Explaining Deep Learning Predictions and Integrating Domain Ontologies

Isaac Ahern

July 16, 2018

Outline

CBL project problem

Human behavior prediction problems, and the issue of explaining deep learning predictions.

- Deep Learning predictions shouldn't be treated as a 'black box' — want to explain classifiers.
- Avoid fitting bias induced from learning 'flat models', using domain ontologies to structure models.

CBL project domains

- PeaceHealth (Electronic Health Records)
- Eli Lilly (Drug Information)
- Baidu (Social Media)

PeaceHealth

- Nonprofit Health Care Network
- Predicting Health outcomes and recurrences: incorporate explicit & implicit social and environmental factors and self motivation into DL model

Eli Lilly

- Global Pharmaceutical Company
- Understanding healthcare outcome relationships between patients and products

Baidu

- Search engine/internet company — "the Chinese Google"
- Incorporate social media user data for human behavior prediction

LIME

Goal: provide an "explanation" for any given classifier, i.e., provide some characteristic which illustrates qualitative understanding of the relationship between an instance in the data, and the corresponding model prediction.

Model Accuracy vs Explanation

It may be desirable to choose a less accurate model for content recommendations based on the importance afforded to different features (e.g., predictions related to 'clickbait' articles which may hurt user retention).

- metrics we can optimize: accuracy
- metrics we might actually care about: user engagement, retention

In this case, it is important to have a heuristic for explaining *how* a model is making predictions, along with the actual predictions themselves.

LIME

An algorithm that can explain model predictions such that:

- explanations are *locally-faithful* to the model.
- explainations are *interpretable*.
- explanations are *model-agnostic*.
- can be extended to a measure of a model's trustworthiness — i.e., extended to *explain the model*.

LIME

- **L**ocal
- **I**nterpretable
- **M**odel-Agnostic
- **E**xplanations

Interpretable vs Features

- text classification:
  - interpretable explanation — binary vector indicating presence/absence of a word.
  - feature — word embedding (i.e. W2V Skipgram).
- image classification:
  - interpretable explanation — binary vector indicating presence/absence of *super-pixels*: contiguous patches of "similar" pixels.
  - feature — representation of image as tensor via ConvNet with 3 color channels / pixel.
- $x \in \mathbb{R}^d$ the representation of an instance $\leftrightsquigarrow$ $x' \in \{0,1\}^{d'}$ is a corresponding interpretable representation.

Optimization Criteria

- $\Omega(g)$ = complexity of the model g.
- $f : \mathbb{R}^d \to \mathbb{R}$ a model, i.e. $f(x)$ is the probability that x belongs to a certain class.
- $\mathcal{L}(f, g, \pi_x)$ = measure of the error in approximation of f by g in the region defined by π_x (locality-aware loss).

Then, the LIME model balances the constraints of interpretability and faithfulness by selecting (locally)

$$\xi(x) = \text{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

where G is a class of potentially interpretable models, such as linear models, decision trees, or falling rule lists

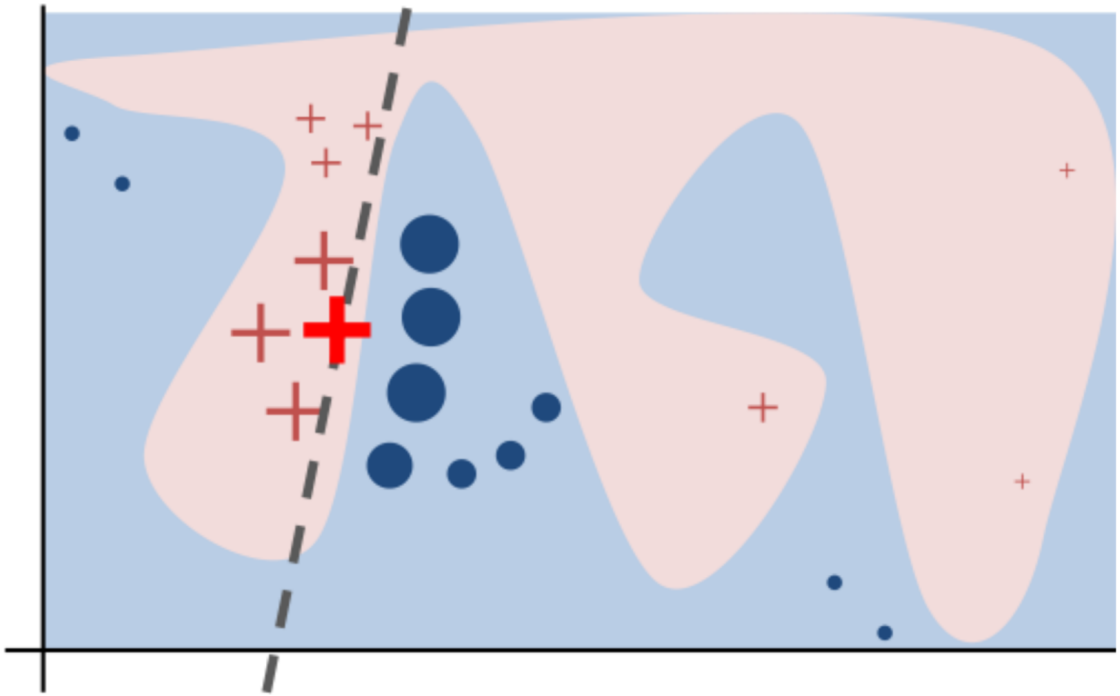
Sampling

To perform minimization as defined by $\xi(x)$, "sample uniformly from instances around x'", weighted according to π_x. This recovers points $z \in \mathbb{R}^d$ to which we apply the label $f(z)$ (model prediction), yielding the dataset $\mathcal{Z} = \{(z, f(z))\}_{\text{sampled } z}$. We then optimize model's $\text{argmin}_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$ for $\mathcal{Z}$.

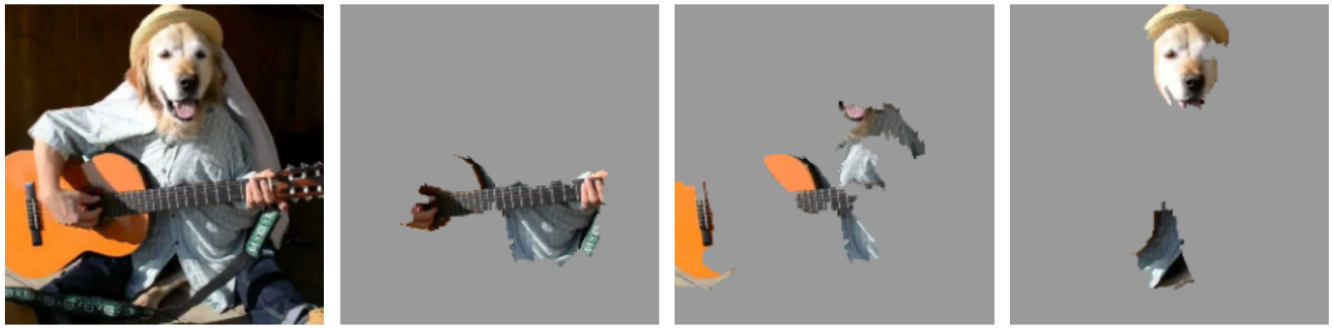
Sparse Linear Explanations

- model class: $G = \{g : z' \mapsto w_g \cdot z'\}$.
- locality distribution: $\pi_x(z) = e^{-D(x,z)^2/\sigma^2}$ (gaussian / exponential kernel) for a domain-appropriate distance measure D. (i.e., cos, L_2, etc.)
- $\mathcal{L}(f, g, \pi_x) = \sum_{z,z' \in \mathcal{Z}} \pi_x(z)(f(z) - g(z'))^2$.

Example: Intuition



Example: ConvNet



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Figure 4: Explaining an image classification prediction made by Google's Inception neural network. The top 3 classes predicted are "Electric Guitar" ($p = 0.32$), "Acoustic guitar" ($p = 0.24$) and "Labrador" ($p = 0.21$)

Explaining Models: SP-LIME

Extend LIME so as to give a global understanding of the model, by explaining a set of individual instances. Problem is to select a set of instances which is simultaneously feasible to inspect and gives non-redundant explanations that represent the model's global behavior.

Given X instances, construct an $|X| \times d'$ *explanation matrix* $W = (|w_{g_{ij}}|)$, where $g_i = \xi(x_i)$ is the LIME-selected interpretable local sparse-linear model approximation.

Interpretation of W

- W represents the local importance of the interpretable components at each instance.
- If $e_j = (0, \ldots, 1, \ldots, 0)^T$, then $I_j := \psi(W_{ij}^T e_j)$ gives a measure of the global importance of component j, where domain-dependent ψ controls weight assigned to column j. Ex (text):

$$I_j = \sqrt{\sum_{i=1}^{|X|} W_{ij}}$$

- Coverage $c_{W,I}(V) = \sum_{j=1}^{d'} \mathbb{1}_{\{i \in V : W_{ij} > 0\}} I_j$ weights each column measure I_j by the number of rows with non trivial weights in column j, giving the total importance of the features that appear in at least one instance in a set V.

Example: Picking from W

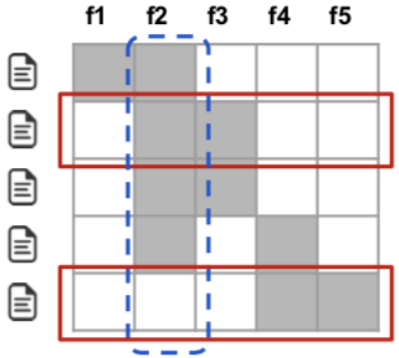


Figure 5: Toy example $\mathcal{W}$. Rows represent instances (documents) and columns represent features (words). Feature f2 (dotted blue) has the highest importance. Rows 2 and 5 (in red) would be selected by the pick procedure, covering all but feature f1.

Pick Step

Toy example (previous slide): If all weights are the same and $V =$ rows 2 & 5, then:

$$c_{W,I}(V) = \sqrt{W_{12} + W_{22} + W_{32} + W_{42}} + \sqrt{W_{23} + W_{33}} + \sqrt{W_{44} + W_{54}} + \sqrt{W_{55}}$$

Given a maximum budget of B inspections, then, the goal is to determine

$$Pick(W, I) = \text{argmax}_{|V| \leq B} c_{W,I}(V)$$

which maximizes coverage $c_{W,I}(V)$ under the restriction $|V| \leq B$.

Pick Step (cont.)

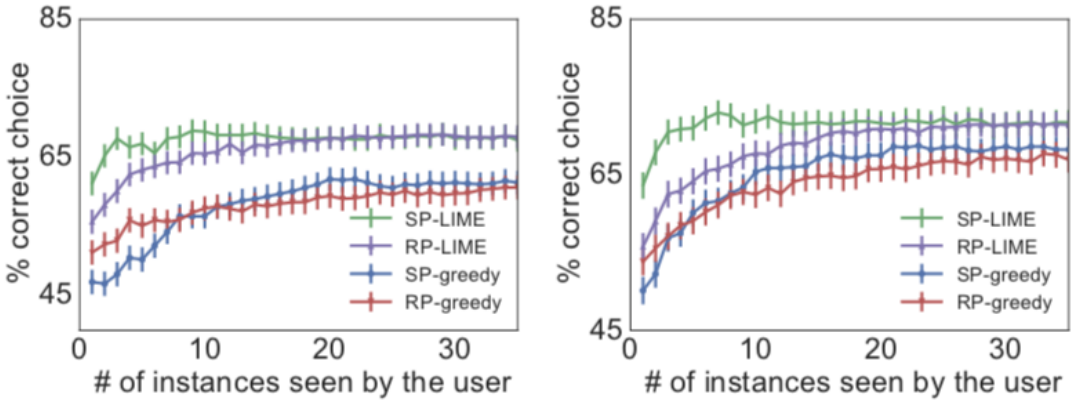
Algorithm: Determining $Pick(W, I)$ is maximizing a weighted coverage function, and is NP-hard. Furthermore, c is submodular, so a greedy algorithm iteratively acting to maximize marginal coverage gain $c_{W,I}(V \cup \{k\}) - c_{W,I}(V)$ offers a constant factor approximation to the optimal coverage.

Results

Test via:

- label a proportion of certain features as "untrustworthy".
- develop oracle "trustworthiness" labeling test set predictions from a black box classifier as "untrustworthy" if the prediction changes when untrustworthy features are removed from the instance, "trustworthy" otherwise.

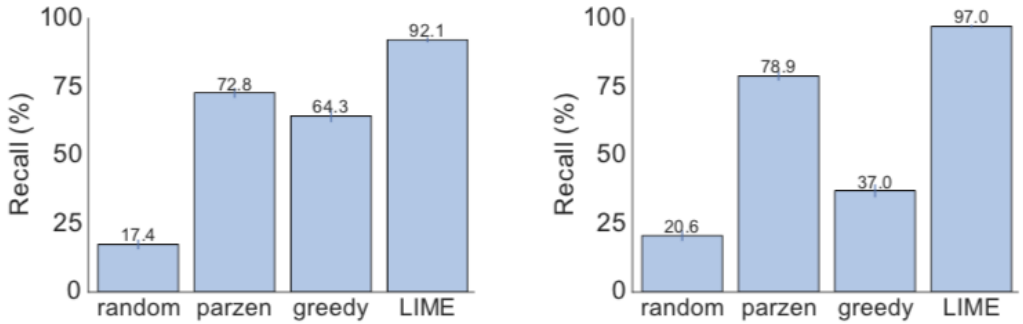
Results (cont.)



(a) Books dataset (b) DVDs dataset

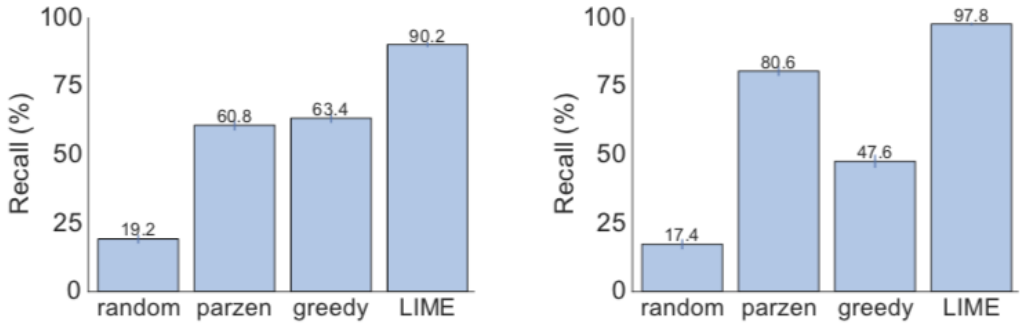
Figure 8: Choosing between two classifiers, as the number of instances shown to a simulated user is varied. Averages and standard errors from 800 runs.

Results (cont.)



(a) Sparse LR (b) Decision Tree

Figure 6: Recall on truly important features for two interpretable classifiers on the books dataset.

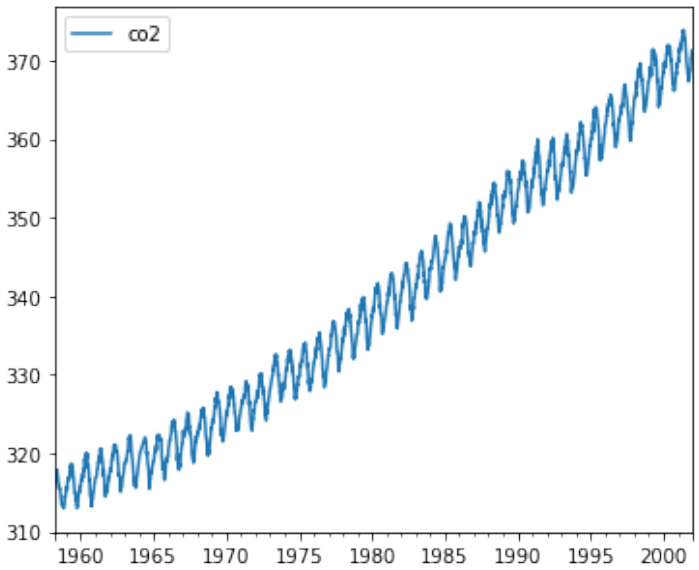


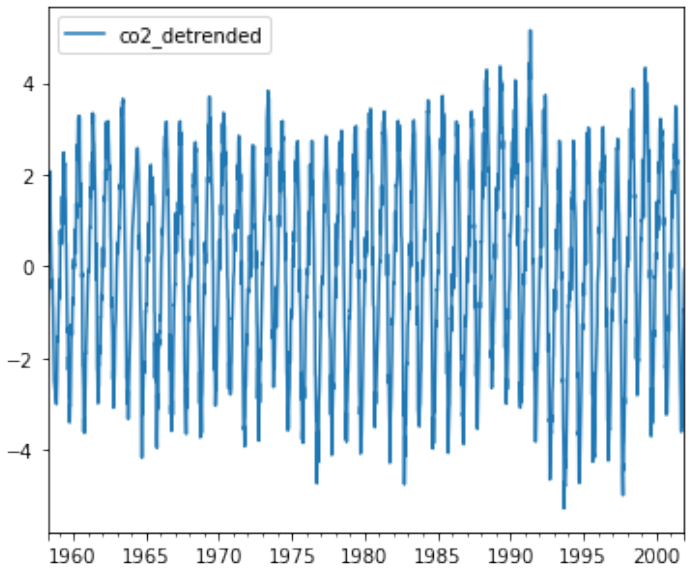
(a) Sparse LR (b) Decision Tree

Figure 7: Recall on truly important features for two interpretable classifiers on the DVDs dataset.

LIME with RNN

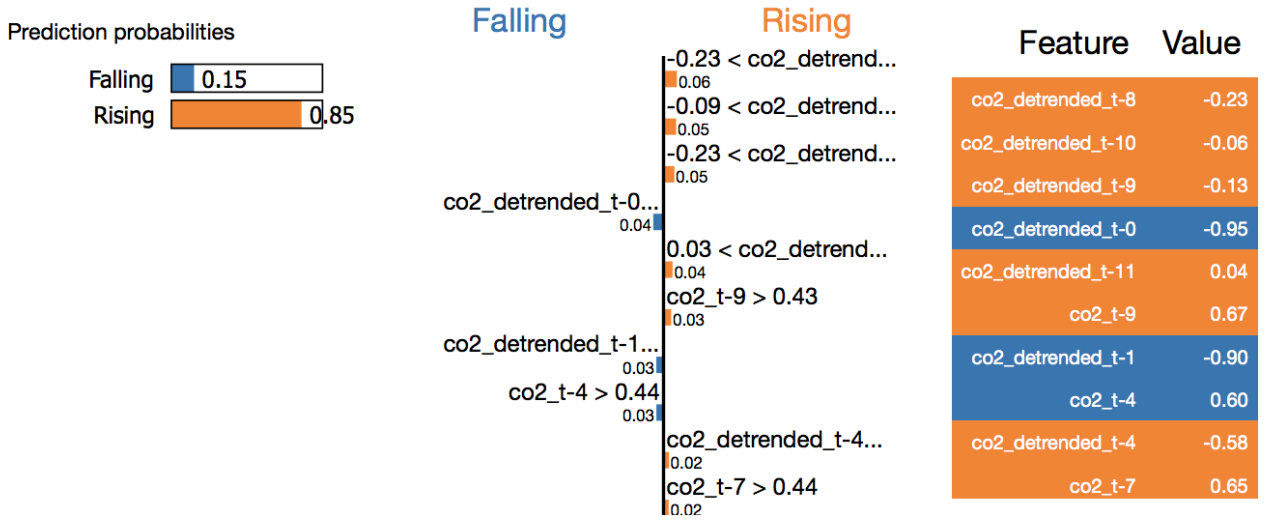
CO2 data:





LIME with RNN (cont.)

Explanation:



We can see that the most important features are the de-trended CO_2 concentration several timesteps in the past. In particular, we see that if that feature is low in the recent past, then the concentration is now probably rising.

Ontological Deep Learning

A network ontology consists of a set of concepts, sub-concepts, and relations between concepts. Each concept can contain sub-concepts as well as characteristics:

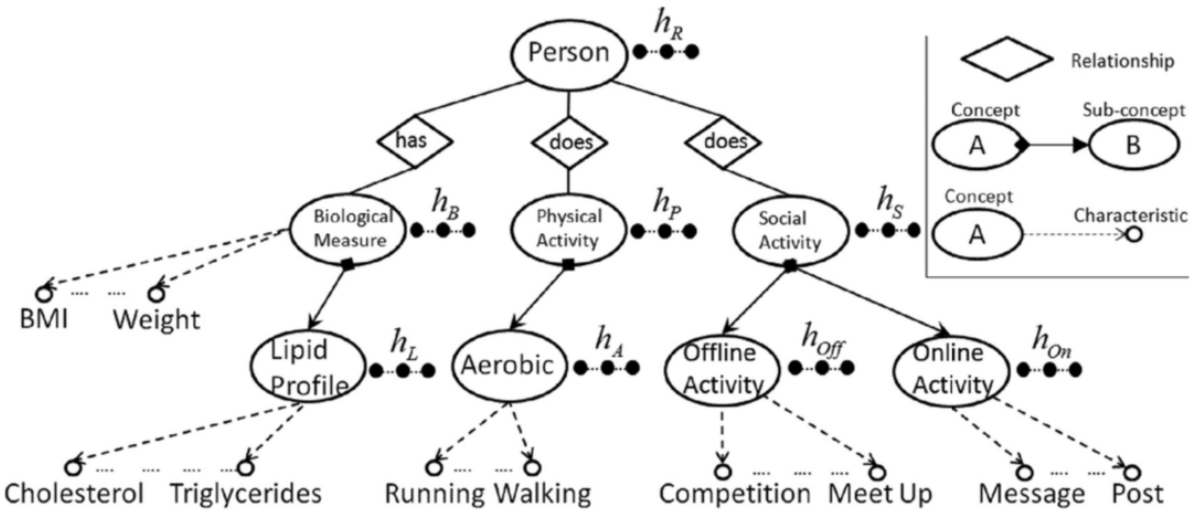


Fig. 2. Partial view of the SMASH ontology and its hidden variables.

Ontology Representation

ORBM$^+$ model extends an RBM with temporal dependencies (connections between hidden, visible layers to historical variables) by first learning a representation of the concepts and relationships in a given ontology H — represent concepts $\mathbb{S} \in H$ by a set of learnable hidden layers $\mathbf{h}_{\mathbb{S}}$: For a concept $\mathbb{S} \in H$,

- $\Psi_{\mathbb{S}} := \bigcup_{F \in F_{\mathbb{S}}} V_F$ is the union of all characteristics from the relationships $F_{\mathbb{S}}$ of $\mathbb{S}$.
- $\Theta_{\mathbb{S}} := \bigcup_{C \in C_{\mathbb{S}}} \mathbf{h}_C$ is the union of all hidden variables from the various sub-concepts $C_{\mathbb{S}}$.

Then, all the variables $v_i \in V_{\mathbb{S}} \cup \Psi_{\mathbb{S}} \cup \Theta_{\mathbb{S}}$ are considered as a visible layer in an RBM, and the $\mathbf{h}_{\mathbb{S}}$ is considered as a hidden layer.

Probabilities

This model has conditional probabilities

$$p(h_j|V_{\mathbb{S}}, C\mathbb{S}, F_{\mathbb{S}}) = \mathcal{N}(b_j + \sum_{v_i \in V_{\mathbb{S}} \cup \Psi_{\mathbb{S}} \cup \Theta_{\mathbb{S}}} v_i W_{ij})$$

$$p(v_i|\mathbf{h}_{\mathbb{S}}) = \mathcal{N}(a_i + \sum_{h_j \in \mathbf{h}_{\mathbb{S}}} h_j W_{ij})$$

which are used to compute the energy function associated with $\mathbb{S}$, for training the model.

ORBM$^+$ Model

To learn representations $\mathbf{h}_{\mathbb{S}}$ of "higher order" concepts $\mathbb{S}$, first learn the representations $\mathbf{h}_{C_{\mathbb{S}}}$ of the related sub-concepts.

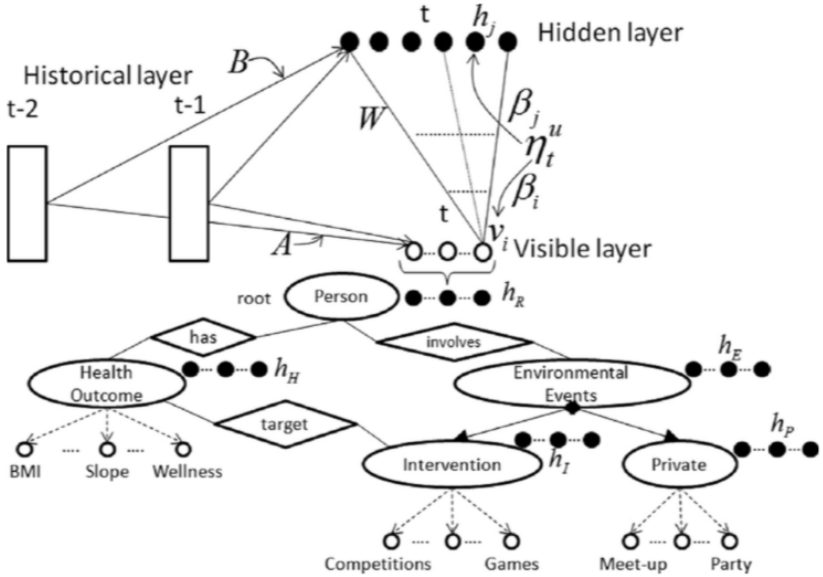


Fig. 3. The ORBM$^+$ model.

ORBM$^+$ Model (cont.)

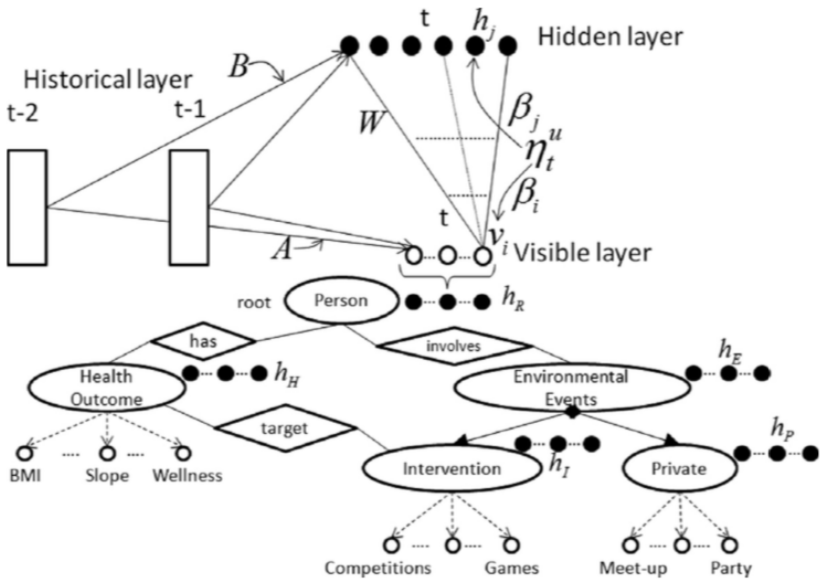


Fig. 3. The ORBM$^+$ model.

- W — used to capture the 'self-motivation factor'
- A, B — used to capture the correlation between past and present states.
- η_t^u — temporal smoothing used to better capture explicit social influences on user u at time t

Explanation Generation

- Model initially used for users in a health program: Explanations increase transparency of the intervention process, contribute to users' satisfaction, and are facilitate in user engagement.
- Explanation as a list of characteristics which maximize the likelihood of a behavior being engaged by a user (or set of users).
- $\log P(Y|X,\theta)$ maximized given the characteristics $X \implies X$ is the best explanation for Y.
- X contains many characteristics in high dimensional data, making it non-transparent and uninterpretable for lay users. i.e., want to find an interpretable explanation

$$X* = \text{argmin}_{X' \subset X} f(\log P(Y|X',\theta), X').$$

Minimum Description Length

Use minimum description length to encode observed data Y via an explanation X', then encode X':

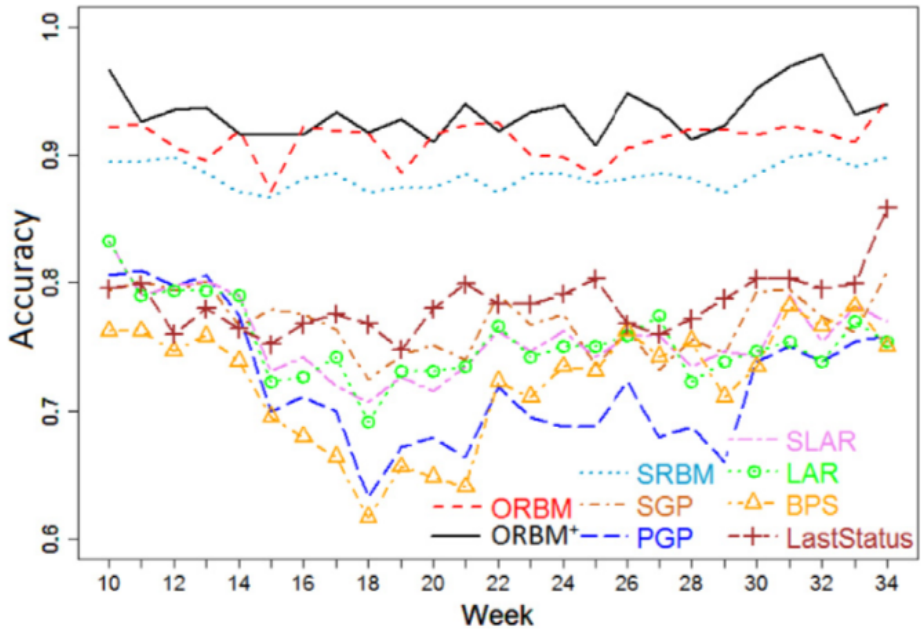
- $L(Y, X') =$ encoding length of Y given X'
- $L(X') =$ encoding length of X'
- MDL minimizes

$$L(Y, X') + L(X') = -\log P(Y|X', \theta) + |X'| \log(|X|)$$

over explanations X'.

- The complexity of explanation generation is NP-hard. Hence, apply a heuristic greedy min algorithm which adds new characteristics into the explanation so that selection model MDL is minimized (stepwise).

Results



(b) ORBM$^+$ vs competitive models in terms of prediction accuracy

References

1. "'Why Should I Trust You?' Explaining the Predictions of Any Classifier" (LIME) — Ribeiro, et al KDD 2016.
2. "Ontology-based deep learning for human behavior prediction with explanations in health social networks" — Phan, et al IS 2017