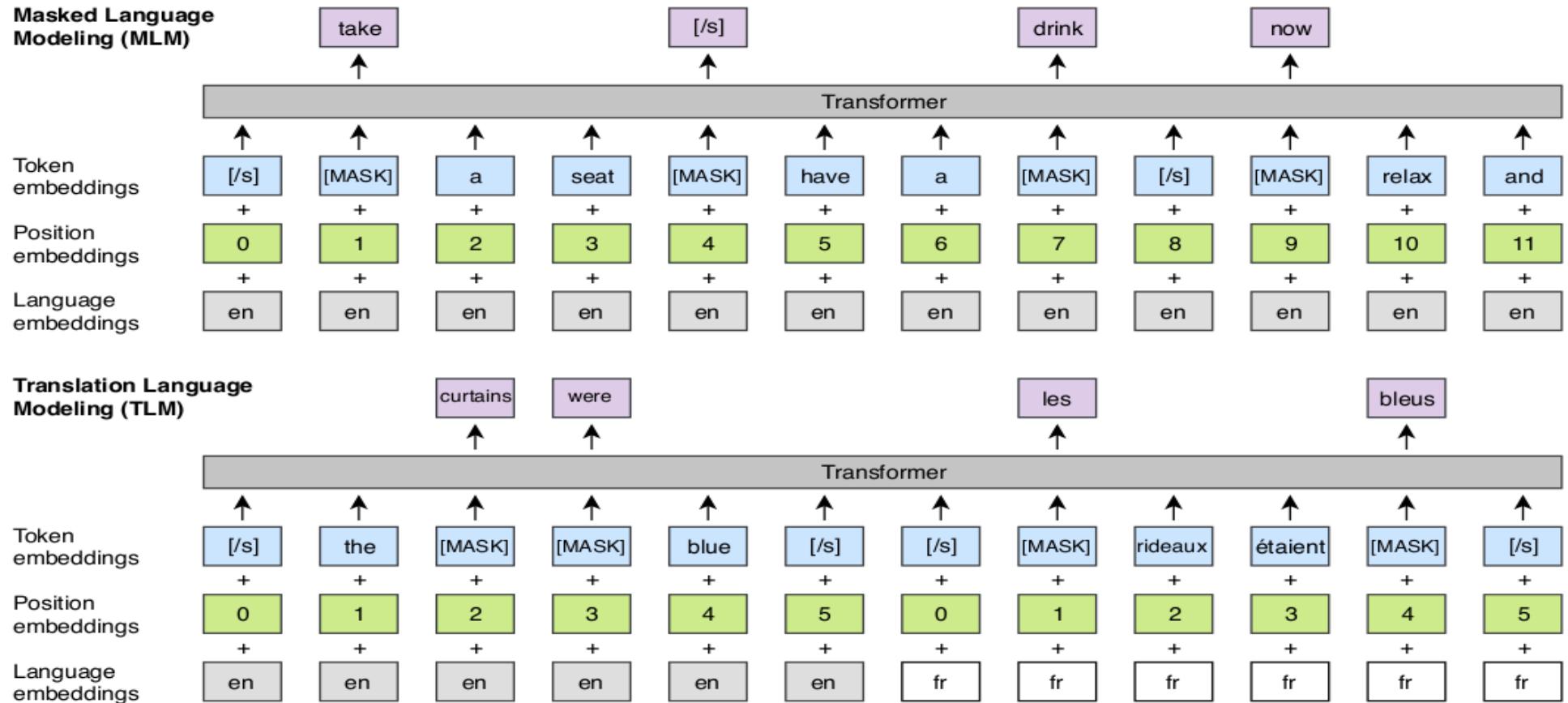


Figure 1: The Transformer - model architecture.

Multilingual Transformer Architecture XLM-R

► [Lample & Conneau ArXiv:1901.07291]



Supervised Symbol Manipulation

- ▶ Solving integrals and differential equations symbolically with a transformer architecture
- ▶ [Lample & Charton
arXiv:1912.01412]
- ▶ Accuracy on various problems →

	Integration (BWD)	ODE (order 1)	ODE (order 2)
Mathematica (30s)	84.0	77.2	61.6
Matlab	65.2	-	-
Maple	67.4	-	-
Beam size 1	98.4	81.2	40.8
Beam size 10	99.6	94.0	73.2
Beam size 50	99.6	97.0	81.0

Equation	Solution
$y' = \frac{16x^3 - 42x^2 + 2x}{(-16x^8 + 112x^7 - 204x^6 + 28x^5 - x^4 + 1)^{1/2}}$	$y = \sin^{-1}(4x^4 - 14x^3 + x^2)$
$3xy \cos(x) - \sqrt{9x^2 \sin(x)^2 + 1}y' + 3y \sin(x) = 0$	$y = c \exp(\sinh^{-1}(3x \sin(x)))$
$4x^4yy'' - 8x^4y'^2 - 8x^3yy' - 3x^3y'' - 8x^2y^2 - 6x^2y' - 3x^2y'' - 9xy' - 3y = 0$	$y = \frac{c_1 + 3x + 3 \log(x)}{x(c_2 + 4x)}$

Natural language understanding & generation [MMBlenderbot]



I would love to take this teddy bear for a spin on my motorcycle.

Ha, would the teddy bear ride in a basket in front of you or in a sidecar?

In front of me, of course. I don't want to get hit by a car.

Yes, good point - the sidecar would take up a lot of room

I think it would be fun to ride on the back of a motorbike with a stuffed animal in the basket.

Do you ride your motorbike often?

I used to ride a lot when I was younger, but I haven't been on one in a long time.



These flowers are so beautiful, I would love to have a garden like this one day.

Yeah, me too! What kinds of flowers would you plant?

I'm not sure, I'd have to do a lot of research to find out what would grow well in my area.

That's a good point - honestly, I'm not sure what would grow well here. Do you have a big backyard for gardening?

I live in the suburbs, so I don't have a huge yard, but I do have a small garden in my front yard.

Oh, that's great. Would you grow vegetables as well or just flowers?

Vegetables, herbs, flowers, fruits and nuts are all things I grow in my garden.

That would be nice! I'd love to have apple trees and then eat fresh apples every fall.

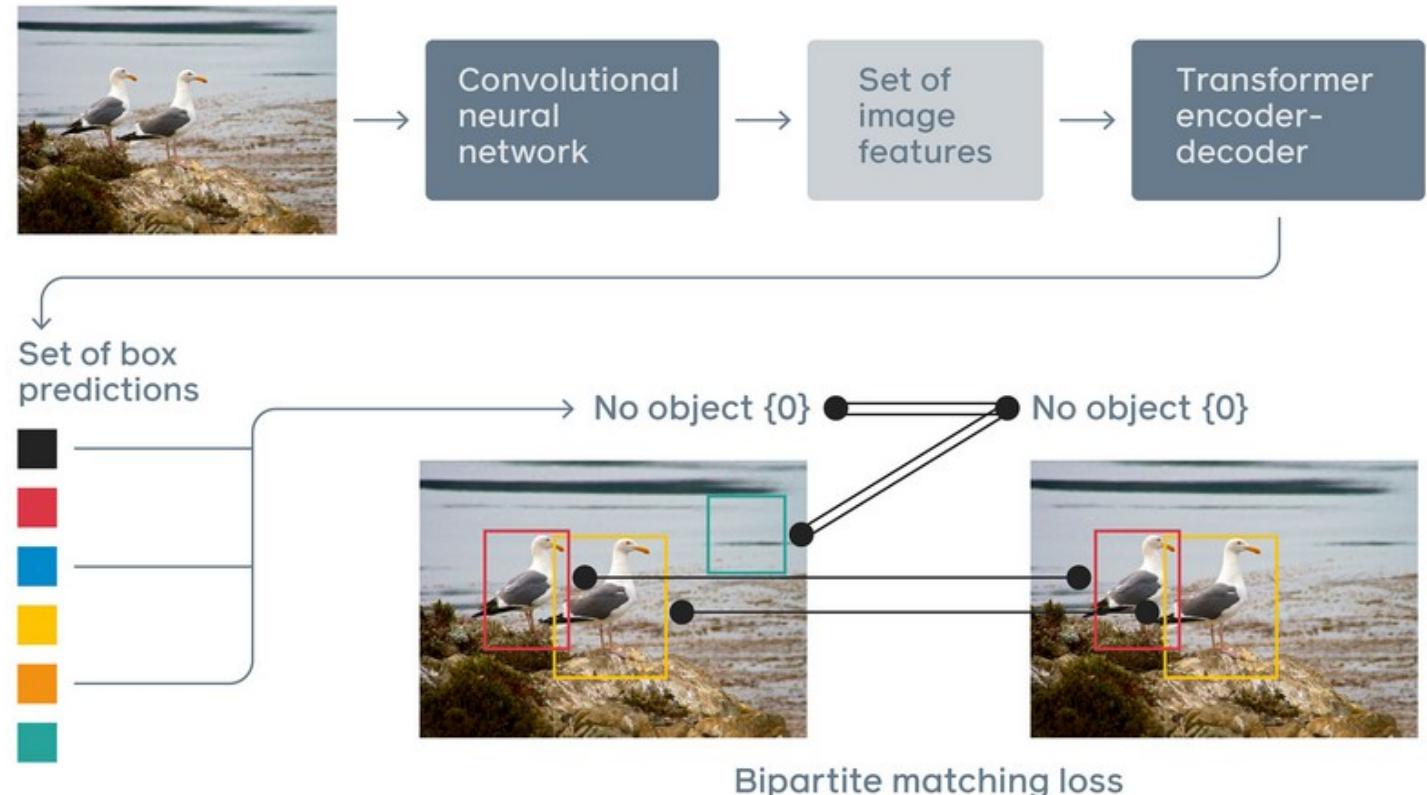
I would grow all of the above. I love fresh fruit and veggies.

Do you like coconuts? People either love or hate them, I think.

I've never had a coconut. I've heard they're good for you though.

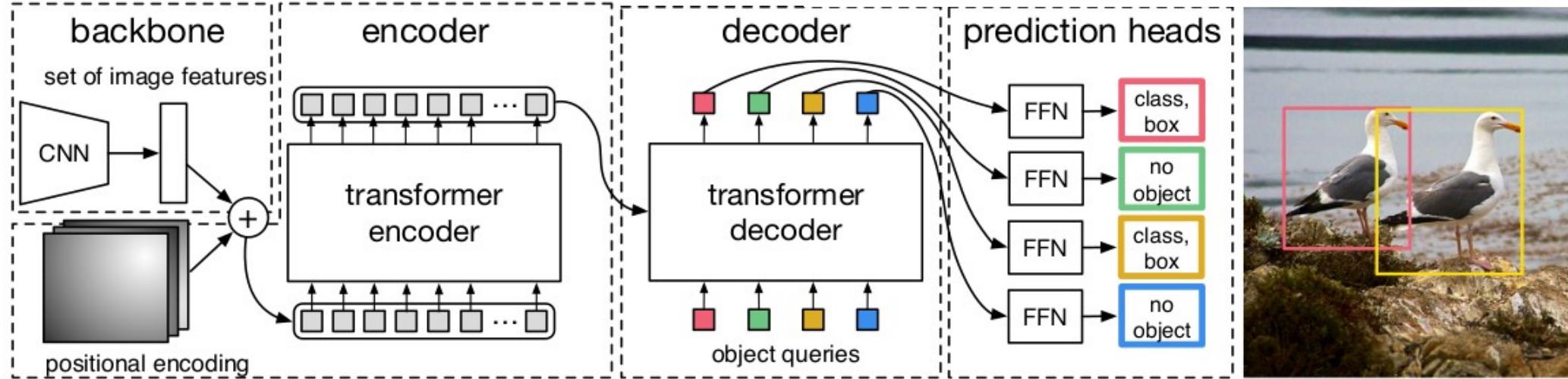
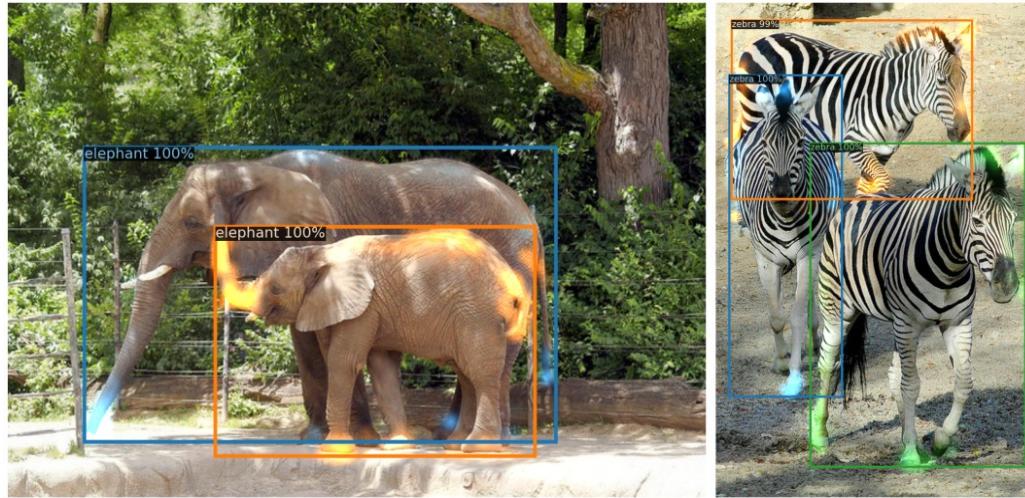
DETR:

- ▶ DETR [Carion et al. ArXiv:2005.12872] <https://github.com/facebookresearch/detr>
- ▶ ConvNet → Transformer
- ▶ Object-based visual reasoning

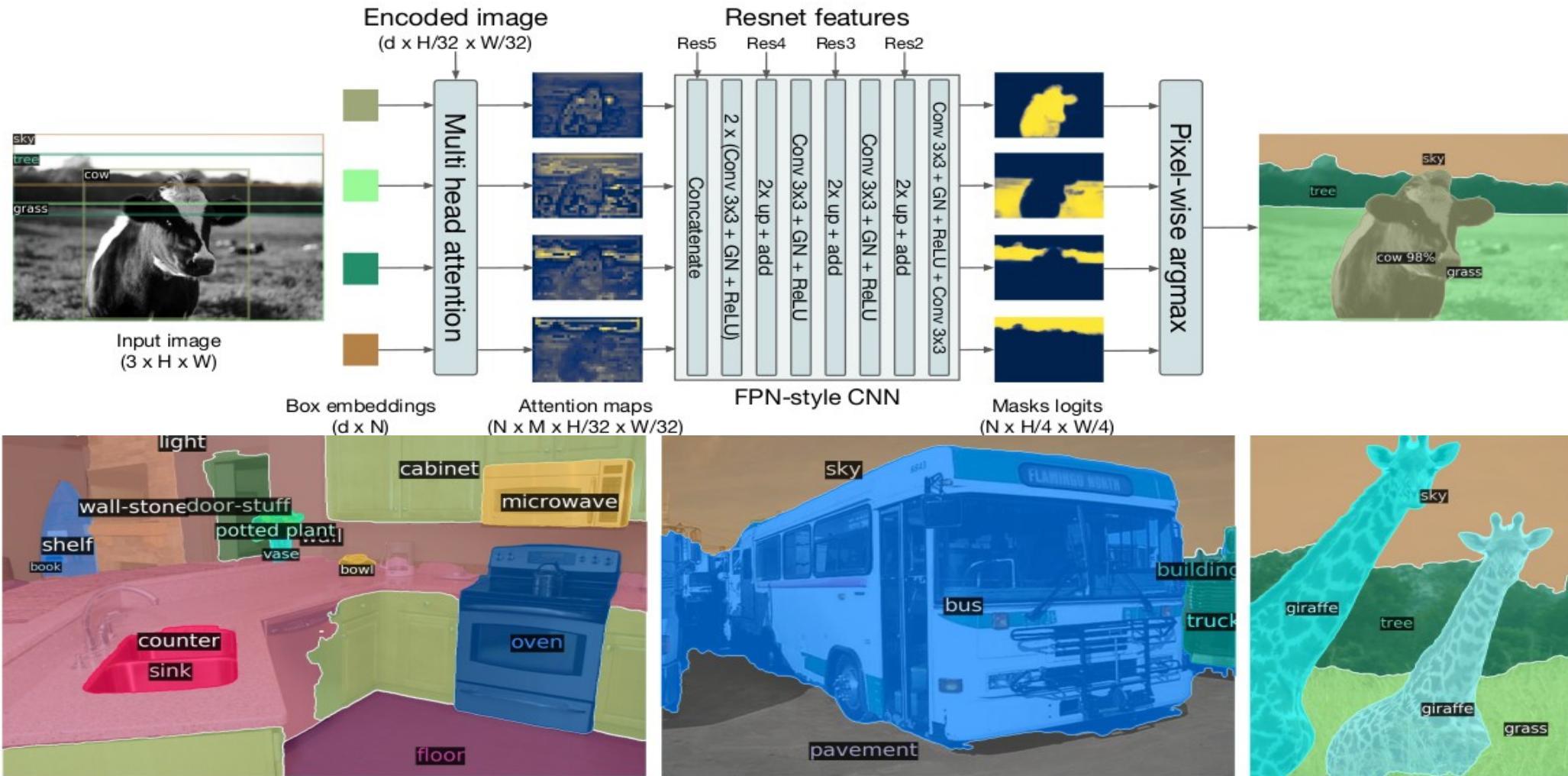


DETR: ConvNet → Transformer for object detection

- ▶ DETR [Carion et al. ArXiv:2005.12872]
- ▶ <https://github.com/facebookresearch/detr>
- ▶ ConvNet → Transformer
- ▶ Object-based visual reasoning
- ▶ Transformer: dynamic networks
- ▶ Through attention/associative mem.

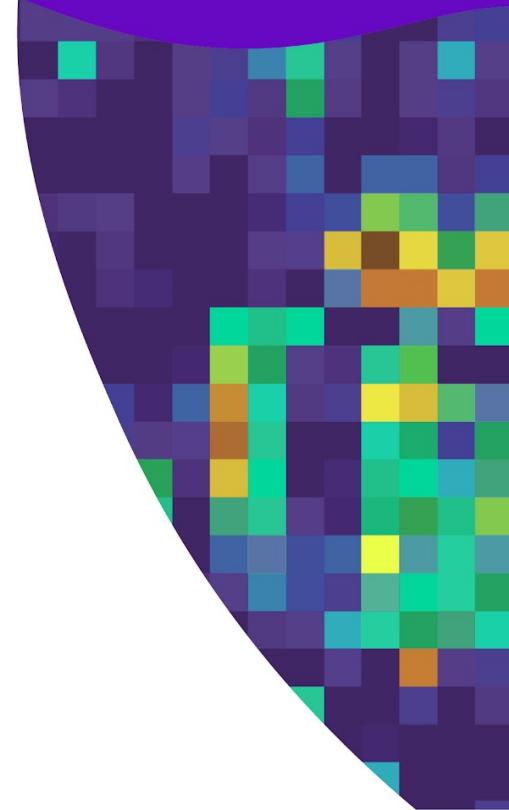


DETR: results on panoptic segmentation



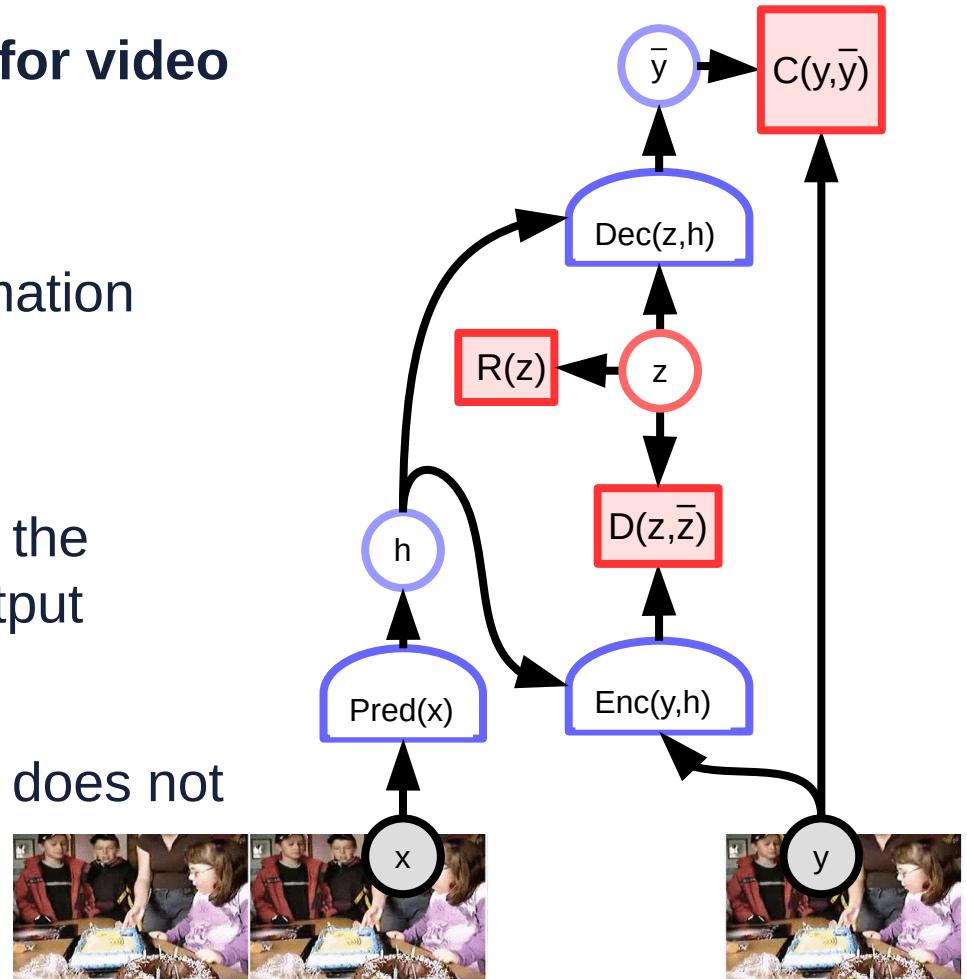
Learning World Models for Planning & Control

Problem: the world is unpredictable



Conditional Regularized AE

- ▶ Regularized Latent Variable EBM for video Prediction
- ▶ Predictor captures the useful information from the past in h
- ▶ Regularized latent variable capture the unpredictable information in the output
- ▶ Regularizer ensures latent variable does not capture all the information.

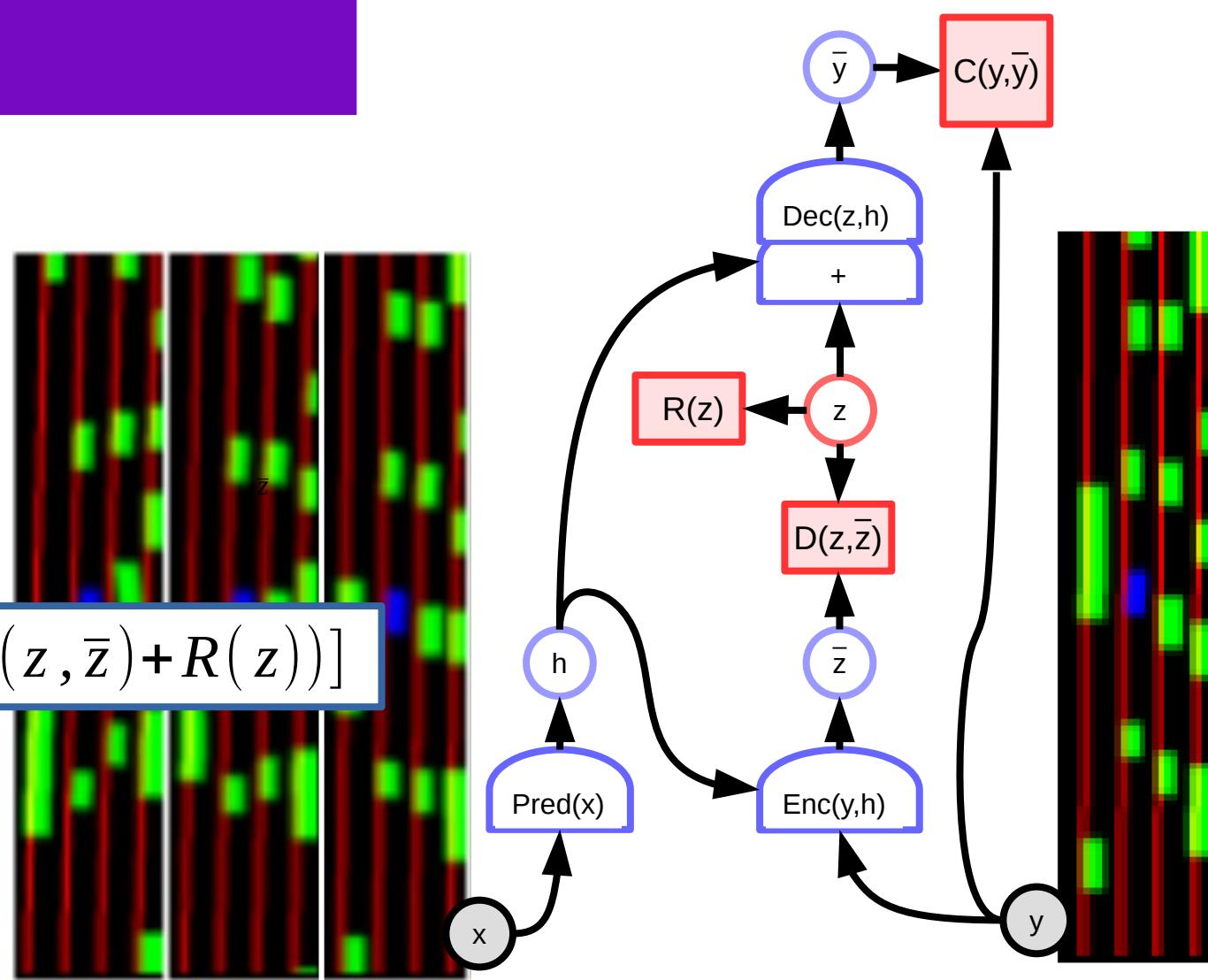


VAE + Drop Out

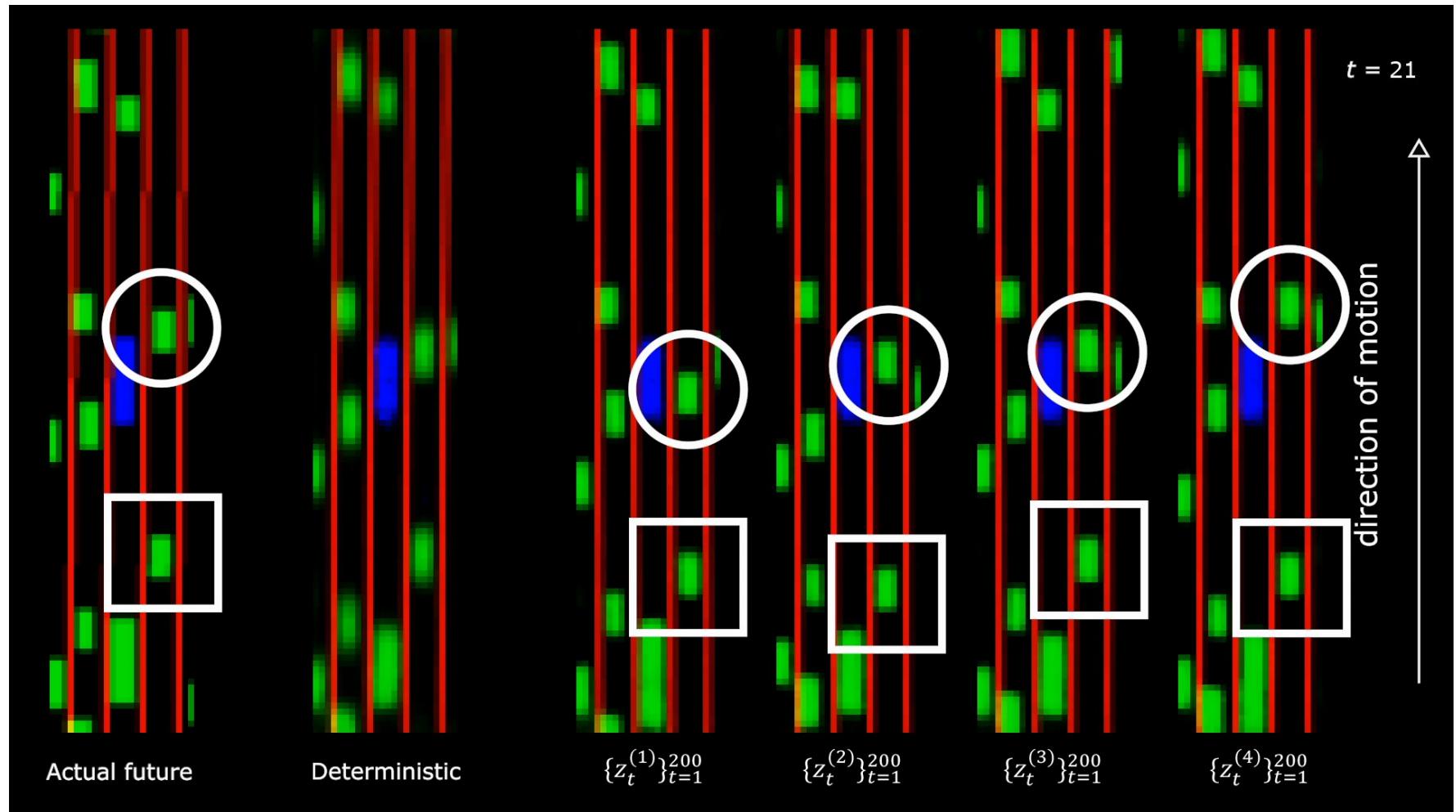
- ▶ **Training:**
 - ▶ Observe frames
 - ▶ Compute h
 - ▶ Predict \bar{z} from encoder
 - ▶ Sample z , with:

$$P(z|\bar{z}) \propto \exp[-\beta(D(z, \bar{z}) + R(z))]$$

- ▶ Half the time, set $z=0$
- ▶ Predict next frame
- ▶ backprop

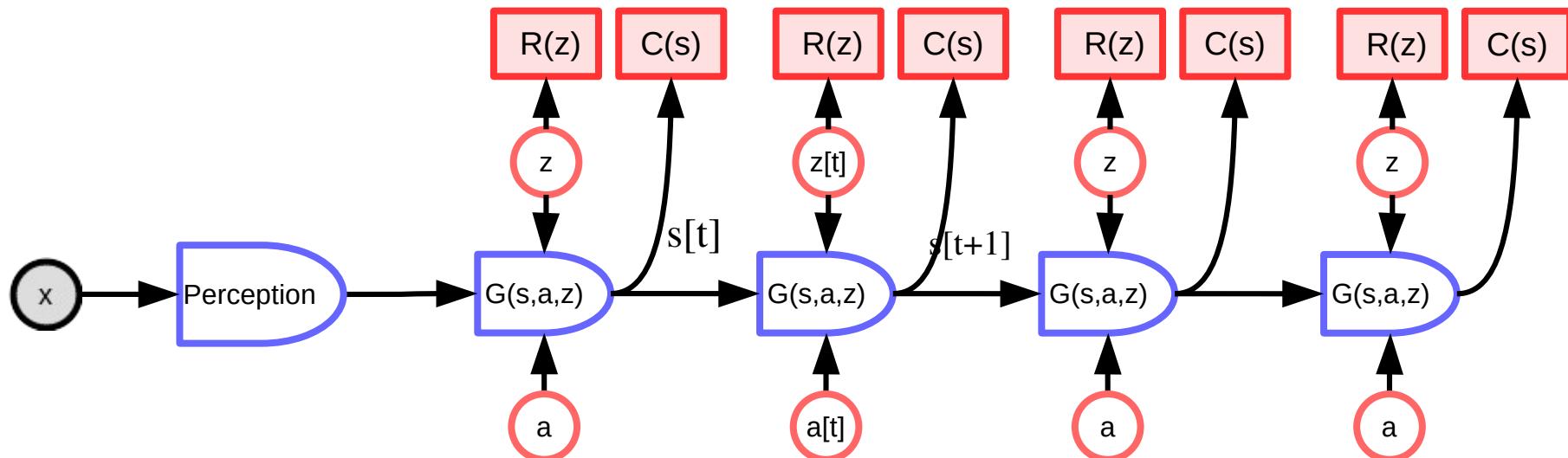


Actual, Deterministic, VAE+Dropout Predictor/encoder



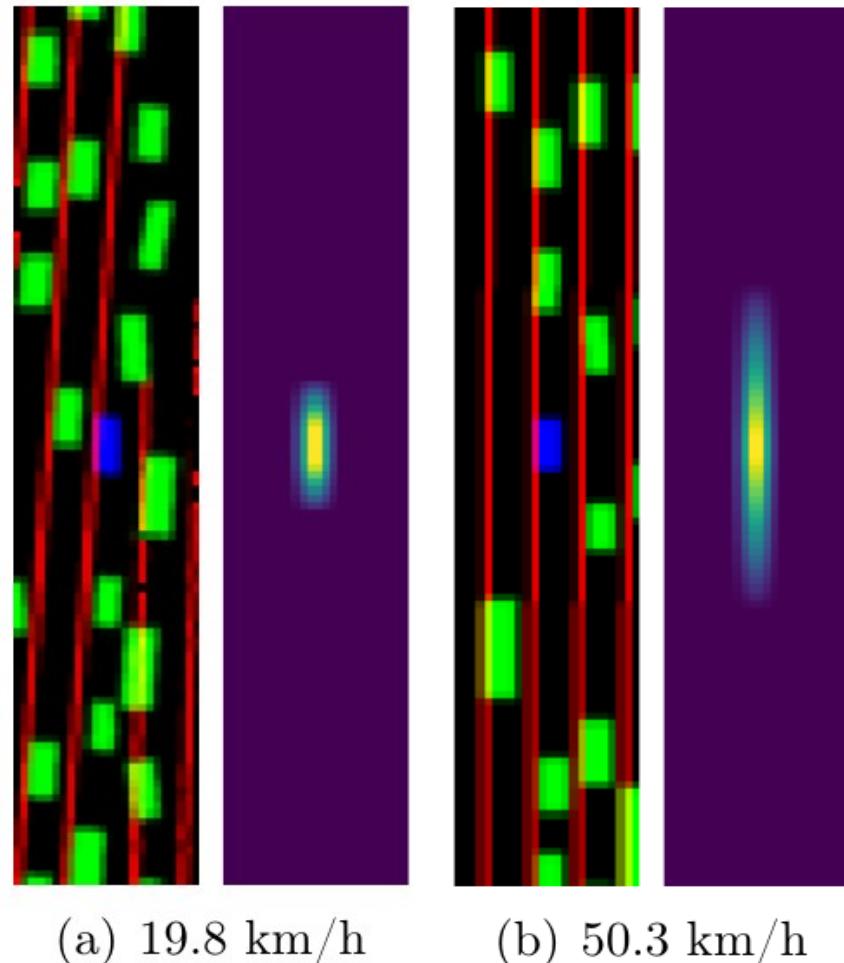
Forward Model for Model-Predictive Control

- ▶ Forward model: $s[t+1] = G(s[t], a[t], z[t])$
- ▶ Cost/Energy: $f[t] = C(s[t])$
- ▶ Latent variable z sampled from $q(z)$ proportional to $\exp(-R(z))$
- ▶ Optimize $(a[1], a[2], \dots, a[T]) = \operatorname{argmin} \sum_t C(s[t])$
through backprop (== Kelley-Bryson adjoint state method)



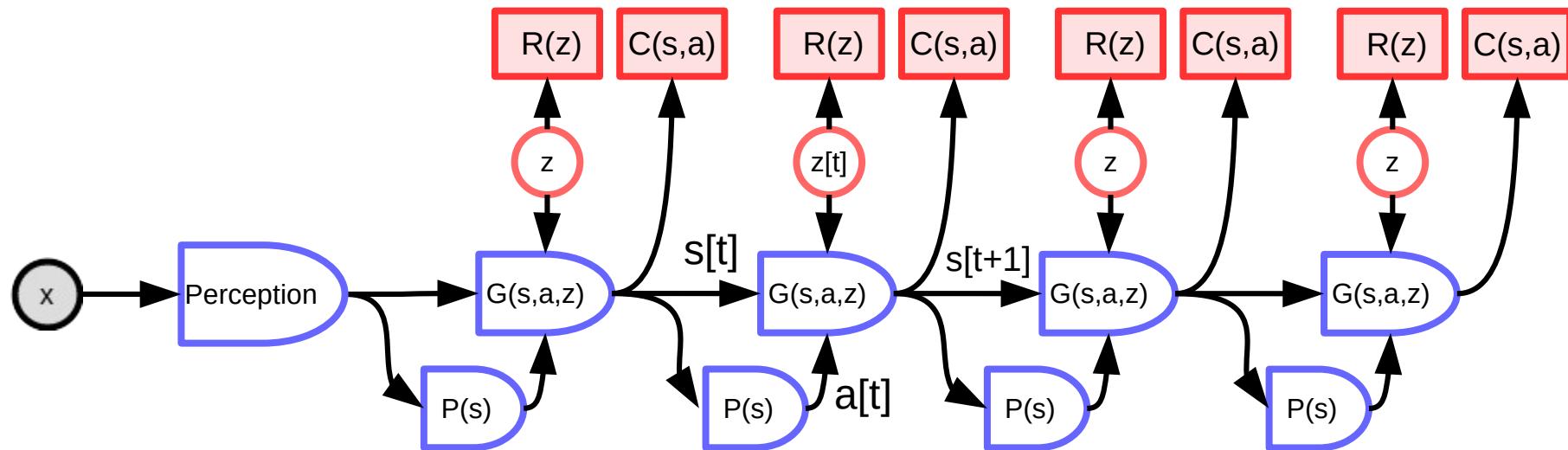
Cost optimized for Planning & Policy Learning

- ▶ **Differentiable cost function**
 - ▶ Increases as car deviates from lane
 - ▶ Increases as car gets too close to other cars nearby in a speed-dependent way
- ▶ **Uncertainty cost:**
 - ▶ Increases when the costs from multiple predictions (obtained through sampling of drop-out) have high variance.
 - ▶ Prevents the system from exploring unknown/unpredictable configurations that may have low cost.



Forward Model for Gradient-Based Policy Learning

- ▶ Forward model: $s[t+1] = G(s[t], a[t], z[t])$
- ▶ Cost/Energy: $f[t] = C(s[t], a[t])$
- ▶ Latent variable z sampled from $q(z)$ proportional to $\exp(-R(z))$
- ▶ Policy: $a[t] = P(s[t])$
- ▶ Learn P through backprop (== Kelley-Bryson adjoint state method)



conclusions

- ▶ **SSL is the future**
 - ▶ Learning hierarchical features in a task-invariant way
 - ▶ Plenty of data, **massive** networks
 - ▶ Learning Forward Models for Model-Based Control
 - ▶ Challenge: handling uncertainty in the prediction: energy-based models
- ▶ **Reasoning/planning through energy minimization**
 - ▶ Energy-Based Models with latent variables
 - ▶ Replace symbols by vectors and logic by continuous functions.
- ▶ **Learning hierarchical representations of action plans**
 - ▶ No idea how to do that!
- ▶ **There is no such thing as AGI.** Intelligence is always specialized.
- ▶ **We should talk about rat-level, cat-level, or human-level AI (HLAI).**

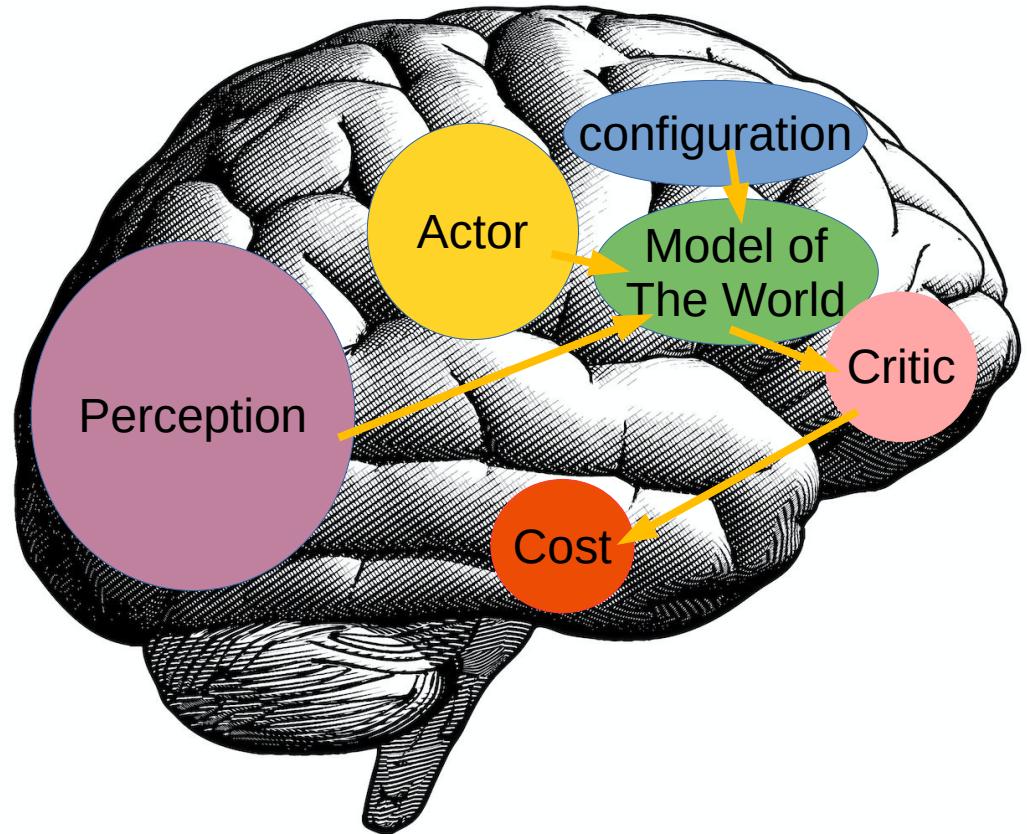
Speculation: Architecture for Autonomous Intelligence

- ▶ **World Model:** predicts future states
- ▶ **Critic:** predicts expected objective
- ▶ **Cost:** computes objective
- ▶ **Perception:** estimates world state
- ▶ **Actor:** computes action

- ▶ Humans only have one world model engine!

- ▶ Configurator: configures the world model engine for the situation at hand.

- ▶ Is this what consciousness is?



Speculations

- ▶ **Self-Supervised Learning is the future of AI / ML**
- ▶ **Models of the world:** Machines need to learn them (through SSL)
- ▶ Perhaps common sense will emerge from learning world models
- ▶ **Emotions are (often) anticipations of outcomes**
 - ▶ According to predictions from the model of the world
- ▶ **Reasoning is finding actions that optimize outcomes**
 - ▶ Constraint satisfaction/cost minimization rather than logic
- ▶ **Consciousness may be the deliberate configuration of our world model?**
 - ▶ We only have one (configurable) model of the world
 - ▶ If our brains had infinite capacity, we would not need consciousness
- ▶ **There is no such thing as AGI.** Intelligence is always specialized.
 - ▶ We should talk about rat-level, cat-level, or **human-level AI (HLAI)**.