Erasmus University Rotterdam

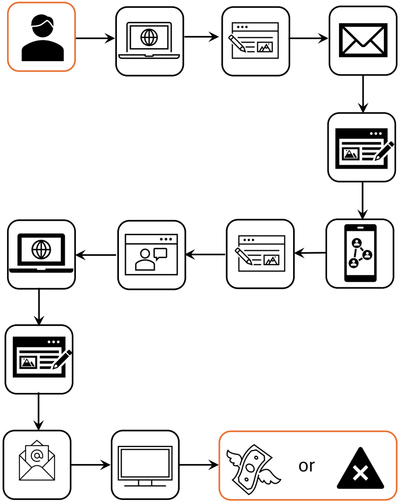
Conversion Attribution Under Uncertainty: A Deep Ensemble Approach*

Kathrin Gruber

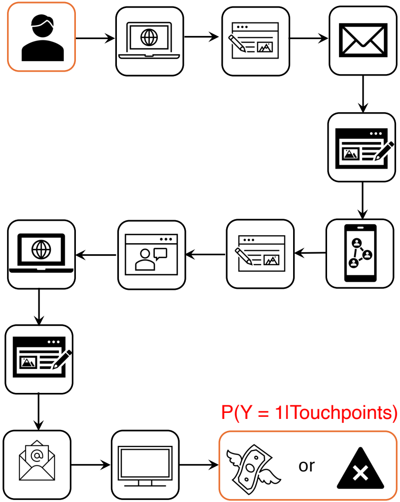
Department of Econometrics
Erasmus School of Economics
gruber@ese.eur.nl

*Joint work with Joram De Vreede (Billy Grace)

The (Class) Prediction Task



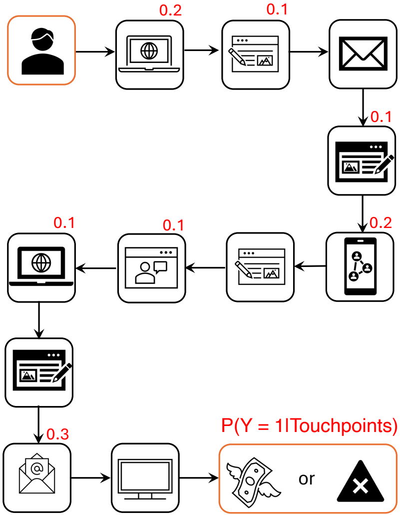
The (Class) Prediction Task



Neural language models (e.g., Ren et al. 2018; Li et al. 2018; Du et al. 2019; Kumar et al. 2020, ...):

"translate" an input sequence of touchpoints to a probability of conversion.

The (Class) Prediction Task

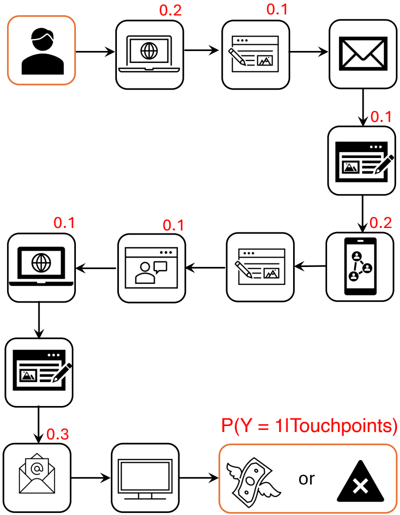


Neural language models (e.g., Ren et al. 2018; Li et al. 2018; Du et al. 2019; Kumar et al. 2020, ...):

“translate” an input sequence of touchpoints to a probability of conversion.

alignment weights between decoder and encoder (hidden states) are the “contributions” of the touchpoints.

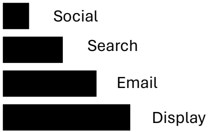
The (Class) Prediction Task



Neural language models (e.g., Ren et al. 2018; Li et al. 2018; Du et al. 2019; Kumar et al. 2020, ...):

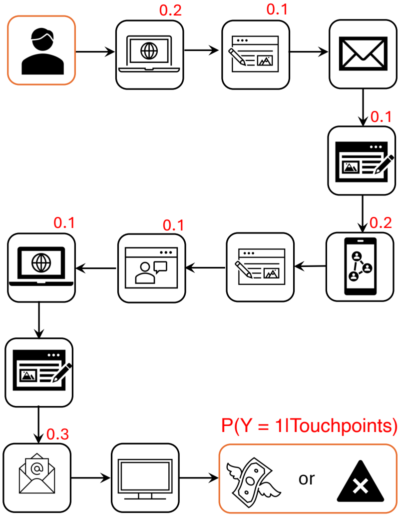
"translate" an input sequence of touchpoints to a probability of conversion.

alignment weights between decoder and encoder (hidden states) are the "contributions" of the touchpoints.



Channel attribution

The Overconfidence Problem



Class imbalance: Conversion events are extremely rare (0.1% – 5%).

Mislabelling: Users routinely jump between numerous digital devices.

Distribution shift: Frequent changes in campaign designs.

Poor generalizability:
No feasible solution with state-of-the-art (i.e., neural language model-based) attribution architectures.

Our Contribution(s)

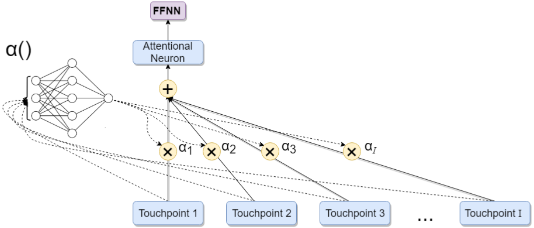
A hierarchical (i.e., transformer-based) machine learning architecture optimized for multi-touch conversion attribution:

- **Easy-to-interpret**: A simplistic feed-forward attention mechanism attributes the conversion credits (attribution scores) directly on specific touchpoints.
- **On-the-go-results:** Enables ensemble techniques (e.g., Bagging, Breiman 1996; Bayesian averaging, Raftery et al. 1997) for improved and robust classification of previously unseen data.

Architecture

Feed-forward attention (Raffel & Ellis, 2015):
For embedding vector $\boldsymbol{z}_i$ (i.e., touchpoint in postion i of a user sequence of length n) of dimension d.

$$
\begin{aligned}
r_i &= a(\boldsymbol{z}_i), && \text{(attention score)} \\
v_i &= \text{softmax}(\boldsymbol{r}), && \text{(attention weight)} \\
\boldsymbol{c} &= \textstyle\sum_i v_i \boldsymbol{z}_i, && \text{(context vector)}
\end{aligned}
$$



Feed-forward attention for touchpoint attribution.

Architecture

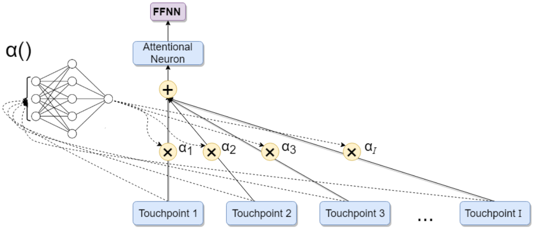
Positional Encodings (Vaswani et al. 2017):

$$\boldsymbol{z}_i' = \gamma \boldsymbol{z}_i + \boldsymbol{p}_i.$$

where

$$\boldsymbol{p}_i = \begin{cases} \sin\left(\omega_k, i\right), & \text{if } i = 2k, \\ \cos\left(\omega_k, i\right), & \text{if } i = 2k+1, \end{cases}$$

with $\omega_k = 1/1000^{(2k/d)}$ for $k = 1, \ldots d/2$.



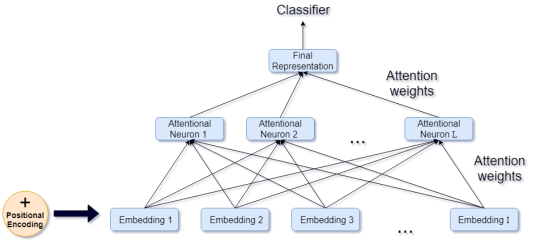
Feed-forward attention for touchpoint attribution.

The Stacked Web of Attentional Neurons:

$$\boldsymbol{c} = \sum_l v_l \sum_i v_{li} \boldsymbol{z}'_i,$$

with attention weight v_{li} for touchpoint i from context vector l, and attention weight v_l for context vector l from the final representation.

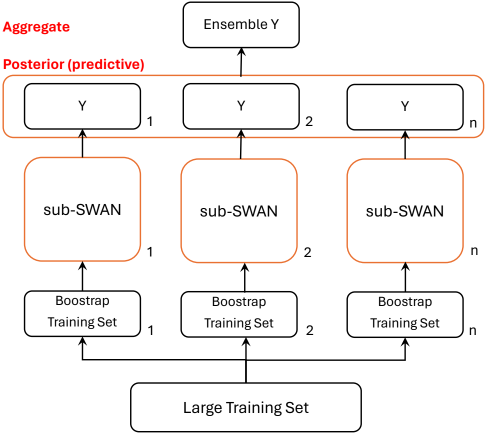
$$P(Y = 1|\boldsymbol{c}) > 0.5$$



Context vector stacking for touchpoint interactions. Note: an attentional neuron refers to a general context vector.

Uncertainty Quantification

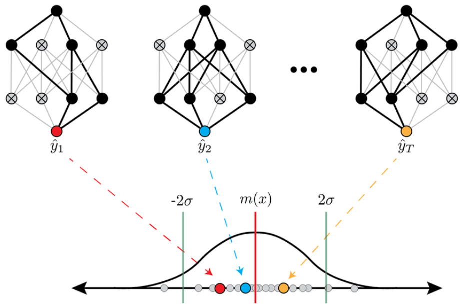
Alleatoric (data-based) uncertainty: N-individual SWAN networks *trained* on N-undersampled datasets. The final classification is the aggregate of the sub-SWANs.



Uncertainty Quantification (2)

Alleatoric (data-based) uncertainty: N-individual SWAN networks *trained* on N-undersampled datasets. The final classification is the aggregate of the sub-SWANs.

Epistemic (model-based) uncertainty: Randomly drop a set of nodes for every forward pass of the data during *testing* (Gal & Ghahramani, 2016).



Experiments

Data (train $=$ 80%, test $=$ 20%)

- Real: 6.1 million user sequences with 59,098 unique touchpoints; 4.7% conversion ratio (Diemert et al., 2017).
- Simulated: 10 million user sequences with 20 unique touchpoints; 2.0% conversion ratio.

Benchmarks

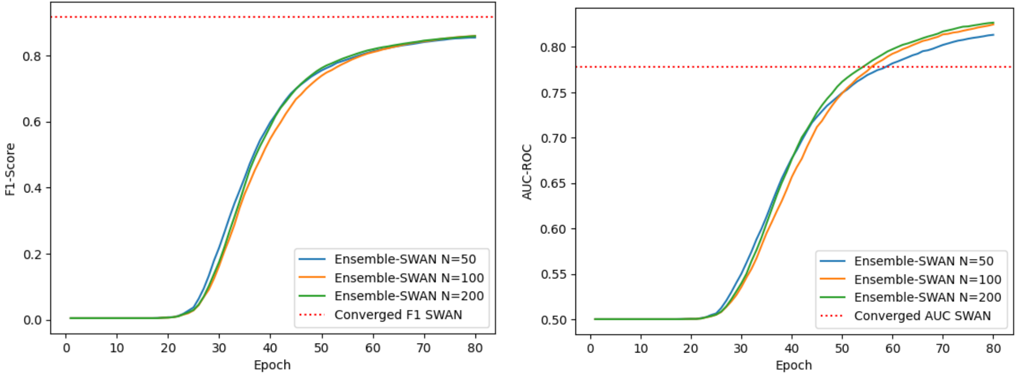
- SWAN: 1 layer of four attentional neurons; trained for 7 epochs (split $=$ 50–50, batch size $=$ 1024, $d = 256$).
- Ensemble-SWAN: 1,000 forward passes and 25% dropout percentage; trained for 80 epochs (split $=$ 50–50, batch size $=$ 1024, $d = 256$).
- ARNN: attention-augmented RNN encoder part (Ren et al. 2018); trained for 7 epochs (split $=$ 50–50, batch size $=$ 1024, $d = 256$).

Conversion Prediction Accuracy

	SWAN	**Ensemble SWAN**	**ARNN**
Accuracy	79.7%	80.5%	79.7%
Precision	74.7%	94.7%	74.5%
F1-score	72.1%	85.5%	72.1%
AUC-ROC	67.8%	82.7%	70.0%

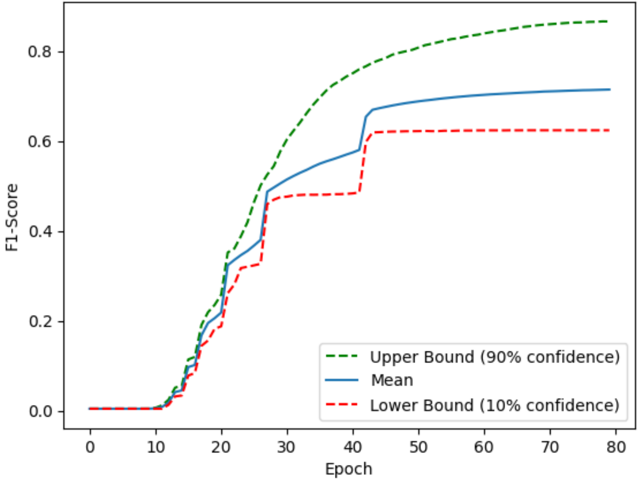
Out-of-sample conversion prediction performance. Ensemble results obtained for $N = 200$ sub-SWANs.

Conversion Prediction Accuracy (2)



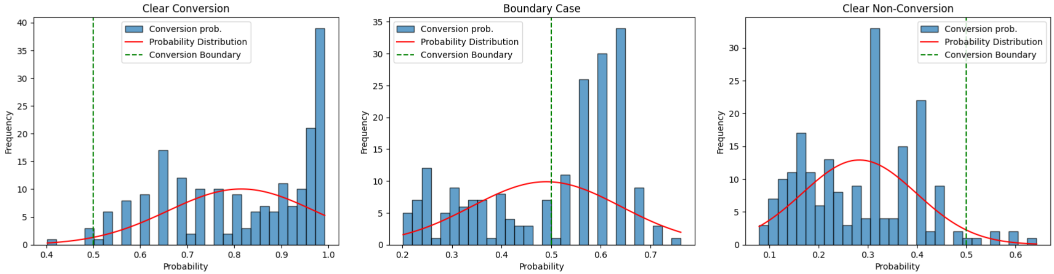
Out-of-sample F1-Score (left) and area under the ROC curve (right) for the converged SWAN (red dotted line) and the Ensemble-SWAN.

Alleatoric Uncertainty



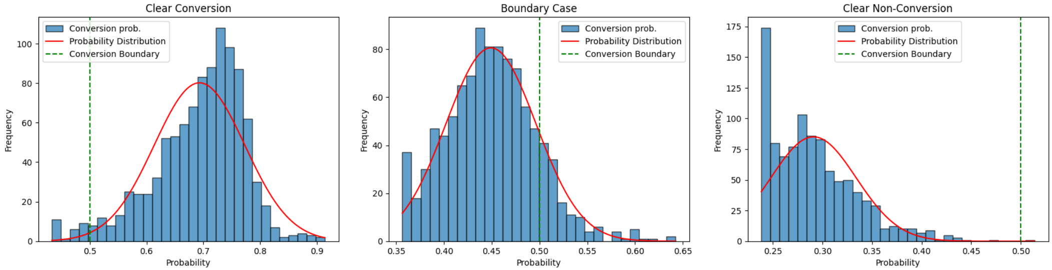
Out-of-sample average F1-score, incl. 90% (upper) & 10% (lower) quantiles for $N = 200$ sub-SWANs before aggregation.

Alleatoric Uncertainty (2)



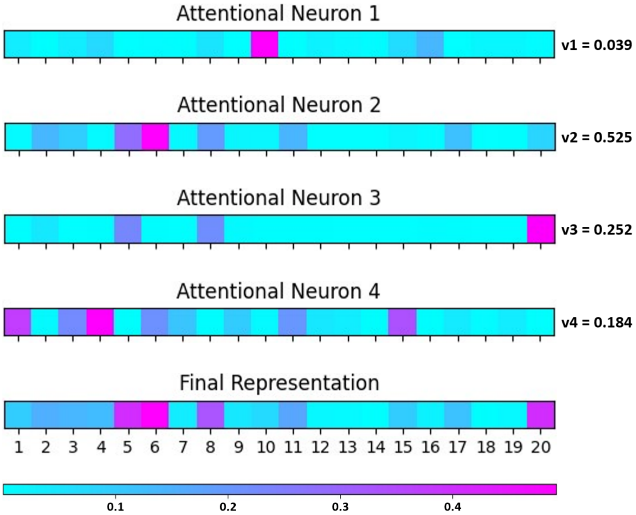
(Non-aggregated) posterior predictive distribution of the $N = 200$ sub-SWANs for three different sequences. The green, vertical, dashed line indicates the chosen conversion probability threshold.

Epistemic Uncertainty

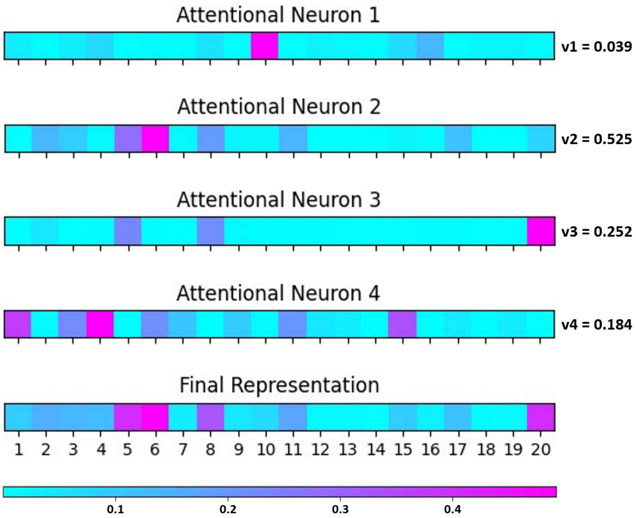


Posterior predictive distribution for 1,000 forward passes for three different sequences. The green, vertical, dashed line indicates the chosen conversion probability threshold.

Attribution Scores



Attribution Accuracy



	MSE 1×10^6
SWAN	2.29
ARNN	6.63
Shapley Value	9.37
Last-touch	9.05
Linear-touch	9.67

Simulated Mean Squared Error. Smaller values indicate better model fit.

Summary

The (Ensemble)-SWAN is an (1) easy to interpret, (2) computational efficient and (3) robust transformer architecture specialised for conversion attribution problems.

Outlook

- Adaption to a more “Bayesian” approach (evidential regression, Amini et al. 2020).
- Uncertainty propagation to touchpoint- and/or channel-specific attributions.
- Field test on more “accessible” data with different additional features (e.g., time between clicks, time spent on a wepage, etc.).

Many thanks for your attention!

Referrences

Amini, A., Schwarting, W., Soleimany, A., & Rus, D. (2020). Deep evidential regression. Advances in Neural Information Processing Systems, 33, 14927-14937.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. The Annals of Statistics 24 (6), 2350–2383.

Diemert Eustache, Meynet Julien, P. Galland, and D. Lefortier (2017). Attribution modeling increases efficiency of bidding in display advertising. In Proceedings of the AdKDD and TargetAd Workshop, pp. 1–6. KDD: ACM.

Du, R., Y. Zhong, H. Nair, B. Cui, and R. Shou (2019). Causally Driven Incremental Multi Touch Attribution Using a Recurrent Neural Network. arXiv preprint.

Gal, Y. and Z. Ghahramani (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In international Conference on Machine Learning, pp. 1050–1059. PMLR.

Haixiang, G., L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing (2017). Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications 73, 220–239.

Referrences

Kumar, S., G. Gupta, R. Prasad, A. Chatterjee, L. Vig, and G. Shroff (2020). Camta: Causal attention model for multi-touch attribution. In 2020 International Conference on Data Mining Workshops (ICDMW), pp. 79–86. IEEE.

Li, N., S. K. Arava, C. Dong, Z. Yan, and A. Pani (2018). Deep neural net with attention for multi-channel multi-touch attribution. ArXiv abs/1809.02230.

Raffel, C. and D. P. Ellis (2015). Feed-forward networks with attention can solve some long-term memory problems. arXiv preprint.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997). Bayesian model averaging for linear
Ren, K., Y. Fang, W. Zhang, S. Liu, J. Li, Y. Zhang, and J. Wang (2018). Learning Multitouch Conversion Attribution with Dual-attention Mechanisms for Online Advertising. Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 1433–1442.

Richardson, M., E. Dominowska, and R. Ragno (2007). Predicting clicks: estimating the click-through rate for new ads. In Proceedings of the 16th International Conference on World Wide Web, pp. 521–530.

Simulated Mean Squared Error

Mean (squared) error in reverse-engineering the simulated conversion attribution process:

$$\text{MSE} = \frac{1}{n} \sum_i \left(C_i - \widehat{C}_i \right),$$

with C_i the total number of conversions attributed to the i-th touchpoint and $\widehat{C}_i$ its estimate.

The true attribution of the i-th touchpoint is the contribuion of its main effect plus half of the effect of its pair-wise interactions:

$$\text{Attr}_i = \frac{e_i + \frac{1}{2} \sum_{j \neq i} e_{i,j}}{S}.$$