Preprint:
Talk to be given in July 2016 at the Biennial Meeting of the International Society for the Empirical Study of Literature (IGEL), International Society for the Learning Sciences (ISLS)

Authors:
Lester C. Loschky, Dept. of Psychological Sciences, Kansas State University
William H. Hsu, Dept. of Computer & Information Sciences, Kansas State University
James R. Hamilton, Dept. of Philosophy, Kansas State University

Key words/phrases:
Bayesian modelling
Causal inferences
Visual Narrative Comprehension

Title:
A Proposed Bayesian Model of Plausible Causal Inference for Visual Narrative Comprehension

Brief Abstract:
We describe a proposed Bayesian computational model that can explain how people comprehend visual narratives that allows viewers to generate typical inferences. The model instantiates aspects of the scene perception & event comprehension theory (SPECT), namely the front-end (entity extraction) and back-end (event model construction). The Bayesian model represents causal inferences (particularly bridging inferences) using a description logic containing an ontology of entities and events. We apply the model to the "Boy, Dog, Frog" visual narratives.

Title:
A Proposed Bayesian Model of Plausible Causal Inference for Visual Narrative Comprehension

Abstract:
We want to create a computational model that can explain how people comprehend visual narratives at a level allowing them to generate typical inferences. That level was called the "situation model" by Kintsch. We look at visual narrative comprehension because it is germane to general narrative recognition. Although, philosophically, the most common way to understand narrative structures is metaphysical, concerned with determining what narratives are (at least minimally), the approach we take is epistemological, concerned with what capacities are engaged when a person grasps a narrative.

We will test hypotheses generated from the Scene Perception and Event Comprehension Theory (SPECT) (Loschky, Hutson, Magliano, Larson, & Smith, 2014, June). SPECT distinguishes front-end processes that occur within single eye fixations, and back-end processes that occur across multiple eye fixations within working memory (WM) and long-term memory. The front-end processes extract information that identify entities, locations,

and events. The back-end processes integrate that information to create event models, and in that process, generate inferences for missing information. To create a computational model of these processes, we will utilize deep learning for the front-end, and Bayesian modeling for the back-end. We can then test the model by comparing its output to various types of human data (e.g., think aloud protocols, picture viewing times, eye movements).

The first step will be to create a basic model of what is going on in the picture stories. Front-end processes in SPECT will be instantiated by deep learning algorithms that can identify entities and actions in the narrative images and label them in the formal language of tagged entities (e.g., "person (animate)").

A second step, which occurs in the back-end of SPECT, will be to infer the visible events within the narrative using causal Bayesian inference. Episodes in the narrative can be discerned in terms of changes in the event structure, such as changes in entities (e.g., characters), locations, time, entities' goals, and causal relationships (other than those subsumed by entities' goals).

A third step, also in the back-end of SPECT, will be to generate inferences of missing sub-events (actions). Inevitably, some actions will be missing in visual or textual narratives, due to time and space constraints of communication. Missing actions can be readily inferred by human viewers of a visual narrative (Magliano, Larson, Higgs & Loschky, 2015). A challenge for the Bayesian model will be to recover missing actions in order to maintain a coherent representation of the narrative.

Regarding step 3, according to SPECT, when we look at each image in a sequential visual narrative, back-end processes in WM not only work with representations of what is currently shown in the image, but also representations from previous pictures in the narrative. These WM representations that span images in a sequence will consist of assertions of events, goals, and types (e.g., animate vs. inanimate entities). We will apply our Bayesian model to create formal representations of these WM contents, as causally interrelated combinations of persistent entities, roles, and goals. The Bayesian model represents plausible inferences over this description logic, which is a decidable fragment of first-order predicate logic. Because the Bayesian model is updated based on new information, it will interpret each image as it is seen in the narrative sequence, so the order in which pictures are processed matters.

Imagine a 3-image narrative sequence showing 1) a boy (B) running down a hill to catch a frog (F) in a pond at the bottom of the hill; 2) the boy tripping over a tree branch (T); 3) the boy (B) having fallen into the pond. The descriptive logic will represent that as "B (animate) trips over T (inanimate) while trying to catch F (animate)". Suppose that instead of a 3-image sequence, the viewer of the visual narrative sees only two images 1) and 3), leaving out middle image 2), showing the boy tripping on the tree branch. Maintaining a coherent narrative representation requires that we explain how the boy came to have his feet sticking out the water. That involves drawing a bridging inference across the visual gap. The Bayesian model will can compute such inferences. The range of alternative causes that we can impute within the description logic will include both visible

causes that the viewer can scan for, and other causes that are not necessarily visible, but are contained within background knowledge (e.g., based on prior frequencies). If we infer B fell, but the cause is unidentified, then we must also infer the cause (namely, T). The Bayesian model can enable both types of inferences, based on both priors and new visual experience. For example, it is much more likely that an animate entity will trip over an inanimate entity than another animate entity. In this case, we can infer that the causal entity was likely T (the tree limb).

References:

Bengio, Y. (2009). Learning Deep Architectures for AI. *Foundations and Trends in Machine Learning* 2. doi:10.1561/2200000006.

Bengio, Y, Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), special issue on Learning Deep Architectures, 35*, 1798–1828. doi:10.1109/tpami.2013.50.

Hinton, G. E., Osindero, S., & Teh, Y. (2006). "A fast learning algorithm for deep belief nets" (PDF). *Neural Computation 18*(7): 1527–1554. doi:10.1162/neco.2006.18.7.1527. PMID 16764513.

T. G., & Papenmeier, F. (2014). Changes in situation models modulate processes of event perception in audiovisual narratives. *40*(5), 1377–1388.

Liu, N., Han, J., Zhang, D., Wen, S., & Liu, T. (2015). Predicting Eye Fixations using Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, 362-370.

Loschky, L. C., Hutson, J., Magliano, J. P., Larson, A. M., & Smith, T. (2014, June). Explaining the Film Comprehension/Attention Relationship with the Scene Perception and Event Comprehension Theory (SPECT). Paper presented at the *2014 annual meeting of the Society for Cognitive Studies of the Moving Image*, Lancaster, PA.

Magliano, J. P., Larson, A. M., Higgs, K., & Loschky, L. C. (2015). The relative roles of visuospatial and linguistic working memory systems in generating inferences during visual narrative comprehension. [journal article]. *Memory & Cognition*, 1-13. doi: 10.3758/s13421-015-0558-7

Magliano, J. P., Miller, J., & Zwaan, R. A. (2001). Indexing space and time in film understanding. *Applied Cognitive Psychology*, 15(5), 533-545.

Magliano, J. P., Taylor, H. A., & Kim, H.-J. J. (2005). When goals collide: Monitoring the goals of multiple characters. *Memory & Cognition*, 33(8), 1357-1367.

Mayer, M. (1980). *Frog, where are you?* New York, NY, US: Dial Books for Young Readers.

Mayer, M. (1992). *A boy, a dog, and a frog.* New York, NY, US: Dial Books for Young Readers.

Palatucci, M., Pomerleau, D., Hinton, G. E., & Mitchell, T. M. (2009) Zero-Shot Learning with Semantic Output Codes. In *Proceedings of Neural Information Processing Systems (NIPS 2009)*, 1410-1418.

Sattar, H., Müller, S., Fritz, M., & Bulling, A. (2015). Prediction of search targets from fixations in open-world settings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2015)*, 981-990.

Shi, T., Liang, M., & Hu, X. (2014). A Reverse Hierarchy Model for Predicting Eye Fixations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2014)*, 2822-2829.

Socher, R., Ganjoo, M., Manning, C. D., Ng, A. Y. (2013). Zero-Shot Learning Through Cross-Modal Transfer. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems (NIPS 2013)*, 935-943.

Yu, A., & Grauman, K. (2015). Just Noticeable Differences in Visual Attributes. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, Dec 2015.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2), 162-185.