The Smart Data Layer

Magnus Sahlgren,[1*] Erik Ylipää,[1] Barry Brown,[2] Karey Helms,[3]
Airi Lampinen,[2] Donald McMillan,[2] Jussi Karlgren[3]
[1]RISE SICS, Kista, Sweden
[2]Stockholm university, DSV, Kista, Sweden
[3]KTH, Stockholm, Sweden
[*]Corresponding author: magnus.sahlgren@ri.se

Abstract

This paper introduces the notion of a *smart data layer* for the Internet of Everything. The smart data layer can be seen as an AI that learns a generic representation from heterogeneous data streams with the goal of understanding the state of the user. The smart data layer can be used both as materials for design processes and as the foundation for intelligent data processing.

IoT and Interaction

One of the more ominous visions of the future Internet of Everything (IoE) is a swarm of loosely integrated systems (e.g. the smart home, social media apps, health and fitness wearables, etc.) that constantly crave our attention with applications bombarding us with notifications and alerts, and devices demanding administration and care. Rather than improving quality of life and efficiency of work, such excessively attention-seeking technology will lead to cognitive overload, adding both stress and complexity to everyday life. The main problem, and risk factor for such a future technological dystopia, is that different forms of smart technology do not blend and cannot interface with one-another, and most importantly, end-users have to learn how to interact with each of the different systems, one by one. In some sense, this is like personal computing before the desktop metaphor, the Internet before the web, or mobile computing before touch interfaces. In short, Internet of Things (IoT) (and IoE) lacks an appropriate interface paradigm.

As one step towards a solution to this interface problem, we investigate the possibility of defining and applying a *smart data layer* that integrates heterogeneous data streams into a coherent representation that can serve as the foundation for further, intelligent, data processing. The idea is not to provide a uniform communication protocol between applications and devices, but to provide a representation of the *state of the user*, to enable more intelligent interface design. The problem we would like to mitigate is for applications and devices to know *when* and *how* to interact with the user. As a simple example, if the user is in a very agitated state, we probably should not send loud audible notifications that the milk in the refrigerator is almost finished and needs refilling, or for that matter send intense tactile vibrations indicating that the user has been stationary for too long and that it is now time to get up and move. Our vision is a data layer that *learns* from the user's behaviors, and that is empathetic to both the current state of the user and the current state of the system. This position statement describes our current research path, and provides some background and motivation for the smart data layer.

AI and Representation Learning

AI will be a critical component in the development of IoT and its various flavors, not only for making sense of the interconnected systems, but also – and equally important – for making sense of the user of the system. The ultimate goal is to *understand* the user; where is the user, what is the user doing, how is she feeling, what are her goals? In short, what is the *state* of the user? Note that we use the term "state" in a broad sense; it can encompass anything from a geographical location, to a task, to the emotional state of the user, to a prediction of the user's next action.

Solving individual tasks such as locating the user, classifying her behavior, or detecting her sentiment are interesting, and potentially useful, tasks in their own right, but they require an ontology to start from. We have to know which are the possible locations, behaviors and sentiments in order to determine which of them the user belongs to. Defining or acquiring such ontologies is typically a task-specific problem, as is the optimization of classifiers. We do not believe that we (at present) can design or learn one generic ontology and one generic classifier that can solve any problem. However, we do believe that we can learn one generic *representation* that can be common for all these problems. Ideally, this representation will capture the causal factors of variation in streaming data of different modalities and rates.

The idea of a generic representation that can be used for various different purposes is not novel in itself, see Bengio et al. (2013) for a review. A good representation simplifies tasks, and a desirable property of a representation is the separation of the causal factors that gives rise to a phenomena. Digital images are an example of representations that are difficult to use directly for solving computer vision tasks. The pixels in the two-dimensional grid explain very little of the scene that generated them. A representation that directly encodes the objects in the scene, their state and surroundings would make automatic decisions based on the scene much simpler. *Representation learning* can be thought of as a generalization of this inverted rendering process, where we infer what the causal factors were that generated the data using methods from statistical learning. Popular methods include latent variable models, deep neural networks and compressed sensing.

Several researchers have published papers in this research direction; one example is Collobert et al. (2011), who propose a unified neural network architecture that can be applied to natural language processing and learns shared representations of language useful for solving a variety of tasks. The field of deep learning to a large extent embodies the idea that it is possible to learn a compositional, generic representation that can be used to solve many different problems; in image recognition for example, it has become customary to use the unit activations of deep neural networks trained on very large datasets (such as AlexNet (Krizhevsky, Sutskever, and Hinton 2012) or ResNet (He et al. 2016)) as the basic representation when building novel classifiers. This method of *transfer learning* is useful where representations learned on large data sets can be used to solve related tasks where data is scarce, see Oquab et al. (2014).

Another recent example of representation learning is the StarSpace framework of Wu et al. (2017), which is a general-purpose neural network representation that can solve a wide variety of problems.

(Word) Embeddings as a Starting Point

Our vision of the smart data layer builds on the prior art discussed in the previous section, and is inspired by the development of *word embeddings* for natural language processing (Turney and Pantel 2010). Embeddings are low-dimensional representations that compress and encode co-occurrence information from the input data. A co-occurrence event is simply the simultaneous occurrence of two (or more) variables. In language data, the variables are typically words, and a co-occurrence is simply a sequence of words. The point of embedding co-occurrence information in a low-dimensional representation is that the resulting representation generalizes from the observed co-occurrence events, and enables quantification of *distributional* (as in a word's distribution over the data) similarity. Since distributional similarity is a proxy for semantic similarity, embedding models can be seen as computational models of meaning (Sahlgren 2006).

Word embeddings have become ubiquitous in both natural language processing and machine learning. However, current embedding models rely on a severely limited, and somewhat naïve, ontology. Most current models are confined exclusively to text data, with words being the only linguistic items under consideration. The fact that two words tend to co-occur is admittedly a useful clue to the meaning of the words, but there may be other types of contextual information that can provide equally useful clues for modeling meaning. In natural discourse, tone of voice, gestures, facial expressions, even time and location are important contextual factors that influence semantic processing. It seems reasonable to assume that this should apply also to computational models and AIs that aim to learn language.

Some recent studies have begun to investigate the possibility to extend the ontology of the co-occurrence model with other modalities such as vision and sound (Bruni, Tran, and Baroni 2014; Vijayakumar, Vedantam, and Parikh 2017). Our aim is more ambitious; one of the goals of the smart data layer is to extend current representation learning models with multi-modal contexts that encompass not only vision and sound, but also other types of contextual data, such as spatio-temporal information, various types of sensor data and infrequently occurring instantaneous events. If our ultimate goal is to build true AI, its representation must be built from more senses than just text.

The Data Sandbox

The type of representation learning mechanisms discussed in the previous sections are data-intensive and require large amounts of data to learn from. We expect the future IoTs to produce tsunamis of data where such models will thrive. However, getting access to such amounts of controlled data for development purposes is currently more difficult. We use the notion of a *data sandbox* (indicating that we start with baby steps) for collecting heterogeneous multimodal data. The data sandbox collects information from a user's computer, and stores the following information:

- Text on the user's screen (using the Google Cloud Vision API[1]).
- Text from the user's keyboard.
- Transcribed speech (using PocketSphinx[2]).
- Sentiment based on the user's facial expression (captured by the computer's camera, and using the Google Cloud Vision API).
- Sentiment based on faces on the user's screen (using the Google Cloud Vision API).
- Labels and categories recognized on the user's screen (using the Google Cloud Vision API).
- Various sensor data, including:
  - CPU usage.
  - Memory and disk usage.
  - Battery life.
  - Temperature.

This heterogeneous data will serve as the foundation for our initial experiments on multimodal representation learning. The idea is to extend embedding models with extralinguistic contexts, such as sentiment labels from facial expressions, or even CPU usage and core temperature. Although the amount of data that we expect to be able to collect using the data sandbox is too small to allow for more advanced techniques such as deep learning or compressed sensing, we plan to use statistical correlation measures to find interesting patterns in the data. As an example, imagine that we

[1] https://cloud.google.com/vision/

[2] https://github.com/cmusphinx/pocketsphinx

use a word embedding technique to learn a concept such as `soccer` based on the text on a person's screen.[3] Next, imagine that we notice that the `soccer` concept often co-occurs with positive facial expressions captured by the computer's camera and low CPU usage. This pattern constitutes a higher-order concept, which we might label something like `taking a break from work`. If instead we notice both high CPU and memory usage in conjunction with the `soccer` concept, we might instead infer that the user is in a state of `waiting for the experiment to finish`.

Possible Use Cases

Producing representations of diverse, textual and non-textual, data provides the possibility to represent user activity in diverse and interesting ways. Yet how could this be made actionable to have an influence on user or system behavior?

One approach taken by Intelligent User Interface research has been to make use of Bayesian models of user activity, "automatically" activating system actions based on predicted desired user outcomes. This has been used to, for example, allow systems to achieve a basic understanding of user intention based on context, and to perform different actions at different times in response to the same input (Wilson and Shafer 2003). Other work has made use of one of our data streams (text scraped from the user interface) to predict users' ongoing "tasks". While potentially interesting, this is a heavily reductionist model of user activity, and user state more broadly which is multifaceted. Clearly, a general representation of user activity has the potential to work in more complex ways.

In conceptualizing different uses of the representations, we have worked with open concepts applicable to varied contexts. Taking a historical view of context, distinctive ways of visualizing user activity from the data streams collected could support searching of past activity by users through looking for similarities between current activity and past activity events. The same interaction paradigm could afford the exploration of activity between users, or groups of users, and could be expanded to not only show the temporal relationship to the membership of a particular class of user, but with expanding the interface to expose the dimensions which relate to each classification. This could show that in one dimension an individual may be an outlier, but similar to many others in the rest of the vector representing this user.

Diverse representations might also enable a richer understanding of contextual inactivity and object appropriation. The insights into user relationships with and through things that this would provide could allow for the development of more sustainable products and systems. A deeper understanding of relationships might also inform a more meaningful design of interactional dialogs with conversational and embodied agents that appropriately act and enact with users and on their behalf. Additionally, a smart data layer might also support the design of experiential narratives that assist multiple user intentions with multimodal interactions for immersive or embodied user experiences. More broadly, building representations of users' ongoing activity may provide ways of supporting ongoing activities, such as speech recognition, Internet search, and advertising. However, we suspect that this would not be in the classic prompting of activity, but in different classes of activity that fit more with the ongoing modeling of action. The openness of these initial concepts enables future investigations into specific use cases across many contexts, from idiosyncratic routines to affective health to enterprise workflows, in which *when* and *how* to interact with the user requires a careful consideration of what can be understood from the systems' understanding the user.

In conclusion, while many of these envisioned use cases for a smart data layer might break with the expected utilitarian forms of use, we propose that in designing such a layer to support more playful, meaningful, and contextually appropriate applications it can be a driver for the development novel paradigms of interaction for the Internet of Things.

Acknowledgments

This work is funded by the Swedish Foundation for Strategic Research (RIT15-0046).

References

Bengio, Y.; Courville, A.; and Vincent, P. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence* 35(8):1798–1828.

Bruni, E.; Tran, N. K.; and Baroni, M. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research* 49(1):1–47.

Collobert, R.; Weston, J.; Bottou, L.; Karlen, M.; Kavukcuoglu, K.; and Kuksa, P. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research* 12:2493–2537.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Computer Vision and Pattern Recognition (CVPR)*, 770–778.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In Pereira, F.; Burges, C. J. C.; Bottou, L.; and Weinberger, K. Q., eds., *NIPS 25*. Curran Associates, Inc. 1097–1105.

Oquab, M.; Bottou, L.; Laptev, I.; and Sivic, J. 2014. Learning and transferring mid-level image representations using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1717–1724.

Sahlgren, M. 2006. *The Word-Space Model*. Ph.D. Dissertation, Stockholm University.

[3] Such concept learning could be accomplished e.g.by clustering the words in an embedding model, resulting in clusters of words that have a semantic relation. A soccer cluster might be populated by words such as "offside", "ball", "goal", "kick" and "Zlatan" (the name of a famous Swedish soccer player).

Turney, P. D., and Pantel, P. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37(1):141–188.

Vijayakumar, A.; Vedantam, R.; and Parikh, D. 2017. Sound-word2vec: Learning word representations grounded in sounds. In *Empirical Methods in Natural Language Processing (EMNLP)*, 931–936. Association for Computational Linguistics.

Wilson, A., and Shafer, S. 2003. Xwand: Ui for intelligent spaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI ’03, 545–552. New York, NY, USA: ACM.

Wu, L.; Fisch, A.; Chopra, S.; Adams, K.; Bordes, A.; and Weston, J. 2017. Starspace: Embed all the things! *arXiv preprint arXiv:1709.03856*.