THE PRINCIPLE OF COGNITIVE ACTION

THE CASE OF VISUAL FEATURES

Science is like sex: sometimes something useful
come out, but that is not the reason we are doing it

– Richard Feynman

Marco Gori
Dipartimento di Ingegneria dell'Informazione e
Scienze Matematiche

Università di Siena

ACDL 2018

Outline (con't)

Part II - The case of visual features

- The role of time and motion invariance
- CAL of convolutional visual features (shallow model)
- Focus of attention as a necessary computational issue
- Discretization in the retina of "shallow models"
- Deep convolutional nets

ROLE OF TIME AND MOTION INVARIANCE IN VISION

Ten questions one would like to ask

1. Why can animals conquer visual skills without intensive supervision?

2. How do animal conquer visual skills by unsupervised learning?

3. Can animals see in a world of shuffled frames?

4. How can humans attach semantic labels at pixel level?

5. How can humans provide a linguistic description of visual scenes?

role of time and motion invariance

Ten questions one would like to ask (con't)

6. Why is the visual cortex arranged according to a hierarchical structure? What about biological plausibility?

7. Why are there two different mainstreams in the visual cortex? (ventral (what) and dorsal (where) visual pathways)

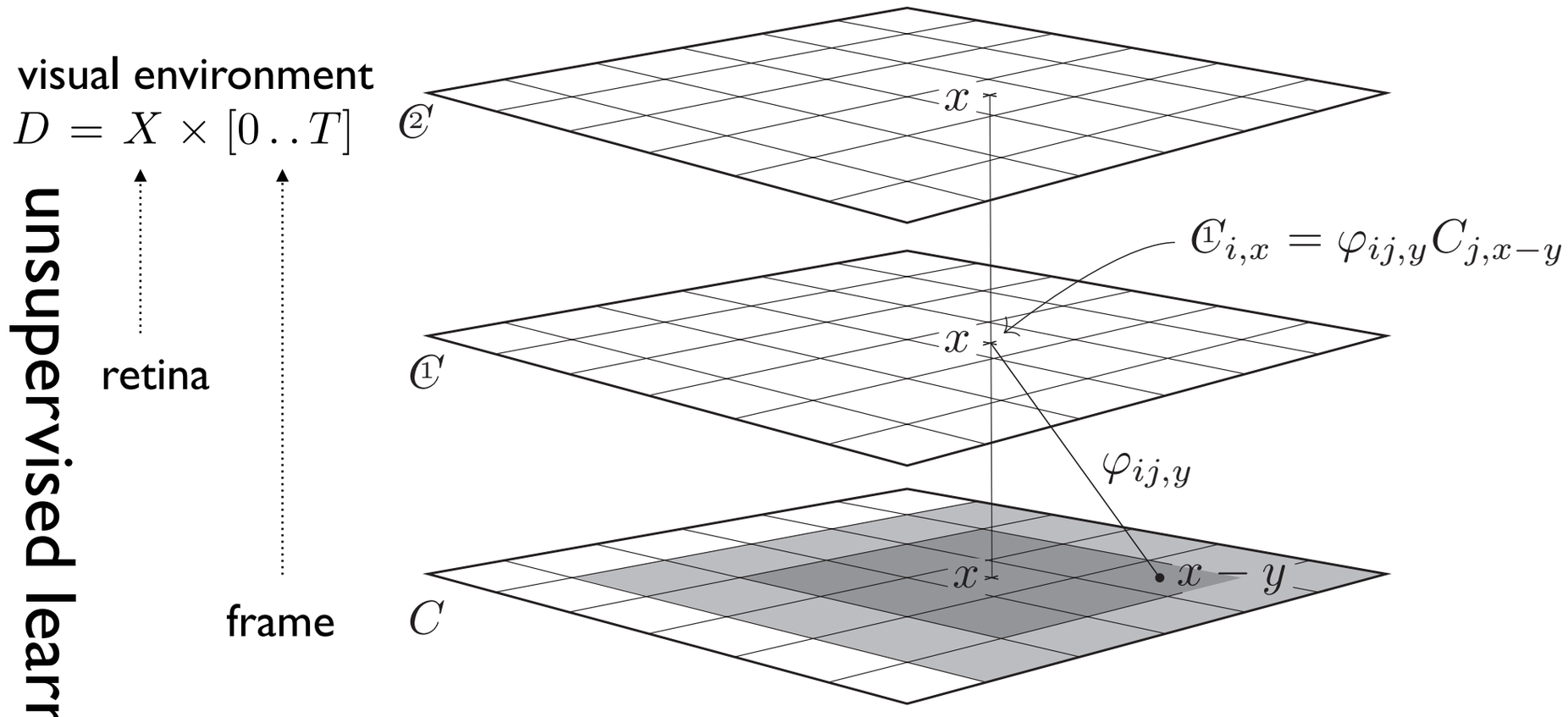
8. Why do foveal animals focus attention?

9. How do foveal animals perform eye movement?

10. Why does it take 8-12 months for newborns to achieve adult visual acuity?

role of time and motion invariance

Convolution to deal with context



$$\mathcal{C}^1_i(x,t) = \frac{1}{n} + \sum_{j=0}^{m-1} \int_X dy\ \varphi_{ij}(x,y,t) C_j(y,t) = \frac{1}{n} + (\varphi_t \times C_t)_i(x)$$

ACDL 2018

Visual constraints

$$V(t, w(t)) = U(w, u)$$

$$D = X \times [0\,..\,T]$$

ergodic assumption $d\mu = f(x,t)dxdt$

convolutional filter weights

$$S(Y \mid X, T, F) = -\int_D d\mu(x,t) \sum_{i=1}^{n} \mathbb{1}_i(x,t) \log \mathbb{1}_i(x,t)$$

$$f(x,t) = g(x - a(t))h(t)$$

focus of attention dissipative term

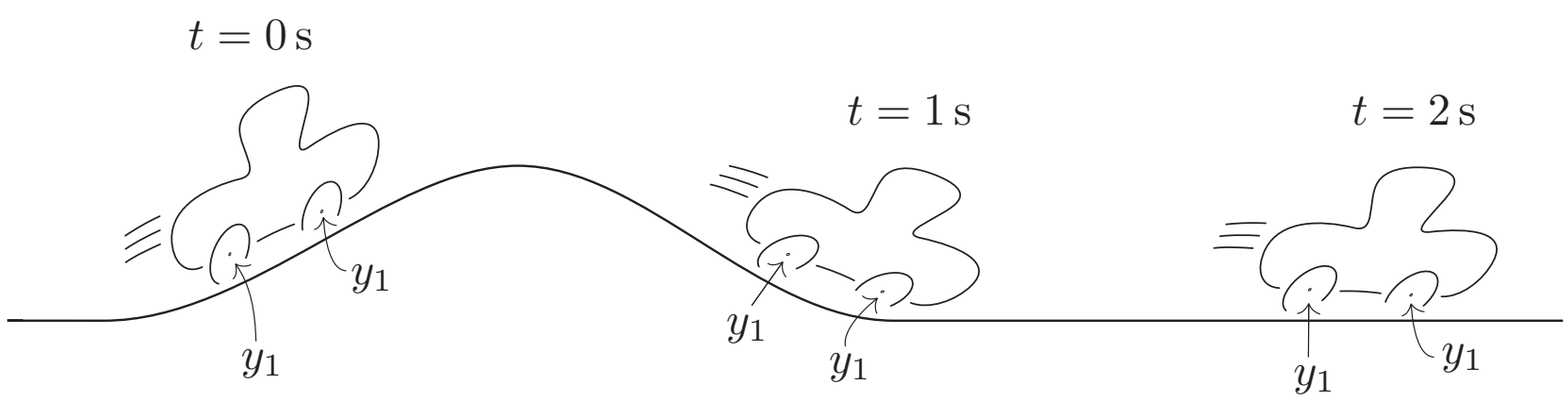
$$S(Y) = -\sum_{i=1}^{n} \Big(\int_D d\mu(x,t)\, \mathbb{1}_i(x,t) \Big) \log \Big(\int_D d\mu(x,t)\, \mathbb{1}_i(x,t) \Big)$$

$$I(Y; X, T, F) = S(Y) - S(Y \mid X, T, F).$$

unsupervised learning

ACDL 2018

Visual constraints (con't)



$$d\mathbb{C}/dt = 0$$

$$\partial_t \mathbb{C}_i + v_j \partial_j \mathbb{C}_i = 0$$

it's the supervision offered for free by Nature!

COGNITIVE ACTION LAWS OF CONVOLUTIONAL VISUAL FEATURES

ACDL 2018

Spatiotemporal regularization

$P_x = \partial_x \qquad P_t = \partial_t$ simplest solution

$$\frac{\lambda_P}{2}\int_D dtdx\, h(t)(P_x\varphi_{ij}(x,t))^2 + \frac{\lambda_K}{2}\int_D dtdx\, h(t)(P_t\varphi_{ij}(x,t))^2$$

Cognitive Action

$$\mathcal{A}_0(\varphi) = \int_D d\mu\, \mathbb{C}_i(\varphi) \cdot \log \int_D d\mu\, \mathbb{C}_i(\varphi) - \lambda_C \int_D d\mu\, \mathbb{C}_i(\varphi) \log \mathbb{C}_i(\varphi)$$
$$+ \lambda_1 \int_D d\mu \left(\sum_{i=1}^{n} \mathbb{C}_i(\varphi) - 1 \right)^2 - \lambda_0 \int_D d\mu\, \mathbb{C}_i(\varphi) \cdot [\mathbb{C}_i(\varphi) < 0]$$
$$+ \frac{\lambda_P}{2} \int_D dtdx\, h(t) (P_x \varphi_{ij}(x,t))^2 + \frac{\lambda_K}{2} \int_D dtdx\, h(t) (P_t \varphi_{ij}(x,t))^2$$
$$+ \lambda_M \int_D d\mu \left(\partial_t \mathbb{C}_i(\varphi) + v_j \partial_j \mathbb{C}_i(\varphi) \right)^2,$$

Cognitive Action

no information-based representation

$$\mathcal{A}(\varphi) = \frac{1}{2}\Big(\int_D d\mu\, \text①\mathcal{C}_i(\varphi)\Big)^2 - \frac{\lambda_C}{2}\int_D d\mu\, \text①\mathcal{C}_i^2(\varphi)$$
$$+ \frac{\lambda_1}{2}\int_D d\mu\, \Big(\sum_{i=0}^{n-1} \text①\mathcal{C}_i(\varphi) - 1\Big)^2 - \lambda_0 \int_D d\mu\, \text①\mathcal{C}_i(\varphi)\cdot[\text①\mathcal{C}_i(\varphi) < 0]$$
$$+ \frac{\lambda_P}{2}\int_D dtdx\, h(t)(P_x\varphi_{ij})^2 + \frac{\lambda_K}{2}\int_D dtdx\, h(t)(P_t\varphi_{ij})^2$$
$$+ \frac{\lambda_M}{2}\int_D d\mu\, \big(\partial_t \text①\mathcal{C}_i(\varphi) + v_j\partial_j \text①\mathcal{C}_i(\varphi)\big)^2,$$

ACDL 2018

Euler-Lagrange equations (con't)

$\lambda = \lambda_0,\ \nu = \lambda_1,\ \eta = \lambda_M,\ \beta = \lambda_C$

$$\lambda_K P_t^\star(h(t)P_t\varphi_{ij}(x,t)) + \lambda_P h(t) P_x^\star P_x \varphi_{ij}(x,t)$$
$$+ c_j(x,t)\cdot\Big(\int d\tau\, d\xi\, c_k(\xi,\tau)\varphi_{ik}(\xi,\tau) - \nu\Big) - \beta\int d\xi\, c_{jk}^2(x,\xi,t)\varphi_{ik}(\xi,t)$$
$$+ \nu\sum_{m=1}^{n}\int d\xi\, c_{jk}^2(x,\xi,t)\varphi_{mk}(\xi,t) - \lambda\int dz\, f(z,t)C_j(z-x,t)[\mathbb{C}_i(z,t) < 0]$$
$$+ \eta\int d\xi\big(\Xi_{jk}(x,\xi,t)\partial_t^2 + \Pi_{jk}(x,\xi,t)\partial_t + \Upsilon_{jk}(x,\xi,t)\big)\varphi_{ik}(\xi,t) = 0,$$

where $c_j(x,t) = \int dz\, f(z,t)C_j(z-x,t)$ and $c_{jk}^2(x,\xi,t) = \int dz\, f(z,t)C_j(z-x,t)C_k(z-\xi,t)$.

temporal locality

spatial locality

yet another result which suggests the importance of deep nets

Euler-Lagrange equations

Terms that are computable from the video

$$\Xi_{jk}(x,\xi,t) = -\int_D dz\, f(z,t) C_j(z-x,t) C_k(z-\xi,t);$$

$$\begin{aligned}\Pi_{jk}(x,\xi,t) = \int_D dz\, \big(& f(z,t) D_t C_j(z-x,t) C_k(z-\xi,t) \\ & - \partial_t (f(z,t) C_j(z-x,t) C_k(z-\xi,t)) \\ & + f(z,t) C_j(z-x,t) C_k(z-\xi,t) \big);\end{aligned}$$

$$\begin{aligned}\Upsilon_{jk}(x,\xi,t) = \int_D dz\, \big(& f(z,t) D_t C_j(z-x,t) D_t C_k(z-\xi,t) \\ & - \partial_t (f(z,t) C_j(z-x,t) C_k(z-\xi,t)) \big).\end{aligned}$$

Temporal locality

Enforce time locality by computing the entropy on frames rather than on the entire life of the agent:

$$\left(\int_D \mathbb{1}_i(x,t) f(x,t)\, dxdt\right)^2 \to \int_0^T dt \left(\int_X \mathbb{1}_i(x,t) f(x,t)\, dx\right)^2.$$

Define a causal entropy

$$s_i(t) = \int_0^t d\tau \int_X dx\, \mathbb{1}_i(x,t) f(x,t)$$

$$\frac{1}{T}\int_0^T s_i^2(t)\, dt + \alpha \int_0^T dt \left(s_i(t) - \int_0^t d\tau \int_X dx\, \mathbb{1}_i(x,\tau) f(x,\tau)\right)^2$$

Temporal locality (con't)

$$\frac{1}{T}\int_0^T s_i^2(t)\,dt + \alpha \int_0^T dt \left(\dot{s}_i(t) - \int_X dx\, \mathbb{G}_i(x,t) f(x,t)\right)^2$$

we gain temporal locality by adding an adjoint system

Space locality

$$Lg = \delta$$

$$L\Gamma_{ij}(x,\xi,t) = \frac{1}{g(\xi - a(t))}\big[c_j(x,t)c_k(\xi,t) - \beta c_{jk}^2(x,\xi,t) + \eta\big(\Xi_{jk}(x,\xi,t)\partial_t^2 + \Pi_{jk}(x,\xi,t)\partial_t + \Upsilon_{jk}(x,\xi,t)\big)\big]\varphi_{ik}(\xi,t),$$

$$L\Lambda_{ij}(x,\xi,t) = \frac{\nu}{g(\xi - a(t))}\sum_{m=1}^{n} c_{jk}^2(x,\xi,t)\varphi_{mk}(\xi,t)$$

$$\lambda_K P_t^\star(h(t)P_t\varphi_{ij}(x,t)) + \lambda_P h(t)P_x^\star P_x\varphi_{ij}(x,t) + \Gamma_{ij}(x,a(t),t) + \Lambda_j(x,a(t),t) - \nu c_j(x,t) - \lambda h(t)\int dz\, g(z)C_j(z-x,t)[\mathbb{C}_i(z,t) < 0] = 0.$$

we gain spatial locality by adding two adjoint systems

CAL equations like for other cognitive tasks: the emergence of receptive fields!

FOCUS OF ATTENTION
AS A NECESSARY COMPUTATIONAL ISSUE

ACDL 2018

Why do we need focus of attention?

$$\int_{\Omega} dP_{X,T,F} \longrightarrow \int_{D} d\mu$$

ergodic assumption $d\mu = f(x,t)dxdt$

$$f(x,t) = g(x - a(t))h(t)$$

focus of attention dissipative term

1. To properly measure the probability distribution

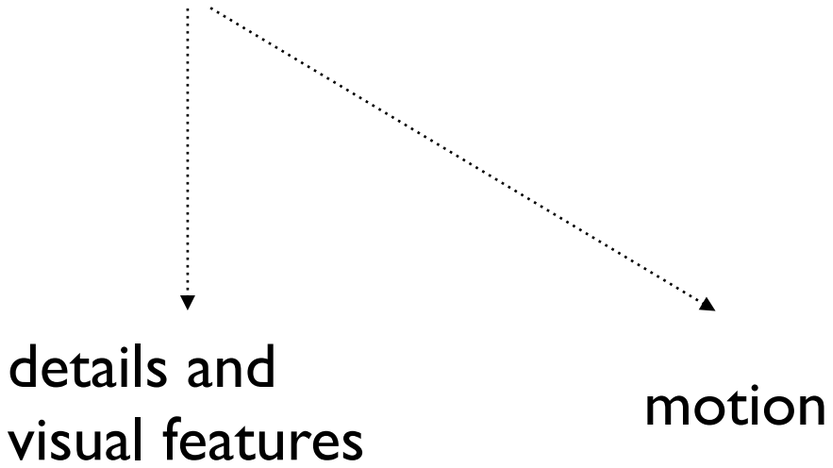
2. To guarantee space locality and facilitate motion invariance

3. One more reason … later

Gravitational eye movement model

What does attract your attention?

focus of attention mass (fam)



ACDL 2018

Gravitational eye movement model

fam

$$V(a) = \frac{Q\psi}{2\pi} \int_{\mathbf{R}} \log \frac{F(x)}{|x-a|} dx \qquad K(a) = \frac{1}{2}(m\psi)\dot{a}^2$$

$$g(x,t) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-a(t))^2}{2\sigma^2}}$$

inhibition map

$$\partial_t I(x,t) + \beta I(x,t) = g(x,t)$$

Gravitational eye movement model (con't)

$$V(t,a) = \frac{Q_1\psi}{2\pi}\int_{\mathbf{R}} \frac{F_1(x,t)}{|x-a|}(1-I(x,t))dx + \frac{Q_2\psi}{2\pi}\int_{\mathbf{R}} \frac{F_2(x,t)}{|x-a|}dx$$

details and features

motion

$F = K - V$

Euler-Lagrange equations

inhibition map

$$\ddot{a} + \theta\dot{a} + \alpha_1 \int_{\mathbf{R}} \frac{x-a}{|x-a|^2} F_1(x,t)(1-I(x,t))dx + \alpha_2 \int_{\mathbf{R}} \frac{x-a}{|x-a|^2} F_2(x,t)dx = 0$$

Gravitational eye movement model (con't)

Zanca & Gori, "Variational Laws of Visual Attention for Dynamic Scenes," NIPS 2017

1. fixations

2. smooth pursuit

3. saccadic movements

4. vestibulo-ocular movements

Saccadic movements carry out a fundamental segmentation to carry out motion invariance

DISCRETIZATION IN THE RETINA OF "SHALLOW MODELS"

One variable for each pixel …

$$A(w) = \int_0^T e^{\theta t}\left(\frac{\mu}{2}|\ddot{w}|^2 + \frac{\nu}{2}|\dot{w}|^2 + \gamma\dot{w}\ddot{w} + \frac{k}{2}|w|^2 + U(w,u)\right)dt$$

Euler-Lagrange Equations

$$\mu w^{(4)} + 2\theta\mu w^{(3)} + (\theta^2\mu + \theta\gamma - \nu)w^{(2)} + (\theta^2\gamma - \theta\nu)w^{(1)} + kw + \nabla_w U(w,u) = 0$$

$$w(0) = w^0, \quad \dot{w} = w^1, \quad \ddot{w}(0) = w^2, \quad w^{(3)}(0) = w^3$$

initial conditions

$$\hat{\mu}\ddot{w}(T) + \hat{\gamma}\dot{w}(T) = 0$$

$$\hat{\mu}w^{(3)}(T) + \dot{\hat{\mu}}\ddot{q}(T) + (\dot{\hat{\gamma}} - \hat{\nu})\dot{w}(T) = 0$$

boundary conditions

consistency needed

Controlling information overloading

- Filtering the input
- Reset of system dynamics

they correspond with the presence of null input

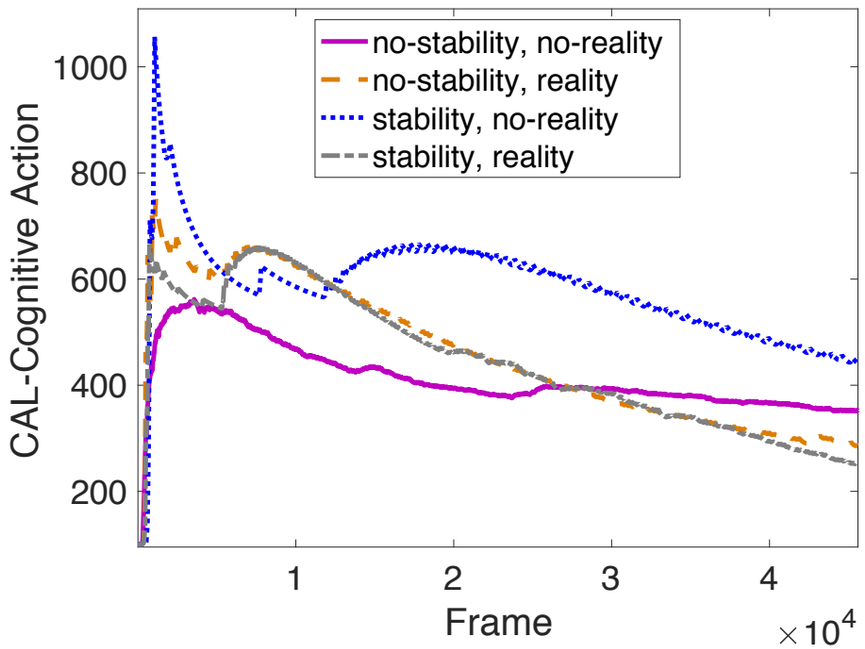
$$w(0), \quad \dot{w}(0) = 0, \quad \ddot{w}(0) = w^2, \quad w^{(3)}(0) = w^3$$

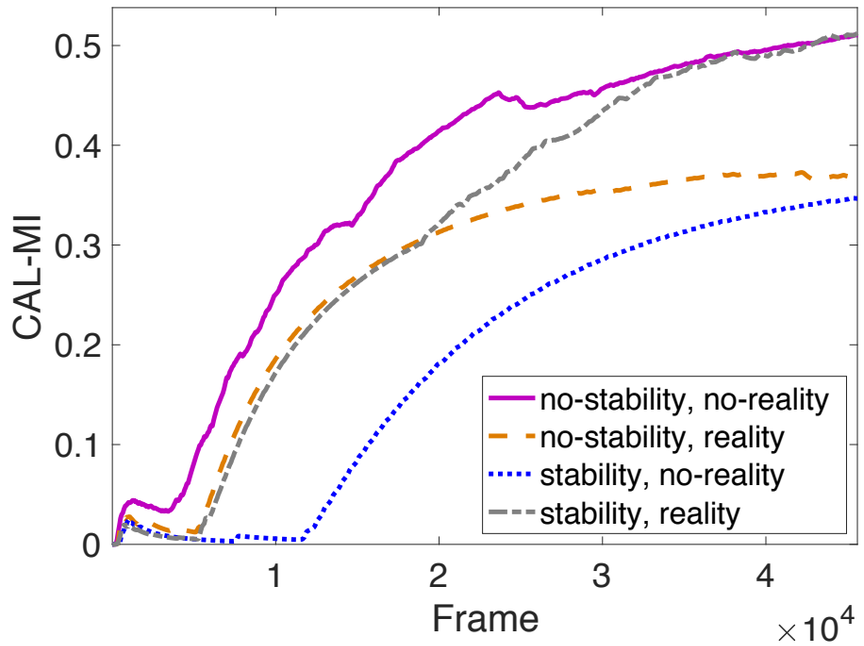
$$\vdots$$

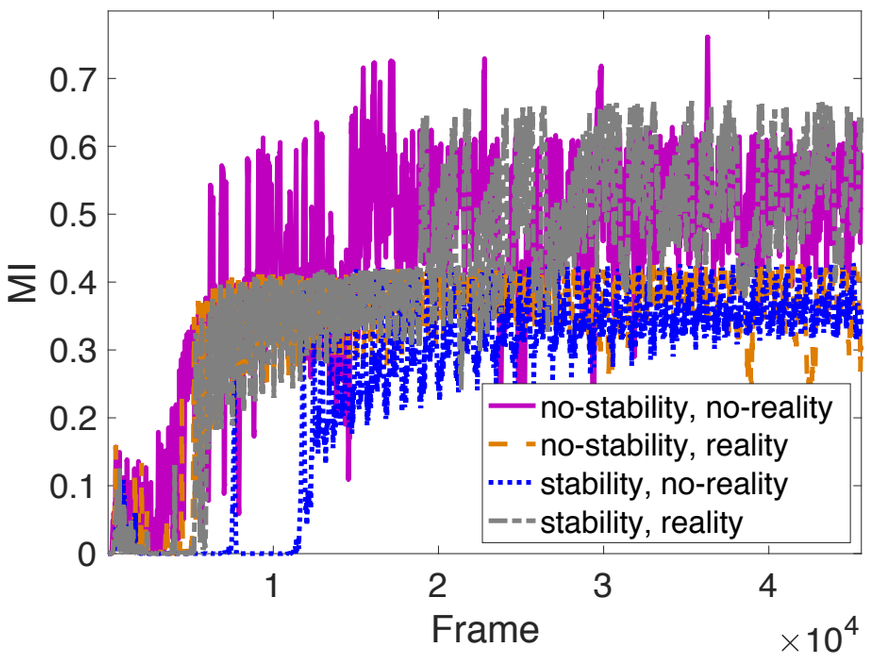
$$w^{(1)}(T) = 0, \; w^{(2)}(T) = 0, \; w^{(3)}(T) = 0$$

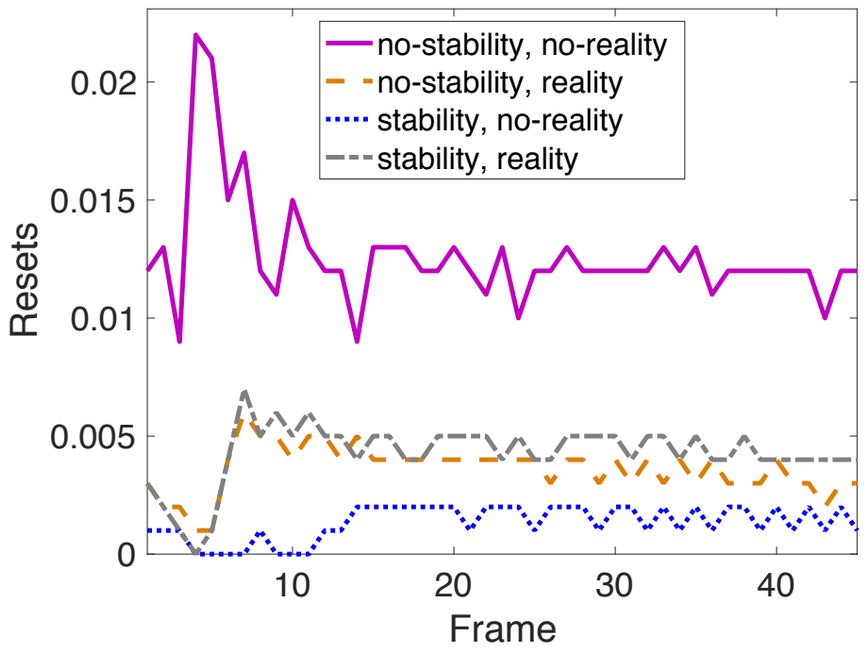
3. the reset can come from saccadic movements

Experiments





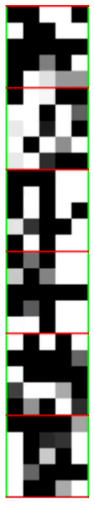




ACDL 2018

Video

Motion



Filters

Feature 1


```
{
  "status": "day",
  "rho": 0.0002999700082000345,
  "action_cur": 832.4277954101562,
  "mi_real_full": 0.00017303228378295898,
  "motion_full": 0,
  "mi_real": 0.00017303228378295898,
  "mi": 0.00010585784912109375,
  "ce": 1.0000181198120117,
  "minus_ge": -0.0001239776611328125,
  "motion": 0,
  "norm_q": 52.456607818603516,
  "norm_q_mixed": 1.5736880687455823e-9,
  "norm_q_dot": 5.2456615440288346e-11,
  "norm_q_dot_dot": 4.7210324538582427e-8,
  "norm_q_dot_dot_dot": 0.0000047209064177
  "eps1": 960,
  "eps2": 960,
  "eps3": 960
}
```


Feature 2

Feature 3

Feature 4

Feature 5

Feature 6

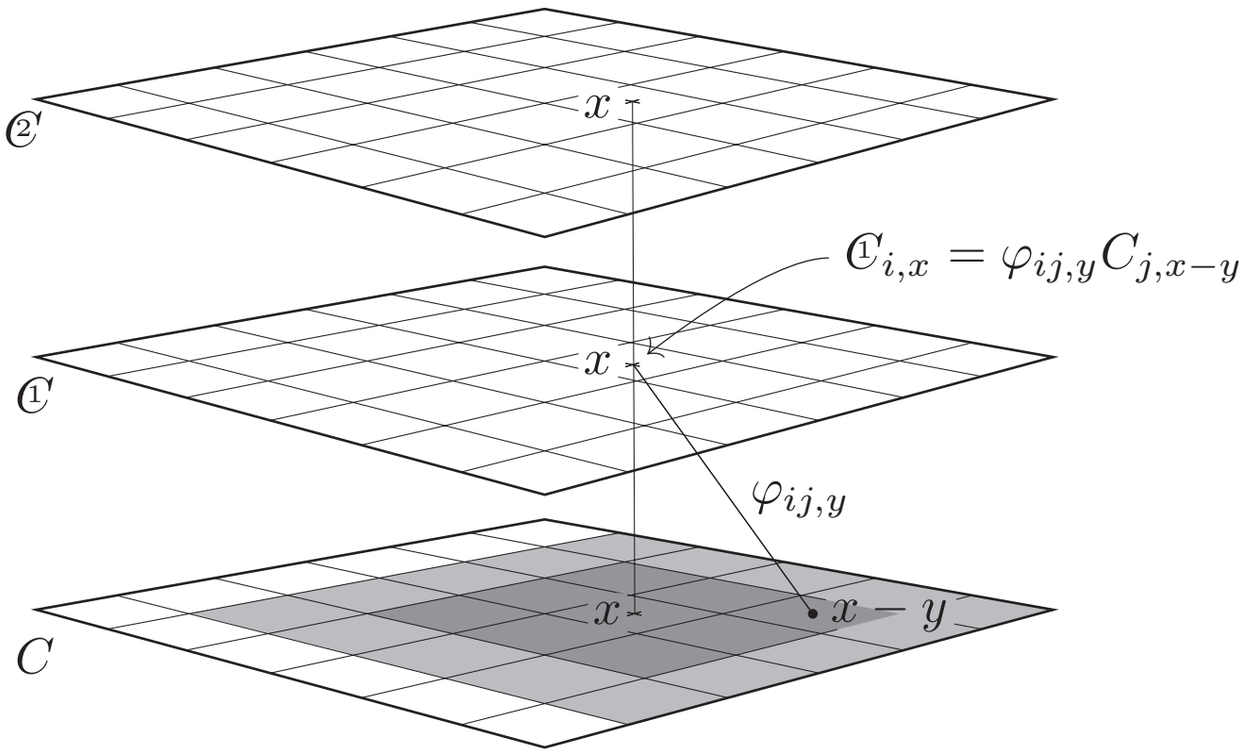
DEEP CONVOLUTIONAL NETS WITH FOCUS OF ATTENTION

Deep Convolutional Nets with focus of attention

$\ell = 1, \ldots, L$

$\varphi_{ij}^{\ell}(x,t),\ A_i^{\ell}(x,t),\ C_i^{\ell}(x,t)$

Euler-Lagrange eq. re-written for each layer! Focus of attention reduces to a 1-D problem by dramatically disentangling the factors of variation underlying visual data



no pooling, always convolution!

Architectural constraints

and biological plausibility

minimize $$E(w) = \sum_{\kappa=1}^{\ell} \sum_{i \in O} V(f(x_{\kappa i}), y_{\kappa i})$$

subject to
$i \in H \cup O$
$\kappa = 1, \ldots, \ell$
$$g_{\kappa i} = x_{\kappa i} - \sigma\left(\sum_{j \in pa(i)} w_{ij} x_{\kappa j} \right) = 0$$

$$L(\lambda, w) = \sum_{\kappa=1}^{\ell} \sum_{i \in O} V(x_{\kappa i}, y_{\kappa i}) + \sum_{i \in H \cup O} \sum_{\kappa=1}^{\ell} \lambda_{\kappa i} \left(x_{\kappa i} - \sigma\Big(\sum_{j \in pa(i)} w_{ij} x_{kj} \Big) \right)$$

Gradient descent/ascent

saddle points of the Lagrangian

$$w_{ik} \leftarrow w_{ij} - \eta_w \partial_{x_{ij}} E$$

$$x_{ik} \leftarrow x_{ij} - \eta_x \partial_{x_{ij}} E$$

$$\lambda_{\kappa i} \leftarrow \lambda_{\kappa i} + \eta_\lambda \partial_{\lambda_{\kappa i}} E$$

$$g_{\kappa i} = x_{\kappa i} - \sigma\left(\sum_{j \in pa(i)} w_{ij} x_{\kappa j} \right) = 0$$

Ten questions one would like to ask

1. Why can animals conquer visual skills without intensive supervision?

Disentangle the factors of variation underlying visual data by the "what" neurons:
A fundamental form of "supervision" is carried out by motion invariance

2. How do animal conquer visual skills by unsupervised learning?

Information-based learning and motion invariance

3. Can animals see in a world of shuffled frames?

My conjecture: No! This means nowadays CN are likely to solve a very difficult problem. There's room for attacking a simpler one …

Ten questions one would like to ask

4. How can humans attach semantic labels at pixel level?

   It's an open problem, but (pixel) semantic labeling looks like a primitive task of fundamental importance for other high level visual cognitive tasks

5. How can humans provide a linguistic description of visual scenes?

   Such a description is likely to emerge from appropriate "aggregations" of pixel-wise cognitive processes. Bounding boxes and other clever computer vision solutions might not play the necessary primitive role for supporting high level linguistic descriptions

Ten questions one would like to ask

6. Why is the visual cortex arranged according to a hierarchical structure? What about biological plausibility?

The "shallow convolutional net" yields a solution which is neither local in time nor in space! The deep net can be trained without Backpropation-like algorithms!

7. Why are there two different mainstreams in the visual cortex? (ventral (what) and dorsal (where) visual pathway)

The "what" neurons are those on which one enforces "motion invariance", whereas the "where" neurons are those on which we don't!

Ten questions one would like to ask

8. Why do foveal animals focus of attention?

   (a) to measure the probability distribution, (b) to favor local in space solutions, and (c) to properly segment object fixations by "resets" that handle information overloading

9. How do foveal animals perform focus of attention?

   focus of attention is driven by details and feature maps as well as by smooth pursuit. The gravitational field that yields the focus of attention (foa) contributes to visual feature extraction that, in turn, drives the foa. Likewise the velocity field can be regarded as "yet another Lagrangian coordinate"

Ten questions one would like to ask

10. Why does it take 8-12 months for newborns to achieve adult visual acuity?

Video blurring in newborns might be rooted in the need of facing information overloading that has been addressed when making Cauchy and boundary conditions properly consistent. Saccadic movements provide and additional "reset" contribution. The blurring stage can be regarded as the learning stage, where a causal optimization takes place.

Conclusions

- The role of time: Learning as an ordering process which requires energy dissipation

- Boundary conditions and information overloading

- Causal learning as an equilibrium problem

- Cognitive action in living systems and machines: we plan to have in the lab a“living visual agent” inspired to information-based laws soon (feature only).