Model generation for machine intelligence

Vadim Strijov

Moscow Institute of Physics and Technology
FRC CSC of the Russian Academy of Sciences

2018, February 7th

To start an applied project an expert and an analyst set

1. Project goal (the expected result of development)
   main purpose of research
2. Project application (how the project result will be applied)
   environment of measures and impacts
3. Historical data description (data formats and timing)
   algebraic structures of data
4. Quality criteria (how the project quality is measured)
   error function
5. Feasibility of the project (how to prove the project feasibility, list possible risks)
   error analysis

How long the model lives after being put on operation? What replaces it after?

Quality criteria for model generation and selection

Three sources of quality criteria

1. Business: model operation productivity, agent impact to environment
2. Theory: statistical hypothesis, bayesian inference
3. Technology: optimization requirements, resources

The main criteria of model quality

- Precision: MAPE, AUC
- Stability (diversity): std deviation for prediction, covariance of parameters
- Complexity: structure complexity, MDL, evidence of model

Problem statement for machine learning

Formal problem statement, an analyst has to set

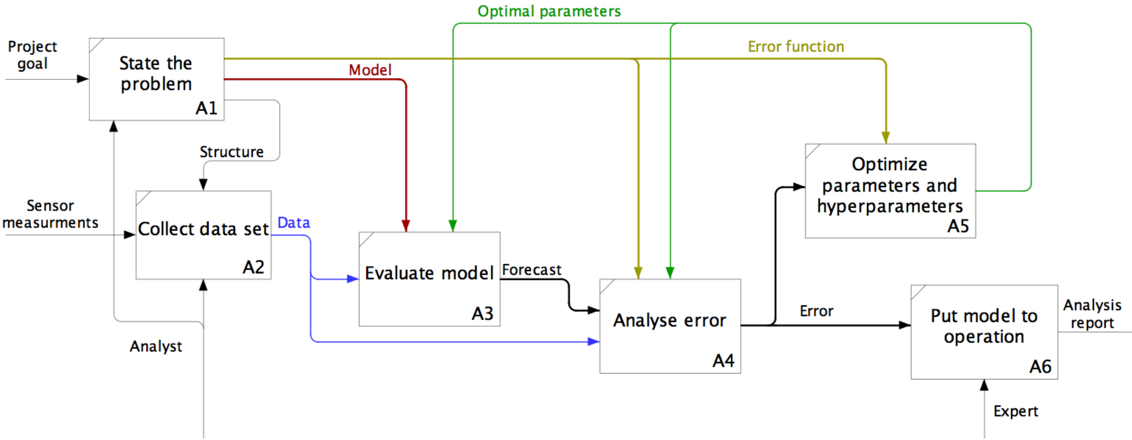
1) an algebraic structure for the dataset from measurements
2) a data generation hypothesis from 1)
3) a model, or a mixture from 2)
4) an error function (quality criteria with restrictions) from 2)
5) an optimization algorithm from 3) and 4)

The result of the model construction is a Cartesian product

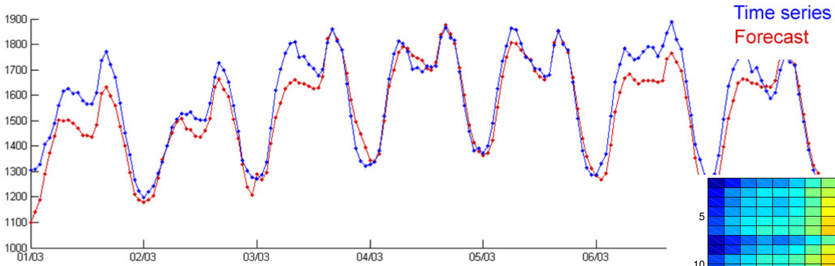
$$\{\textbf{models} \times \textbf{datasets} \times \textbf{quality criteria}\}.$$

Def: *Big data rejects the i.i.d. (independent and identically distributed random variables) data generation hypothesis from 2). It requests a mixture model.*

Analyst creates a model for expert to put it to operation



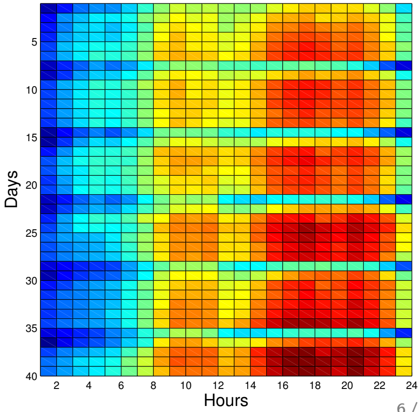
Model selection in forecasting



In terms of regression
$\hat{\mathbf{y}} = \mathbf{f}(\mathbf{X}, \mathbf{w}) = \mathbf{X}\mathbf{w}$, a class of linear models.

Classes of models to select from are RBF, NN, SVM, CNN, etc.

$$\left[\begin{array}{c|c} \underset{1\times 1}{\hat{s}_T} & \underset{1\times n}{\mathbf{x}_{m+1}} \\ \hline \underset{m\times 1}{\mathbf{y}} & \underset{m\times n}{\mathbf{X}} \end{array}\right] =$$



Binary representation of the model structure

Select a model f from a class $\mathfrak{F}$ by optimizing binary vector $\mathbf{a} \in \mathbb{B}^n$,

$$\hat{y} = f(\mathbf{w}, \mathbf{x}) = a_1 w_1 x_1 + \cdots + a_n w_n x_n$$

for the linear model

$$f(\mathbf{w}, \mathbf{x}) = \mathbf{x}^\mathsf{T}\mathbf{w}$$

and for the neural network

$$\mathbf{f}(\mathbf{w}, \mathbf{x}) = \frac{\exp\big(\mathbf{h}(\mathbf{x})\big)}{\sum_j \exp\big(h_j(\mathbf{x})\big)}, \qquad \mathbf{h}(\mathbf{x}) = \mathbf{W}_2^\mathsf{T}\mathbf{tanh}(\mathbf{W}_1^\mathsf{T}\mathbf{x}), \qquad \mathbf{w} = \mathbf{vec}(\mathbf{W}_1 \vdots \mathbf{W}_2),$$

according to the optimal brain damage method the structure vector

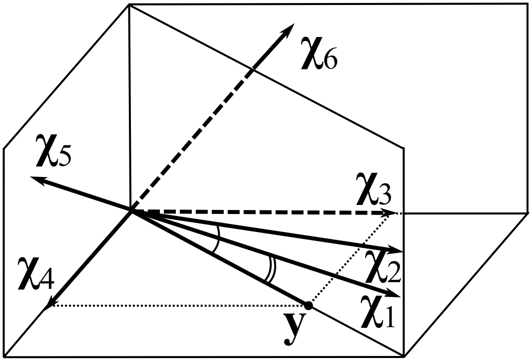
$$\mathbf{e}_i^\mathsf{T}\Delta\mathbf{w} + w_i = 0$$

with i-th element of $\mathbf{e}$ equals 1, the rest equal 0.

The model is defined by a vertex on the n-dimensional cube.

Select a stable and precise model given set of features

The sample contains multicollinear χ_1, χ_2 and noisy χ_5, χ_6 features, columns of the design matrix $\mathbf{X}$. We want to select two features from six.

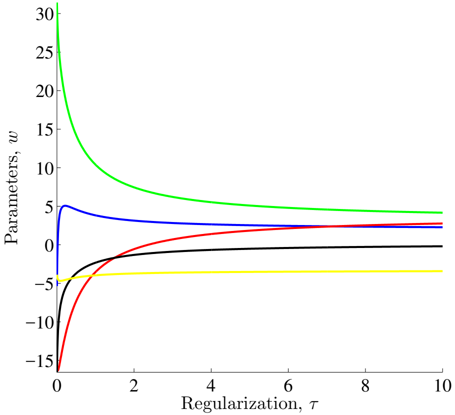


Stability and accuracy for a fixed complexity

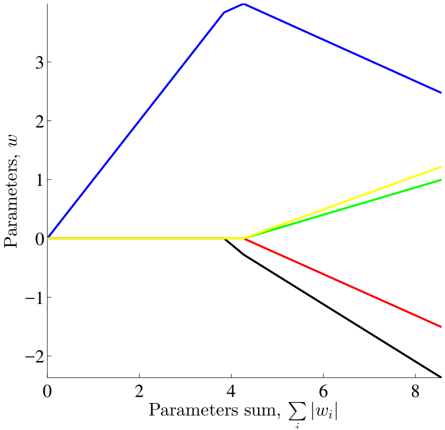
The solution: χ_3, χ_4 is an orthogonal set of features minimizing the error function.

Model parameter values with regularization

Vector-function $\mathbf{f} = \mathbf{f}(\mathbf{w}, \mathbf{X}) = [f(\mathbf{w}, \mathbf{x}_1), \dots, f(\mathbf{w}, \mathbf{x}_m)]^\mathsf{T} \in \mathbb{Y}^m$.



$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2 + \gamma^2\|\mathbf{w}\|^2$$



$$S(\mathbf{w}) = \|\mathbf{f}(\mathbf{w}, \mathbf{X}) - \mathbf{y}\|^2, \; T(\mathbf{w}) \leqslant \tau$$

Minimize number of similar and maximize number of relevant features

The model is defined by a vertex point in the n-dimensional cube.

Introduce a feature selection method QP(Sim, Rel) to solve the optimization problem

$$\mathbf{a}^* = \arg\min_{\mathbf{a}\in\mathbb{B}^n} \mathbf{a}^{\mathsf{T}}\mathbf{Q}\mathbf{a} - \mathbf{b}^{\mathsf{T}}\mathbf{a},$$

Number of correlated features Sim $\to$ min, number of correlated to the target Rel $\to$ max.

where matrix $\mathbf{Q} \in \mathbb{R}^{n\times n}$ of pairwise similarities of features $\boldsymbol{\chi}_i$ and $\boldsymbol{\chi}_j$ is

$$\mathbf{Q} = [q_{ij}] = \mathsf{Sim}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) = \left|\mathrm{Cov}(\boldsymbol{\chi}_i, \boldsymbol{\chi}_j) \div \sqrt{\mathrm{Var}(\boldsymbol{\chi}_i)\mathrm{Var}(\boldsymbol{\chi}_j)}\right|$$

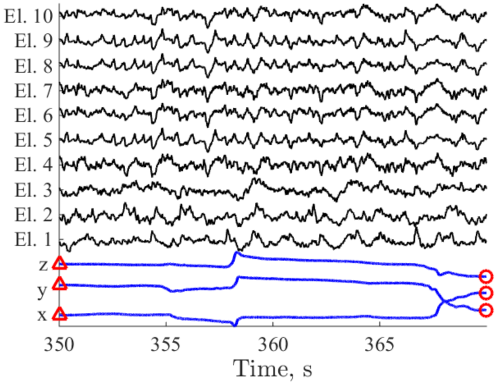
and vector $\mathbf{b} \in \mathbb{R}^n$ of feature relevances to the target is

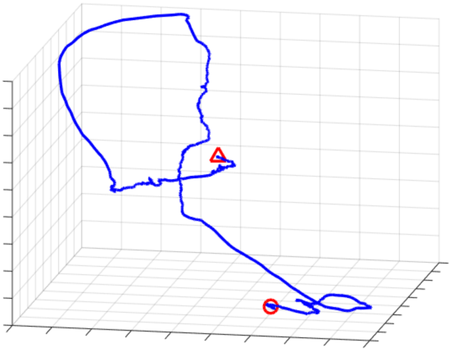
$$\mathbf{b} = [b_i] = \mathsf{Rel}(\boldsymbol{\chi}_i),$$

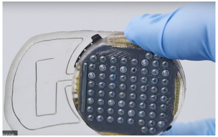
elements b_i are absolute values of the correlation between feature $\boldsymbol{\chi}_i$ and the target $\mathbf{y}$.

Katrutsa, Strijov. 2017. Comprehensive study of feature selection methods to solve multicollinearity problem // Expert Systems with Applications

WIMAGINE (clinatec.fr) 64-Channel ECoG implant and physical motion



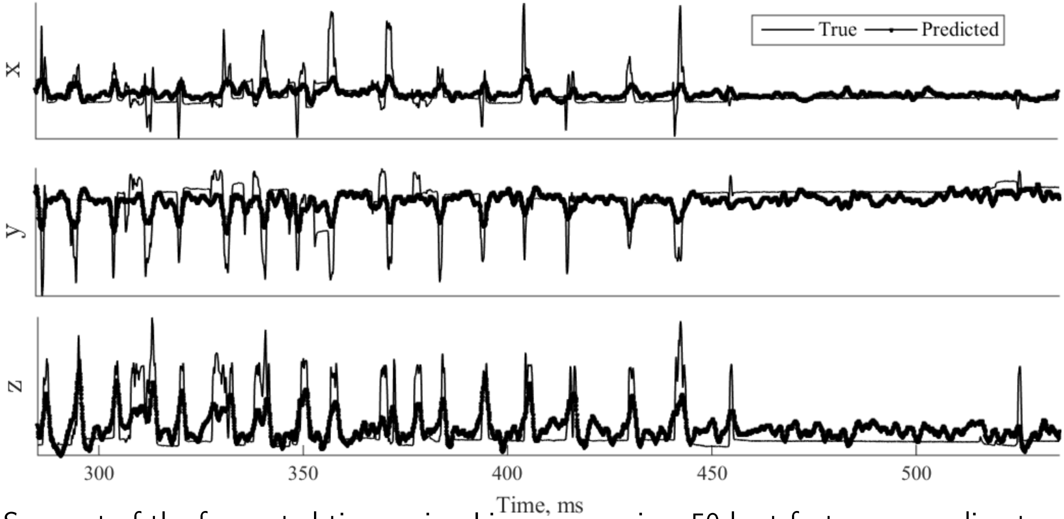




Extracts (350–370s) from voltage and wrist position time series for monkey A and 3D wrist trajectory for the same extract.

Motrenko, Strijov, 2018. Multi-way feature selection for ECoG-based BCI //Expert Systems with Applications, sub.

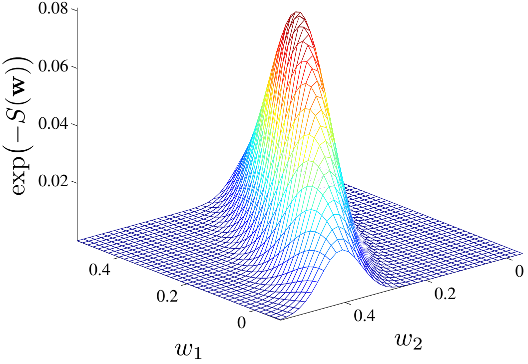
The wrist motion trajectory prediction with ECoG

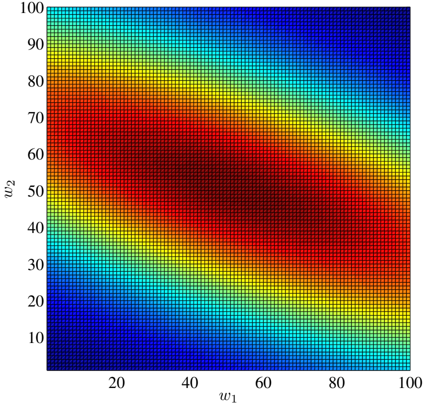


Segment of the forecasted time series. Linear regression, 50 best features according to multi-way QPFS (from 1000 highly-correlated features).

Empirical distribution of model parameters

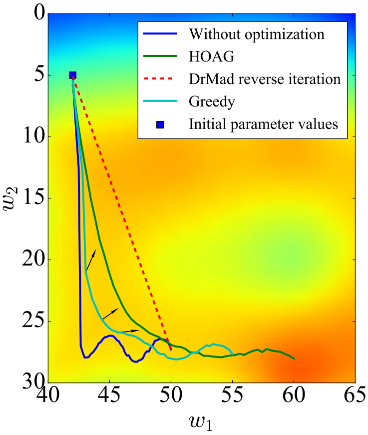
There given a sample $\{\mathbf{w}_1, \ldots, \mathbf{w}_K\}$ of realizations of the m.r.v. $\mathbf{w}$ and an error function $S(\mathbf{w}|\mathfrak{D}, \mathbf{f})$. Analyze the set $\{s_k = \exp\bigl(-S(\mathbf{w}_k|\mathfrak{D}, \mathbf{f})\bigr) | k = 1, \ldots, K\}$.

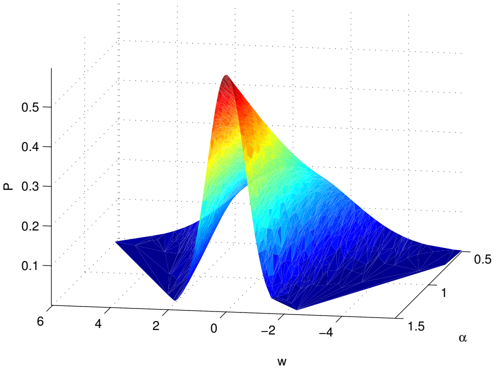




Kuznetsov, Tokmakova, Strijov. 2016. Analytic methods of structure parameter // Informatica

No one expected convergence for various priors...



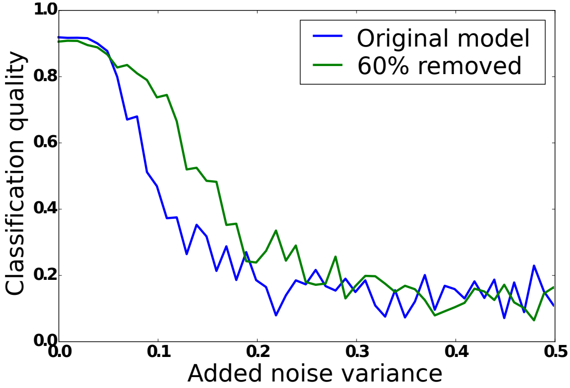


... since there is no convergence even for a single prior.

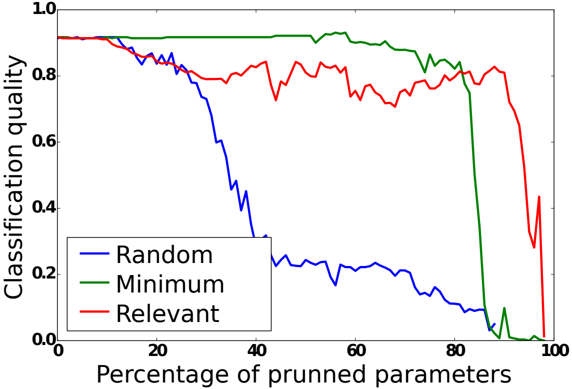
Prior of parameters $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1})$ with inverted parameter variance $\mathbf{A} = \alpha \mathbf{I}$ versus posterior distribution $p(\mathbf{w}|\mathfrak{D}, \mathbf{A}, \mathbf{B}, \mathbf{f})$.

Forecasting quality does not change until almost all connections removed

Model stability

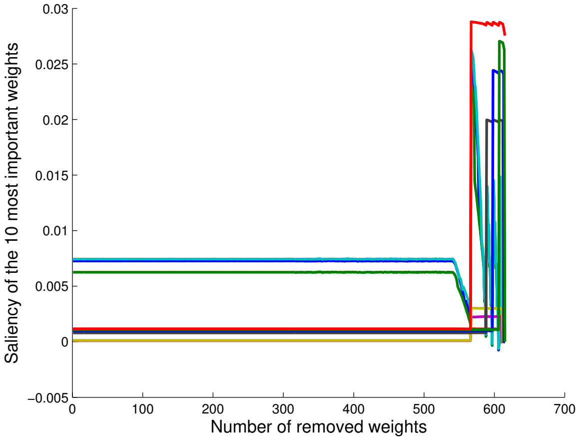


Redundancy of parameters



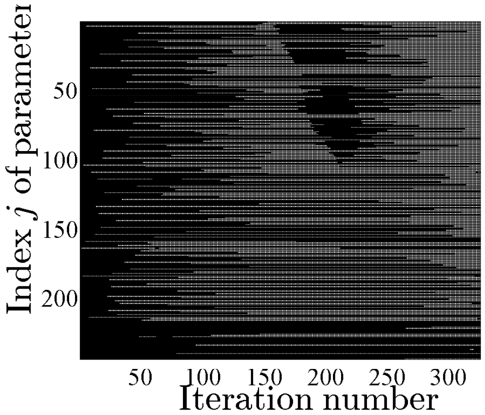
Def: *Deep neural network is a model of exceeding complexity. It ignores the universal approximation theorem (George Cybenko 1989, Kurt Hornik 1991).*

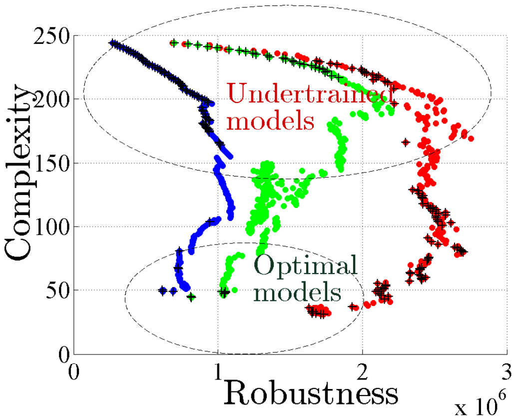
Neural network optimal brain damage procedure



Salience function $L_j = \frac{w_j^2}{2\mathbf{H}_{jj}^{-1}}$ versus number of removed parameters

Consequent model generation

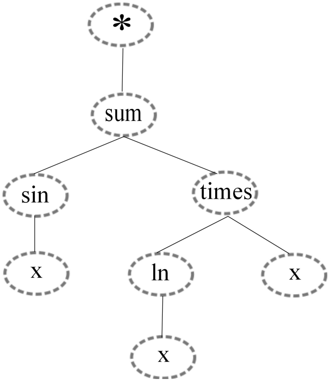




Popova, Strijov. 2015. Selection of optimal physical activity classification model // Informatics and Applications

Let the universal model be a mixture of superpositions of primitives

The tree Γ_f corresponds to some superposition $f \in \mathfrak{F}$



$f = \sin(x) + (\ln x)x$

Construct a superposition f

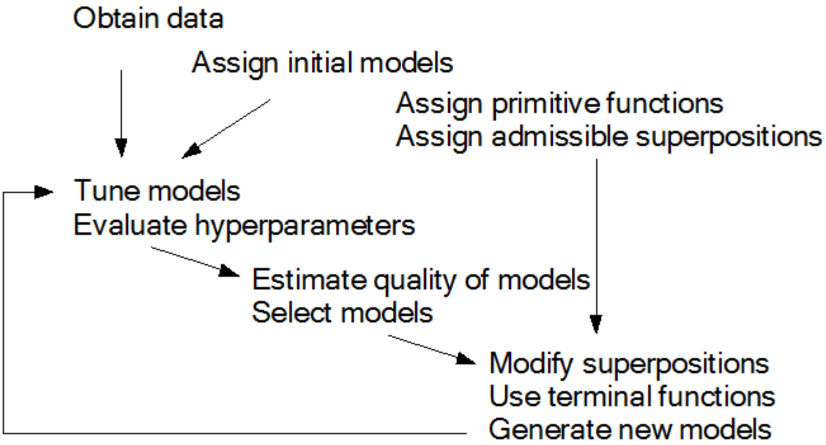
1) primitive functions $\mathfrak{G} \ni g : (\mathbf{w}', \mathbf{x}') \mapsto \mathbf{x}''$,
2) generation rules Gen and simplification rules Rem,
3) an admissible superposition is $\mathrm{cod}(g_{k+1}) \subseteq \mathrm{dom}(g_k)$, for any k.

A model is the superposition $f(\mathbf{w}, \mathbf{x}) = (g_1 \circ \cdots \circ g_K)(\mathbf{w})(\mathbf{x})$.

Construct a tree Γ_f

1) the root $*$ of the tree Γ_f has the single vertex,
2) other vertices V_i correspond to the functions $g_r \in \mathfrak{G}$: $V_i \mapsto g_r$,
3) the leaves Γ_f correspond to elements of the vector $\mathbf{x}$.

Consequent model generation

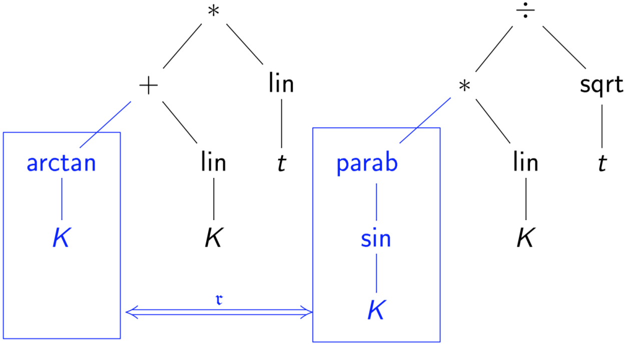


Add-delete strategy modifies a model to select it from a class, it searches around the maximum model evidence.

Genetic optimization constructs symbolic regression model structure

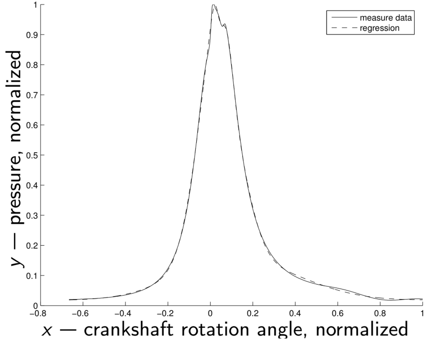
To create a model as a superposition of primitive functions

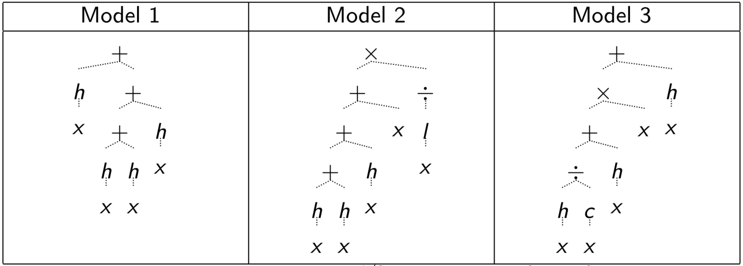
1) exchange random sub-trees between two models,
2) replace a random primitive for another one,
3) select the best models and repeat.



Simple superposition has 14 parameters versus 2-NN has 64 parameters

Approximate the pressure in the combusting camera of a diesel engine





Legend: h — gaussian $y = \lambda(2\pi\sigma^{-1/2})\exp(-(x-\xi)^2(2\sigma^{-2}) + a)$,
c — cubic $y = ax^3 + bx^2 + cx + d$, l — linear $y = ax + b$.

$$f_2 = g_1(g_2(g_3(g_4(g_5(x), g_6(x)), g_7(x)), x), g_8(x)).$$

The full representation of the Model 2 is

$$y = (ax+b)^{-1}\left(x + \sum_{i=1}^{3} \frac{\lambda_i}{\sqrt{2\pi}\sigma_i}\exp\left(-\frac{(x-\xi_i)^2}{2\sigma_i^2}\right) + a_i\right).$$

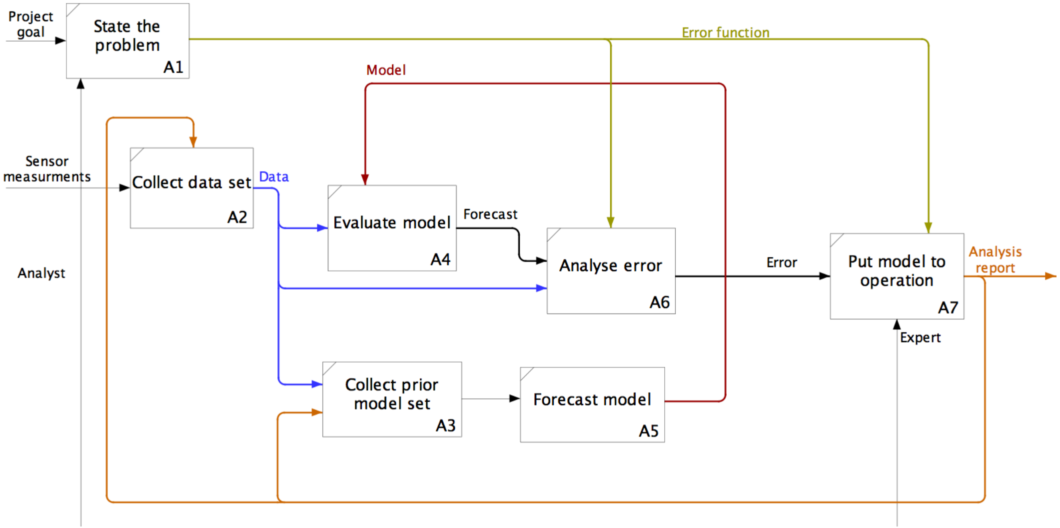
TREC text document collection has 2M documents times 200K requests

f_1	$e\sqrt{\ln(x/y)}$	h_1	$g\left(\frac{g(x)}{\sqrt{\ln(x)+x}}\right)-\ln(y)$
f_2	$\sqrt{\frac{\ln(x)}{\sqrt{y}}}$	h_2	$g\left(\frac{g(x)}{\sqrt{\frac{1}{2}\ln(x)+x}}\right)-\ln(y)$
f_3	$\sqrt[4]{\frac{x}{y}}$	h_3	$g\left(\ln\left(\frac{g(x)}{\sqrt{\frac{1}{2}\ln(x)+x}}\right)-\ln(y)\right)$
f_4	$\sqrt{y+\sqrt{\frac{x}{y}}}$	h_4	$g\left(\frac{g(x)}{\sqrt{g(\sqrt{x})+x}}\right)-\ln(y)$
f_5	$\sqrt{\sqrt{\frac{x}{y}}\cdot e^{-y}}$	h_5	$g\left(\frac{g(x)}{\sqrt{\ln(x)+\ln(y)}}\right)-\ln(y)$
f_6	$\sqrt{\sqrt{x}+\sqrt{\frac{x}{y}}}$	h_6	$g\left(\frac{g(\ln(x))}{\sqrt{\ln(x)+x}}\right)-\ln(y)$

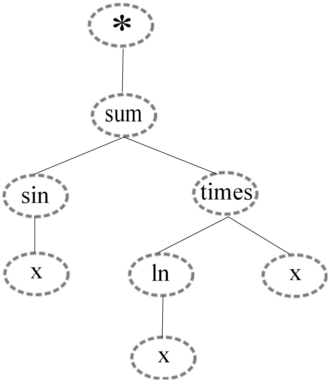
The information retrieval rank models with quality of Mean Average Precision $= 14.03$ for TREC-8 by the USA National Institute of Standards and Technology.

Kulunchakov, Strijov. 2017. Generation of simple structured Information Retrieval functions by genetic algorithm without stagnation // Expert Systems with Applications

One model to forecast models



Link matrix $\mathbf{Z}_f$ estimation limitations



$f = \sin(x) + (\ln x)x$

The link matrix $\mathbf{Z}_f$ for the tree Γ_f

	sum	times	ln	sin	x
$*$	1	0	0	0	0
sum	0	1	1	0	0
times	0	0	0	1	1
ln	0	0	0	0	1
sin	0	0	0	0	1

The link probability matrix $\mathbf{P}_f$ for the tree Γ_f

	sum	times	ln	sin	x
$*$	0.7	0.1	0.1	0.1	0.2
sum	0.2	0.7	0.8	0.1	0.2
times	0.1	0.3	0	0.8	0.8
ln	0.2	0.1	0.3	0.1	0.9
sin	0.1	0.2	0.1	0	0.8

$\mathfrak{Z}$ is a set of matrices corresponding to the superpositions from $\mathfrak{F}$.

There is given a sample $\mathfrak{D} = \{(\mathbf{D}_k, f_k)\}$ where the element $\mathbf{D}_k = (\underset{m\times n}{\mathbf{X}}, \underset{m\times 1}{\mathbf{y}})$, there given $\mathfrak{G}$ and $\mathfrak{F} = \{f_s \mid \mathbf{f}_s : (\hat{\mathbf{w}}_k, \mathbf{X}) \mapsto \mathbf{y}, s \in \mathbb{N}\}$.

The goal

to find an algorithm $a : \mathbf{D}_k \mapsto f_s$ following the condition

$$\mathbf{Z}_{f_s} = \arg\max_{\mathbf{Z}\in\mathfrak{Z}} \sum_{i,j} P_{ij} \times Z_{i,j}.$$

The index $\hat{s}$, что $f_{\hat{s}}$ provides a minimum for the error function S:

$$\hat{s} = \arg\min_{s\in\{1,\dots,|\mathfrak{F}|\}} S(f_s \mid \hat{\mathbf{w}}_k, \mathbf{D}_k),$$

where $\hat{\mathbf{w}}_k$ is an optimal vector of parameters f_s for each $f_s \in \mathfrak{F}$ with the fixed $\mathbf{D}_k$:

$$\hat{\mathbf{w}}_k = \arg\min_{\mathbf{w}\in\mathbb{W}_s} S(\mathbf{w} \mid f_s, \mathbf{D}_k).$$

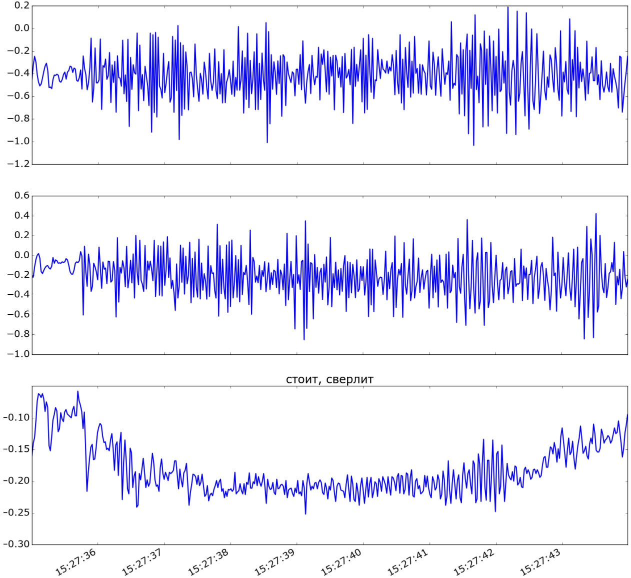
Complex action: workers construct a rack (Forecsys.ru, behavioral analysis)

Complex movement: the worker is drilling while standing

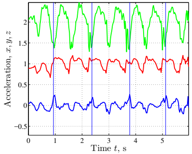
x

y

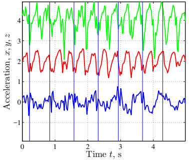
z



Acceleration time series $[x_t, y_t, z_t]^\mathsf{T}$.



Slow walking



Jogging

Time series samples for physical activity monitoring

Instances

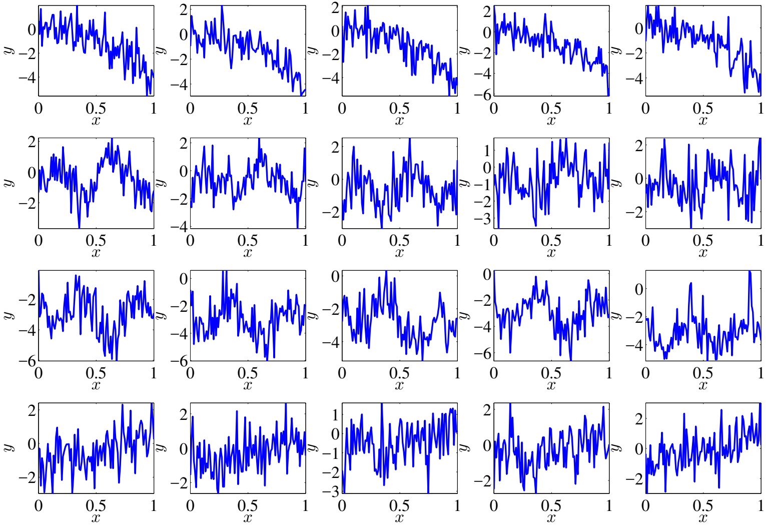
Classes

One

Two

Three

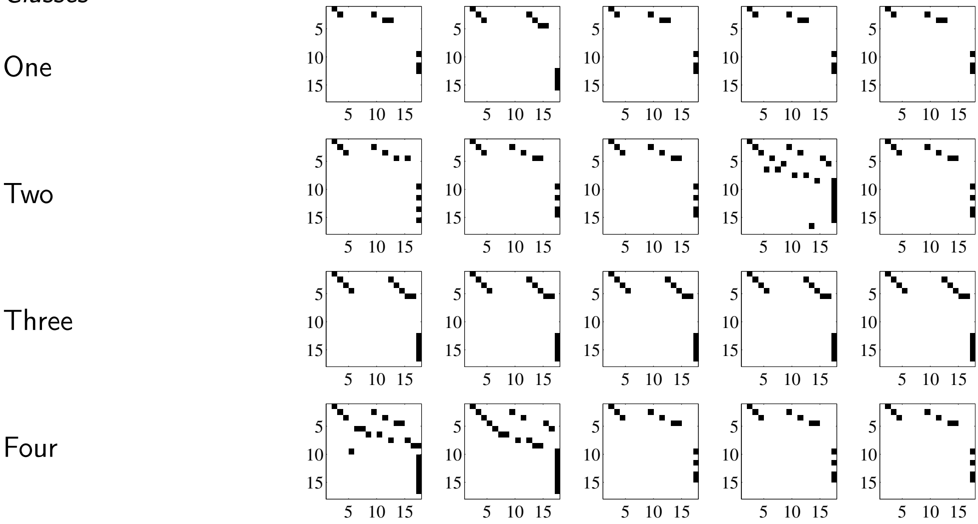
Four



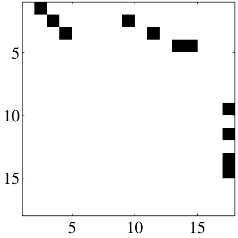
Time series samples for physical activity monitoring

Instances

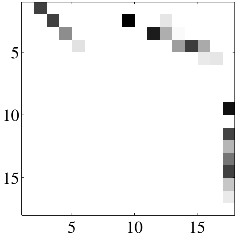
Classes



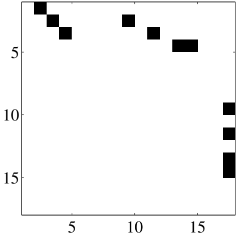
The initial and the forecasted superposition



Ground truth



Forecasted probabilities

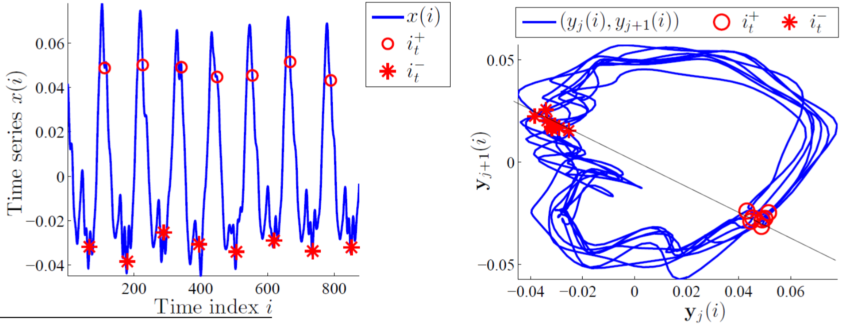


Forecasted superposition tree (model)

Human gait detection with time series segmentation

Find dissection of the trajectory of principal components $\mathbf{y}_j = \mathbf{H}\mathbf{v}_j$, where $\mathbf{H}$ is the Hankel matrix and $\mathbf{v}_j$ are its eigenvectors:

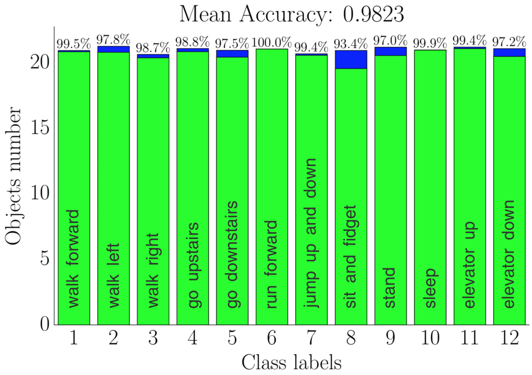
$$\frac{1}{N}\mathbf{H}^{\mathsf{T}}\mathbf{H} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\mathsf{T}}, \quad \Lambda = \text{diag}(\lambda_1, \dots, \lambda_N).$$



Motrenko, Strijov. 2016. Extracting fundamental periods to segment human motion time series // IEEE Journal of Biomedical and Health Informatics

Replace universal models for interpretable superposition: NN $\to$ SSA+LgR

Neural network replaced by Singular Structure Analysis + Linear regression boosts quality and puts the model into a wristwatch.

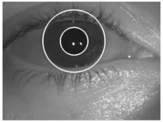


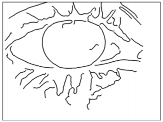
Performance of the human physical activities classification

Ignatov, Strijov. 2015. Human activity recognition // Multimedia Tools and Applications

Discover the iris by linear mixture (possible example)

Replace a proprietary algorithm or CNN for mixture of linear models to drop the computational complexity.

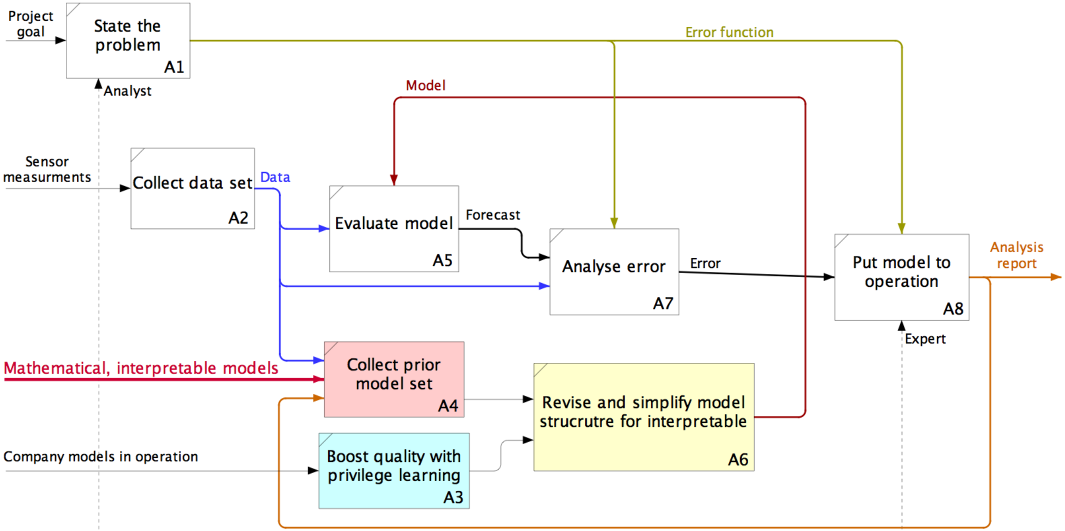




Example of interpretable modelling

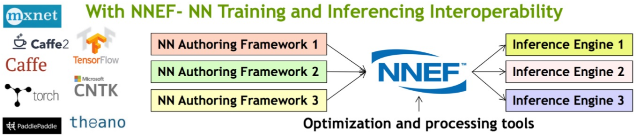
Chigrinsky. 2017. Modeling of the iris movement by optical flow//BS Thesis, adv. by Matveev

Put interpretable models to operation along with privilege learning models



List of the model generation paradigms

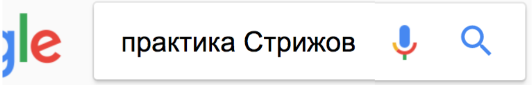
1. Binary/continuous/graph optimization of model structures
2. Neural networks forecast hyperparameters of neural networks (ref. NIPS 2017)
3. Networks forecast superpositions
4. Interpretable models replace neural network blocks
5. Company models boost quality of neighbor models by privilege learning



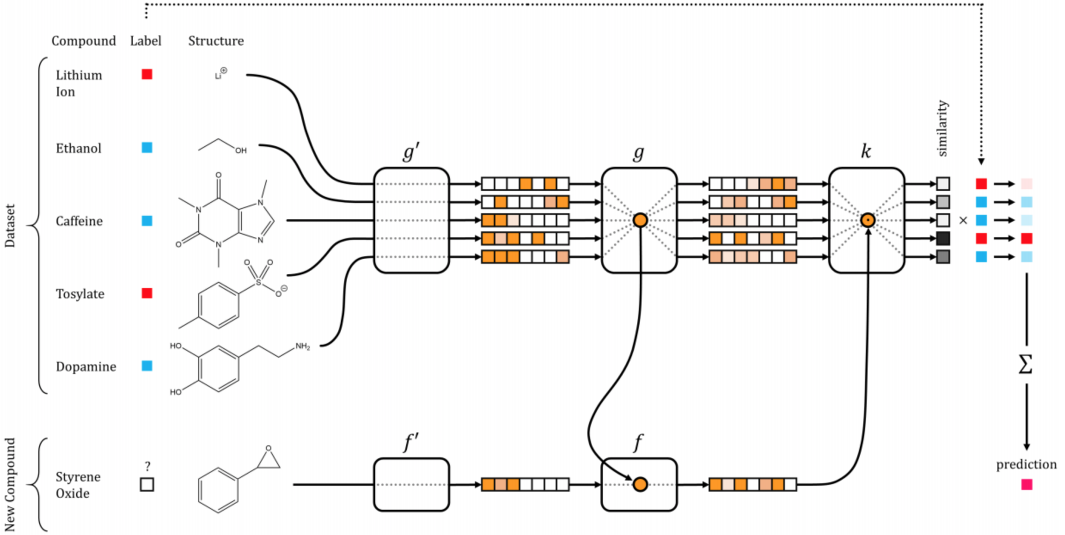
Our research challenges

1. Lay the foundations for the forecasting of model structures
2. Develop the theory of local modeling for signals of wearable devices
3. Deploy standards to exchange local and universal models

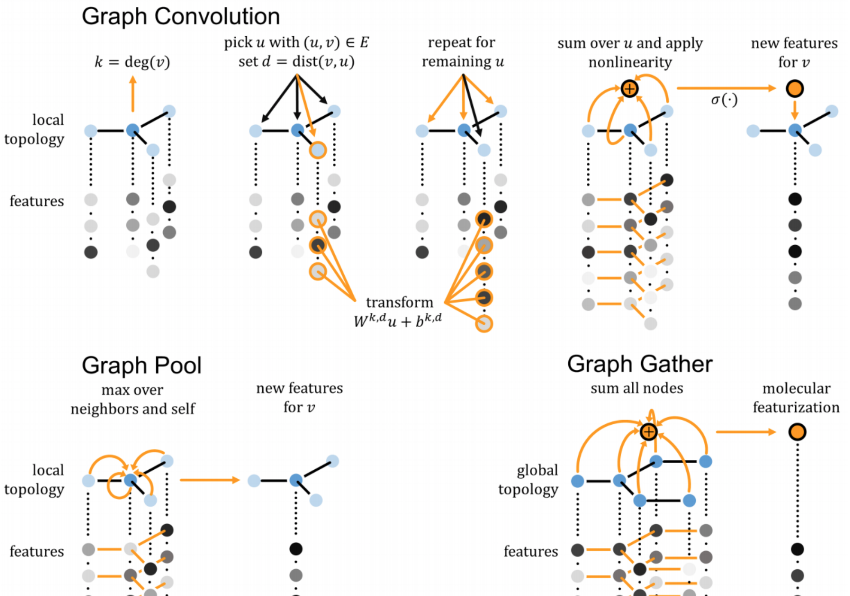
30+ projects start 14.2.18. with 60 analysts, experts and MIPT students:



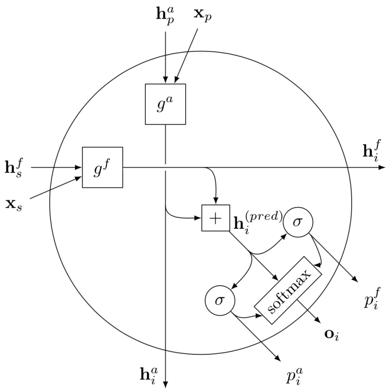
Selected papers to discuss



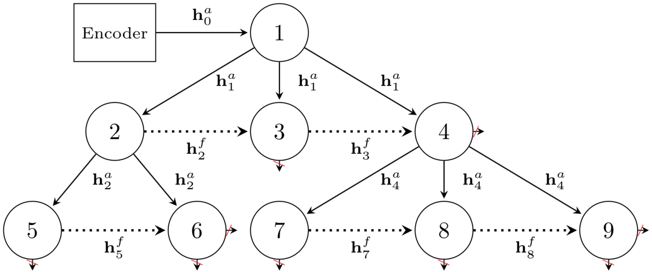
Compound
Label
Structure
Dataset
Lithium Ion
Ethanol
Caffeine
Tosylate
Dopamine
New Compound
Styrene Oxide
?
g′
g
k
f′
f
similarity
×
Σ
prediction



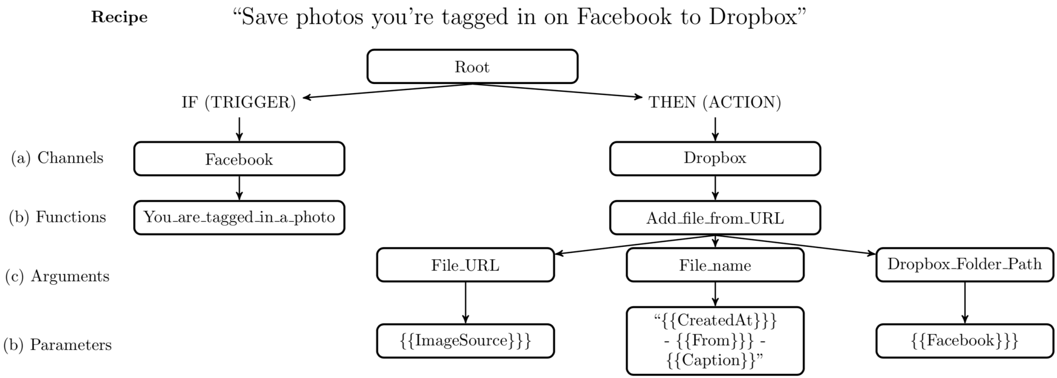
Graph Convolution
$k = \deg(v)$
pick u with $(u, v) \in E$
set $d = \mathrm{dist}(v, u)$
repeat for
remaining u
sum over u and apply
nonlinearity
new features
for v
$\sigma(\cdot)$
local
topology
features
transform
$W^{k,d}u + b^{k,d}$
Graph Pool
max over
neighbors and self
new features
for v
local
topology
features
Graph Gather
sum all nodes
molecular
featurization
global
topology
features



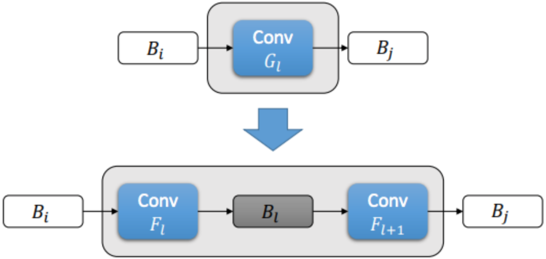
A cell of the doubly-recurrent neural network corresponding to node i with parent p and sibling s.



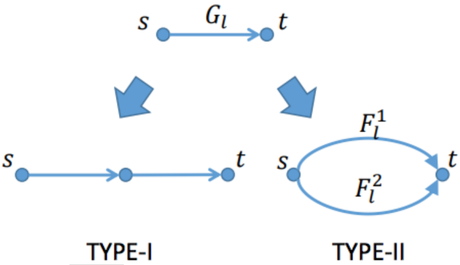
Structure-unrolled DRNN network in an encoder-decoder setting. The nodes are labeled in the order in which they are generated. Solid (dashed) lines indicate ancestral (fraternal) connections. Crossed arrows indicate production halted by the topology modules.



Example recipe from the IFTTT dataset. The description (above) is a user-generated natural language explanation of the if-this-then-that program (below).

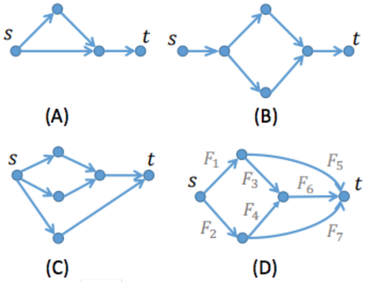


Network morphism in depth.

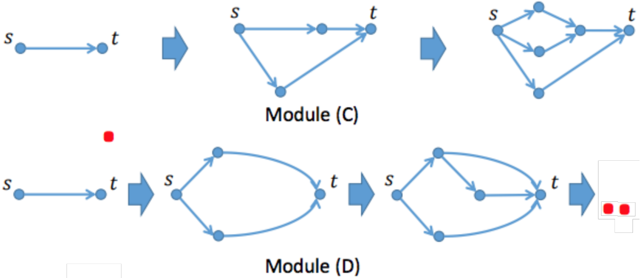


Atomic morphing types.

Illustration of atomic morphing types. (a) One convolutional layer is morphed into two convolutional layers; (b) TYPE-I and TYPE-II atomic morphing types.



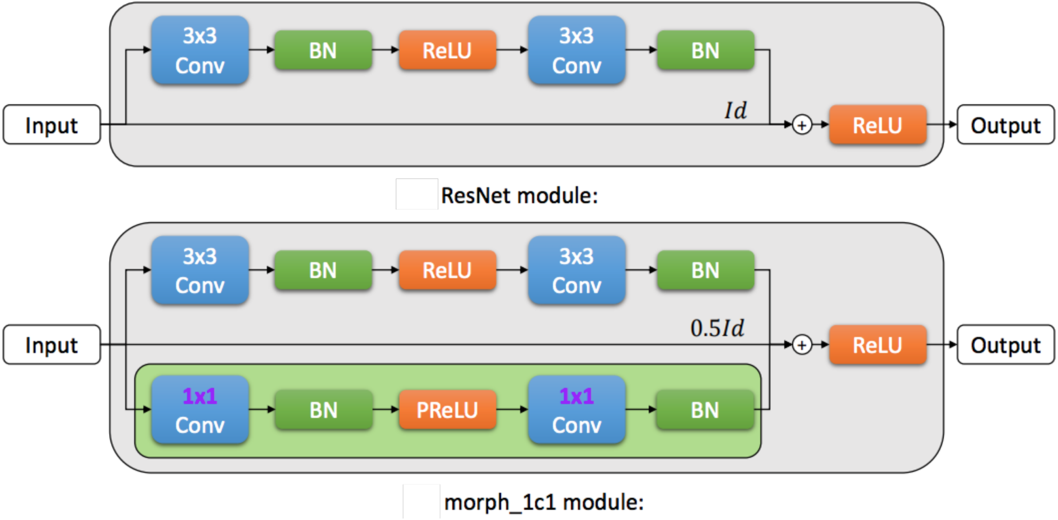
Example modules.



Morphing process for module (C) and (D).

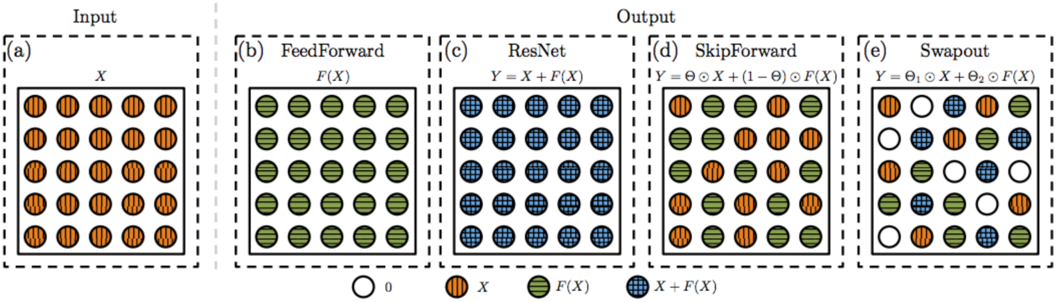
Example modules and morphing processes. (a) Modules (A)-(C) are simple morphable, while (D) is not; (b) a morphing process for module (C), while for module (D), we are not able to find such a process.

Tao Wei. 2017. Modularised morphing of neural networks //ICLR



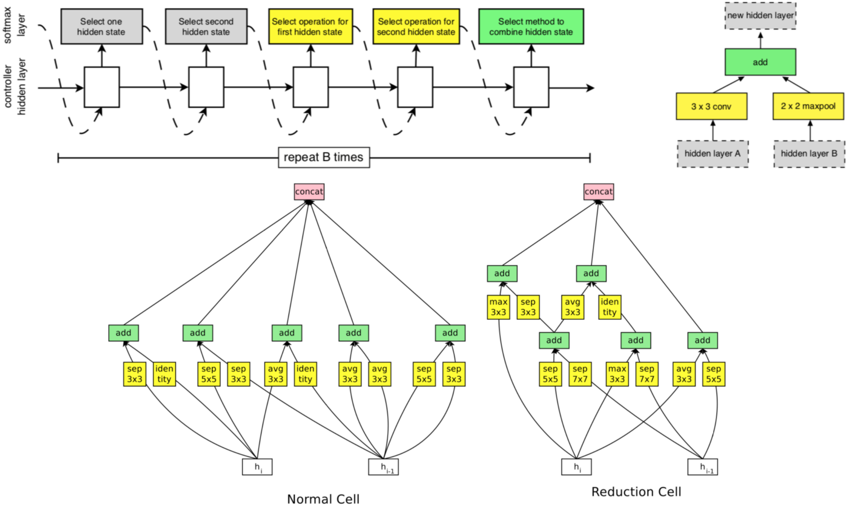
Detailed architectures of the ResNet module and the morph 1c1 module

Saurabh Singh et al. 2016. Swapout: Learning an ensemble of deep architectures // NIPS

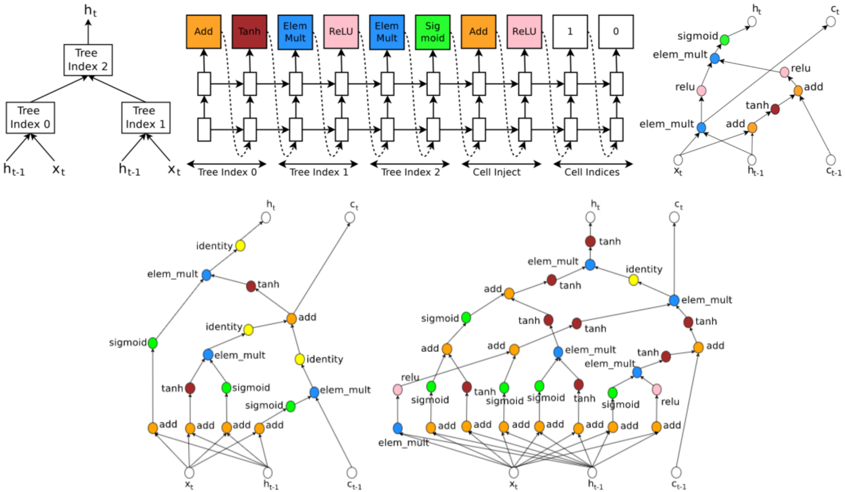


Visualization of architectural differences, showing computations for a block using various architectures. Each circle is a unit in a grid corresponding to spatial layout, and circles are colored to indicate what they report. Given input X (a), all units in a feed forward block emit F(X) (b). All units in a residual network block emit X + F(X) (c). A skip-forward network randomly chooses between reporting X and F(X) per unit (d). Finally, swapout randomly chooses between reporting 0 (and so dropping out the unit),

Zoph B. et al. 2016. Neural architecture search with reinforcement learning // arXiv preprint arXiv:1611.01578



Zoph B. et al. 2017. Learning Transferable Architectures for Scalable Image Recognition // arXiv preprint arXiv:1707.07012



Bello Irwan et al. 2017. Neural optimizer search with reinforcement learning // arXiv:1709.07417