IBM **Research** – Industries & Solutions

Correcting Forecasts with Multifactor Neural Attention

Matthew Riemer, Aditya Vempaty, Flavio P. Calmon, Fenno F. Heath III, Richard Hull, and Elham Khabiri

IBM Watson Research Center

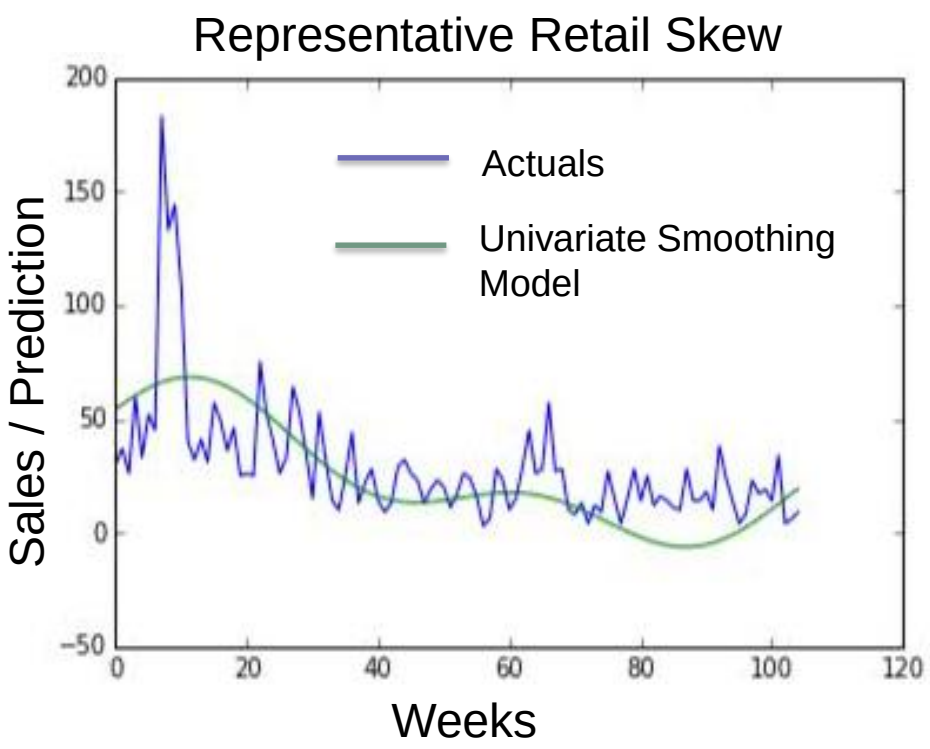
Contact: mdriemer@us.ibm.com

Contributions of This Work

- We extend mixture of experts neural network models proposed in the 90s to include a modern soft attention mechanism
- We are the first to show value of neural attention in the domain of time series forecasting
  - We consider a setting with thousands of exogenous variables
- We demonstrate a model that has high empirical accuracy and an unprecedented level of descriptive ability for a neural network forecast
- We propose multiple novel technical contributions that qualitatively and quantitatively improve mixture of experts models:
  - Sparse Attention
  - Dropout of Experts

© 2015 International Business Machines Corporation

A New Approach is Needed for Forecasting



- Prominent univariate forecasting algorithms attempt to model the longer term trends in data by smoothing out the peaks and valleys
  - They assume that short term spikes are random anomalies!
  - However, in actuality fluctuations happen for real reasons determined by countless external factors that business do not attempt to model.
- Considering the widespread use of these models, it is perhaps not surprising that many industries have seen marginal or no improvement in forecasting ability over recent decades.
- In practice, it is the job of human analysts to incorporate in the external factors and the forecast is only used as guidance.
- In fact, (Franses & Legerstee, 2009) conducted a case study where 90% of all forecasts were found to be manually adjusted.

© 2015 International Business Machines Corporation

Human Derived Forecasts

- Large amounts of human analyst time spent on forecast correction can be very costly!
- It is not even clear that humans do a particularly good job at estimating the impact of external events because they are known to have various biases (Lawrence et al., 2006).
  - For example, humans were found to often have an optimism bias in their projections of the impact of promotions in (Fildes et al., 2009), (Trapero et al., 2011), and (Trapero et al., 2013).
- Humans also may not be knowledgeable about external factors such as local or national events in advance when adjusting a forecast.

© 2015 International Business Machines Corporation

Automatic End to End Systems

- For some high frequency forecasting applications, there is little time for humans to intervene and decisions need to be made automatically.
- As we are starting to live in the era of Big Data and the Internet of Things, we can certainly conceive of a computer being consistently tapped into more relevant impactful sources than human analysts today:
  - Various signals related to Weather
  - Various signals related to Social Media
  - Various signals related to internet activity
  - Various signals related to cell phone movement data
  - Various relevant economic indicators
  - Information about National events
  - Information about Local events
  - Ongoing promotions

- Our experience is that a team of data scientist is needed to work on feature engineering in order to achieve acceptable performance incorporating these factors with traditional models.

© 2015 International Business Machines Corporation

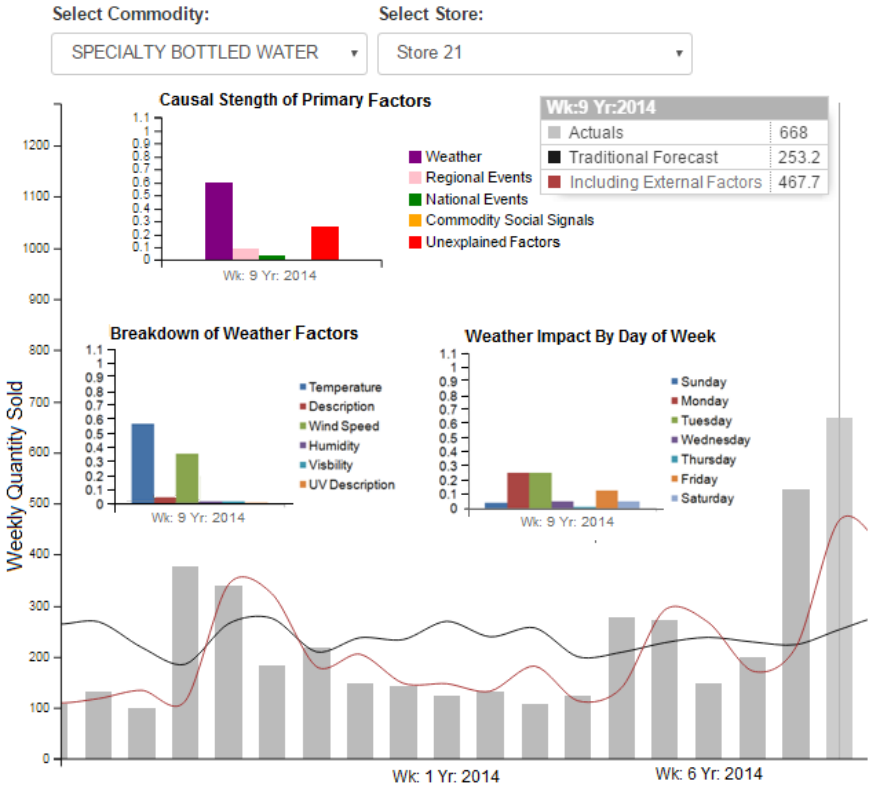
Neural Networks and the “Black Box” Complaint

- Neural Networks have been shown to be highly beneficial for solving problems without feature engineering when sufficient data quantities are available.

- In fact, Neural networks have a substantial history of quantitatively superior performance to industry standard forecasting techniques in literature.
  - Far from exhaustive list: (Xu et al., 2005), (Castillo et al., 2006), (Sunaryo et al., 2011), (Cortez et al., 2012), (Marvuglia & Messineo, 2012), (Neupane et al., 2012).

- They have ultimately not been widely adopted in industry because their modest improvements have come at the cost of not being interpretable.

Visualizing Our Proposed Model

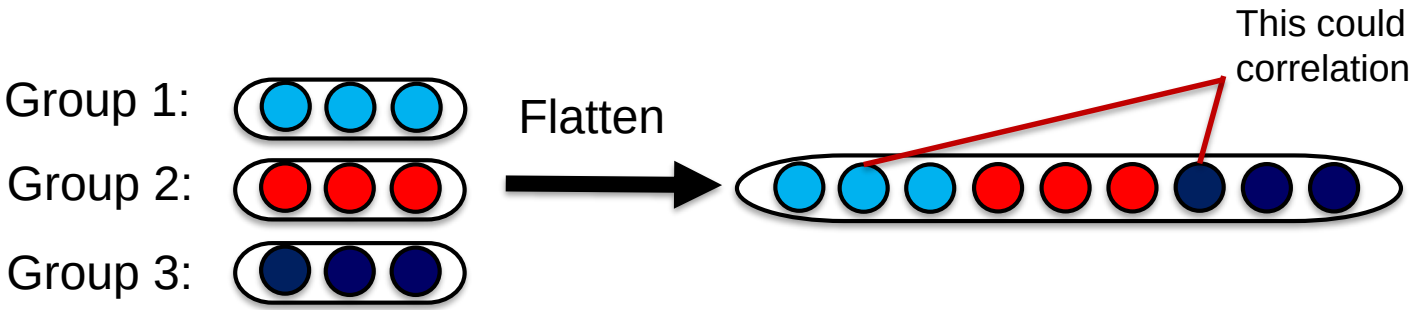


© 2015 International Business Machines Corporation

Observation 1: Input Can Be Rich with Structure That We Don't Leverage

- One limitation of most predictive modeling approaches is that they require input to be presented as a vector.
- This may not seem like a problem because any tensor can be flattened into a vector, but information about semantic boundaries within the input is lost!

With input in this form, many learners don't understand the significance of the original semantic grouping of the data!

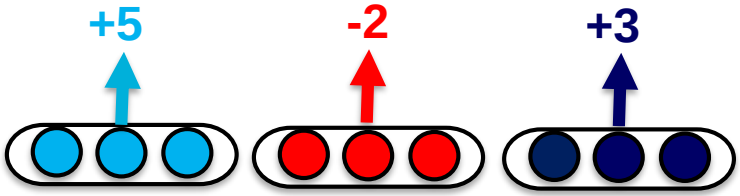


© 2015 International Business Machines Corporation

Observation 2: Humans Can Provide Additive Interpretations

- Humans can perform additive decompositions based on semantic groups to explain their thought process:
  - i.e. “I expect we will sell 50 less product on Friday because of the predicted weather showing a Thunder Storm” (“weather” and “Thunder Storm” refer to different hierarchical levels of semantic groupings of multiple input readings)

- With this motivation, we explore a model where each group of features is mapped to a predicted additive outcome given proper context.



Hierarchical Interpretability:
- All “blues” account for +8.
- Total contribution is +6.

© 2015 International Business Machines Corporation

A General Baseline Correction Formulation

Current time step τ Baseline Forecast: $B(\tau)$ Current Factor: f

Current Observation: i Observation Vector: $x_{if}(\tau)$ Impact Vector: $y_{if}(\tau)$ Prediction at Current Time: $p(\tau)$

Set of Factors: F Previous Period of Relevant Time for Factor: P_f Total Observations of Current Factor and Time: N_f

$$y_{if}(\tau) = G(x_{if}(\tau))$$

$$p(\tau) = B(\tau) + \sum_{f}^{F} \sum_{\tau}^{P_f} \sum_{i}^{N_f} y_{if}(\tau)$$

- The input observations alone for a feature group is generally not expressive enough. As such, we create the observation vector by concatenating the raw observations with an appropriate context vector.

$$x_{if}(\tau) = concatenate(r_{if}(\tau), context_{if}(\tau))$$

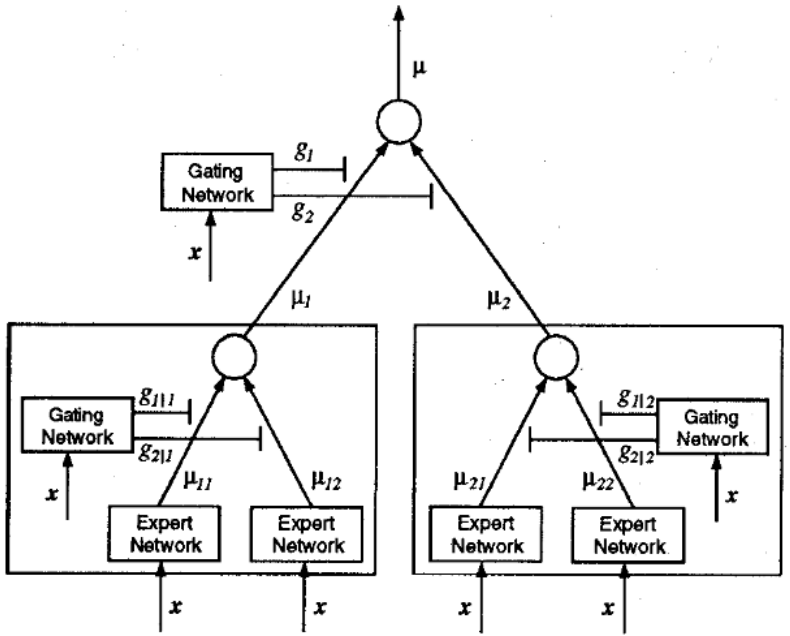
- The context vector can be explicit (i.e. the time of year) or learned (i.e. the hidden layer of an RNN).

© 2015 International Business Machines Corporation

Hierarchical Mixture of Experts Models (Jacobs et al., 1991), and (Jordan and Jacobs, 1994).

- A model in this form can also be defined through a particular instantiation of a one-level mixture of experts model.
- The key addition is that of a gating network which computes a weighted average of each expert network's outputs based on each network's input.
- Our architecture seems to benefit greatly from a gating network. The network follows a new logical procedure:

1. Each observation of each feature group is considered in isolation to determine how interesting the observation is.
2. All of the observations are considered in context and a small subset is picked that are most likely to have influence on the forecast.
3. The individual impact of each important observation, given its importance in the context, is assessed and added together to determine a prediction.



A two-level hierarchical mixture of experts model as depicted in (Jordan and Jacobs, 1994)

© 2015 International Business Machines Corporation

Content Based Attention Mechanisms

- Since, it was first proposed in the field of Machine Translation in (Bahdanau et al., 2014), content based attention has proven to be a useful idea in many cases.
- It is noted in (Cho et al., 2015) that it useful for problems with richly structured inputs and outputs:
  - Speech recognition (Chorowski et al., 2015)
  - Image caption generation (Xu et al., 2015)
  - Reading comprehension (Hermann et al., 2015)
  - Video description generation (Yao et al., 2015).
- Attention can also be highly beneficial when the input has a rich structure and the output is simple:
  - Textual entailment (Wang and Jiang, 2015), and (Rocktäschel et al., 2015).
- In our setting, attention can be seen as serving the same purpose as a gating network while being based on the hidden units and not the inputs.

$$m_{if}(\tau) = sigmoid(W_{mf} h_{if}(\tau) + b_{mf}) \qquad a_{if}(\tau) = \frac{m_{if}(\tau)}{\sum_f^F \sum_\tau^{P_f} \sum_i^{N_f} m_{if}(\tau)}$$

$$d_{if}(\tau) = tanh(W_{df} h_{if}(\tau) + b_{df})$$

$$y_{if}(\tau) = a_{if}(\tau) d_{if}(\tau)$$

© 2015 International Business Machines Corporation

Attention Should be More Powerful Than a Gating Network

- Some theoretical advantages of using a hidden representation:
  - It may be more powerful due to more learnable parameters
  - It may generalize better because the hidden layer tends to be small relative to the input feature size in our setting
  - It may generalize better because the hidden layer weights are shared between both the attention score and output vector, the representation is biased in a potentially favorable way
- We see experimentally that the attention formulation yields much improved empirical results in our setting.
- We also propose a novel “sparse attention” paradigm motivated by the fact that few factors should be impactful at each moment (with mean squared error and positive attention amplitudes):

$$Loss(\tau) = (t(\tau) - p(\tau))^2 + \beta \sum_{f}^{F} \sum_{\tau}^{P_f} \sum_{i}^{N_f} m_{if}(\tau)$$

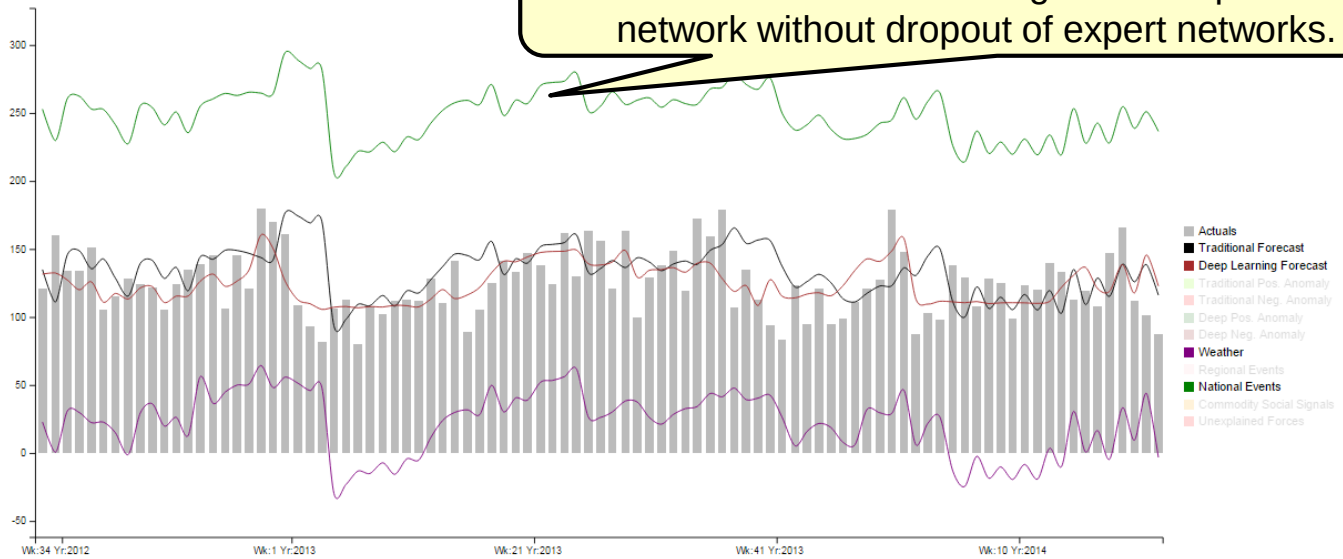
- We find this to significantly improve both gating networks and attention networks in our setting!

© 2015 International Business Machines Corporation

Preventing Co-Adaption of Features

- We apply random dropout (Hinton et al., 2012) of features both on the input representation and on the output of the expert networks.
- Dropout of the expert networks is vital to avoiding uninterpretable co-adapted relationships between the output of the expert networks:

Here the uninterpretable relationship between National Events and Weather factors is a great example of co-adaption in a network without dropout of expert networks.



© 2015 International Business Machines Corporation

Experimental Setup

Prominent Retailer

Transactional Data and Product Hierarchy Data for 107 stores in mid-west from May 2012 to May 2014.

A univariate forecast based on transactional data re-learned every week adjusting for seasonality.

The Weather Channel

Weather Data for nearest station to each store from May 2012 to May 2014.

Local Event Metadata across local area from May 2012 to May 2014.

10% of all Tweets from May 2012 to May 2014:
- ✓ Mentions of 51 national events & holidays
- ✓ Mentions, sentiment, and intent related to each product category

© 2015 International Business Machines Corporation

Empirical Results

Model	Features	Total: MAPE	Total: Anomaly%	Hardest 5: MAPE	Hardest 5: Anomaly%
Baseline Forecast	N/A	26.79%	7.13%	50.76%	16.77%
Our Model	Independent Observations	**20.40%**	**5.11%**	**34.87%**	**11.56%**
- Sparse Attention		23.69%	5.65%	38.66%	12.03%
- Soft Attention		33.49%	11.05%	55.72%	25.98%
Our Gating Network	Independent Observations	24.98%	6.74%	38.29%	12.95%
+ Sparse Attention		24.01%	6.23%	35.89%	11.69%
Random Forest	Flattened Feature Vector	24.87%	5.77%	43.14%	13.14%
Feed-forward Network	Flattened Feature Vector	28.27%	5.78%	49.42%	12.86%
Support Vector Regression	Flattened Feature Vector	31.46%	6.60%	61.60%	19.58%
Decision Trees	Flattened Feature Vector	34.17%	9.62%	52.91%	19.89%
Bayesian Regression	Flattened Feature Vector	38.74%	14.24%	74.87%	31.20%
Lasso Regression	Flattened Feature Vector	46.76%	16.49%	89.67%	34.48%

© 2015 International Business Machines Corporation

Conclusion and Acknowledgements

In our work we explored the following ideas that yielded positive results:

- It seems that when supplied with a gating or attention mechanism, neural networks can constructively utilize input data in its semantically parsed form for superior empirical results.
- Our work supports the hypothesis that a neural network can overcome the simplicity of an additive output layer to achieve strong empirical results if it is sufficiently powerful before that layer.
  - This can be a compelling approach in many domains as additive layers are particularly interpretable when they can be mapped directly to inputs
  - We explore some new ideas that can help this kind of network generalize despite its tendency to over fit:
    - Sparse Attention
    - Dropout of Experts

We would like to thank and acknowledge Krishna Ratakonda (IBM Research), Rakesh Mohan (IBM Research), Matthew McNaughten (IBM Commerce), and Mark Andrews (IBM Analytics) for helpful conversations and guidance that made this work possible!

More questions? Visit our poster Wednesday morning from 10am to 1pm!

© 2015 International Business Machines Corporation

BACKUP

© 2015 International Business Machines Corporation

A Simple Neural Network Formulation

- Here is a simple example applied to a neural network with tanh units:

$$h_{if}(\tau) = tanh(W_{hf}x_{if}(\tau) + b_{hf})$$
$$y_{if}(\tau) = tanh(W_{yf}h_{if}(\tau) + b_{yf})$$

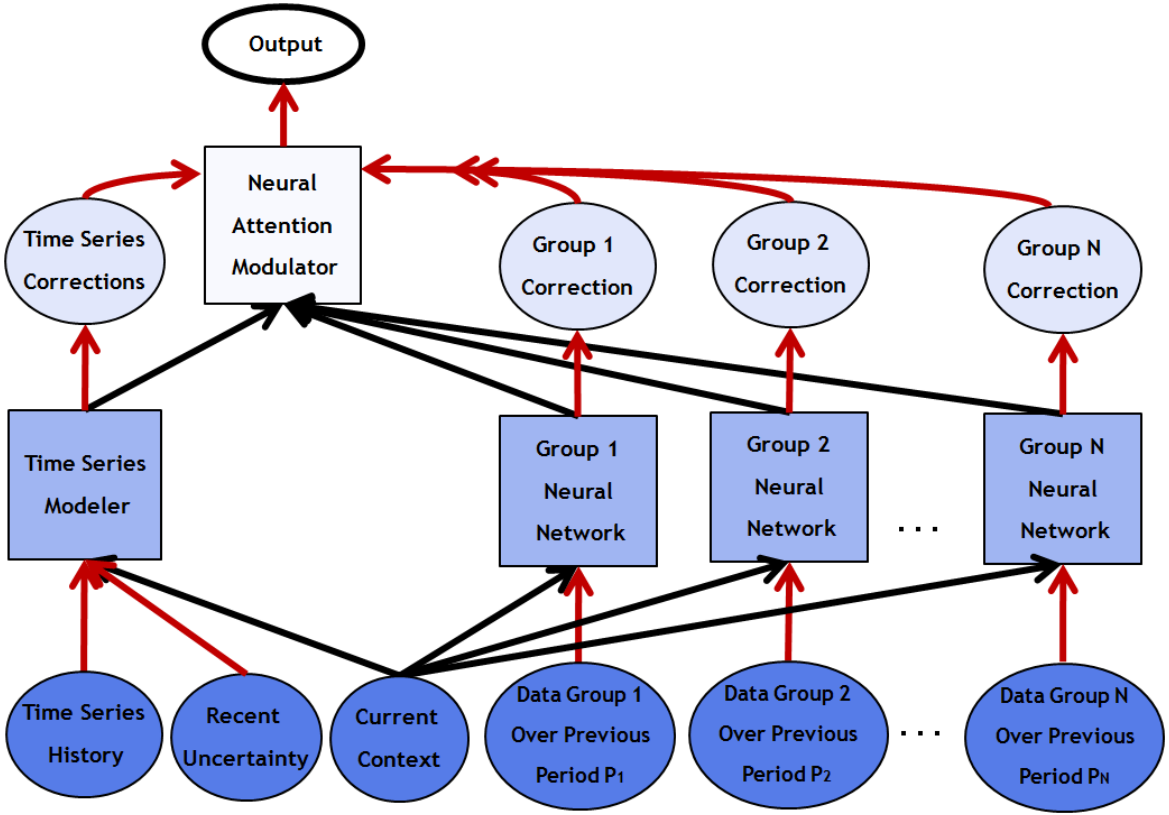
- An obvious issue to avoid in our setting of modifying a baseline forecast is overfitting the effect of external factors when a large part of the error is not accounted for by these factors.

- As such, we also provide all of our neural network models with recent error signals and the last actual value.

- This leads to substantial increases in performance when the baseline error is large. However, this formulation still performs very poorly in our experiments!

© 2015 International Business Machines Corporation

Process Flow of Our Proposed Model



© 2015 International Business Machines Corporation

Analysis of Results

- Networks with gating and attention are able to consistently find generalizable representations utilizing semantically structured feature groups.
- Sparse attention seems to improve results considerably and to some degree for the case of gating networks.
- Models that are more simple seem to over fit significantly on highly volatile commodities.

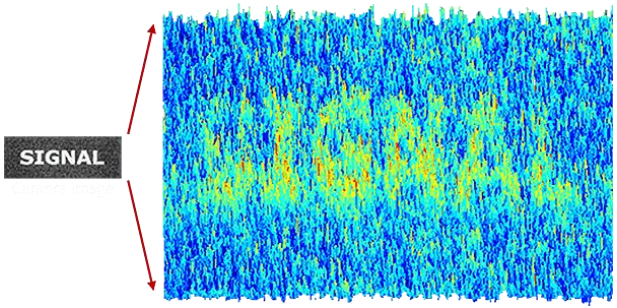
Component	All 20	Hardest 5
Unexplained Correction	54.7%	48.7%
Weather	22.8%	26.1%
Commodity Social Signal	18.6%	22.4%
National Events	2.4%	1.1%
Local Events	1.5%	1.6%

Relative total contribution of each group of observations to the prediction of all 20 commodities and the hardest 5 commodities to model over 2 years of data.

© 2015 International Business Machines Corporation

Intuition About “Noisy” Data

- In our experience the bulk of the separation between our attention models and traditional approaches came on the most volatile commodities.
- They are “noisy” in that their sales are highly volatile, with little training data, and thousands of possible explanatory features to consider.
- **Our intuition:** attention based neural networks should play a role in combating this “noisy” data problem.
  - The sparse attention mechanism forces entire observation vectors to have zero influence on the prediction early in training.
  - This effectively shrinks the number of explanatory variables considered by the model at that point.
  - A small number of values in an observation vector may by chance have a high correlation with the volatility in the signal over a small period (this is more probable as volatility increases).
  - The attention mechanism makes a holistic judgment based on a group of features to dismiss the entire group and shield the model from reacting to spurious correlations in a small subset of the observation vector.
- Our experiments seem to support this hypothesis, but a more rigorous theoretical analysis of the properties of this model will be left to future work.



© 2015 International Business Machines Corporation

End to End Model

- Our focus has been on a neural network module that corrects an existing time series signal with no sharing of the latent parameters used for time series prediction.
- We experiment with a GRU (Cho et al., 2014) recurrent neural network with sparse regularization as our time series modeler that is sent the entire prior history of the store's time series concatenated with a one hot store encoding at each time step.

$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1}) \quad (1)$$

$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1}) \quad (2)$$

$$\tilde{h}_t = tanh(W_{xh}x_t + r_t \circ W_{hh}h_{t-1}) \quad (3)$$

$$h_t = z_t \circ h_{t-1} + (1 - z_t) \circ \tilde{h}_t \quad (4)$$

- We adjust our architecture slightly to allow for sharing of latent parameters:
  - We concatenate both the output and last hidden representation of the GRU to the context vector for all observations for each external feature group.
  - The uncertainty vector and unexplained factors are not needed.
- Empirically, we find this model achieves 20.28 MAPE with a 5.05 anomaly percentage.
- This at least seems indicative that there is some value in tighter integration of the expert models to regularize the univariate forecast.

© 2015 International Business Machines Corporation