Interpretability in Multimodal Deep Learning

Purvanshi Mehta[1] Antoine Ledent[1] Waleed Mustafa Robert Vandermeulen Marius Kloft

TU Kaiserslautern, Contact - purvanshi.mehta11@gmail.com

Abstract

We focus on the processing of heterogeneous sources of information with the aim of determining which modalities or their combinations play an important role in prediction. The interaction of high level features of combinations is modelled through tensor products. We provide:

- A generalized l_p-norm approach to Multimodal Deep Learning.
- A novel iterative batch normalisation procedure to be applied to the tensor fusion model. The aim of this procedure is to disentangle the modeling of interactions in the tensor products by making higher order products unable to model lower degree information.

Basic network combination

- Take weighted contribution of each sub-network to form the score function

$$N_\beta = \sum_{m=1}^{M} \beta_m \left\langle f^m_{w^m_1, w^m_2 .. w^m_{L-1}}(x_i), W^m_L \right\rangle + b.$$

The optimal weights $\beta_1, \beta_2, \ldots, \beta_m$ are importance factors for each subnetworks.

- The formulation of our *optimization problem* is

$$\min_{W_1,\ldots,W_L,b} \sum_{i=1}^{n} \ell\left(\sum_{m=1}^{M} \beta_m \left\langle f^m_{W^m_1,\ldots,W^m_{L-1}}(x_i), W^m_L \right\rangle + b, y_i\right) + \lambda \sum_{m=1}^{M}\sum_{l=1}^{L} \|W^m_l\|_2^2$$

- This can be shown to be equivalent to training a combined network with last layer loss

$$\sum_{i=1}^{n} \ell\left(\sum_{m=1}^{M} \left\langle f^m_{w_1,w_2\ldots,w_{l-1}}(x_i), W^m_L \right\rangle + b, y_i\right),$$

using the modified regularizor

$$\lambda\left(\sum_{l=1}^{L-1} \|W_l\|_2^2 + \left[\sum_{m=1}^{M} \|W^m_L\|_2^q\right]^{2/q}\right), \quad \text{with}$$

$q = (p+1)/2p$.

Batch Normalisation

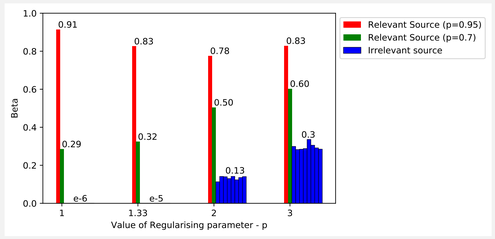
Problem with Beta weights - If last layer features are very large, a low value of Beta might correspond to highly relevant features.

- **Vector BatchNorm (VBN)** A slight adaptation of batch normalization to ensure that the L_2-norm of the features, *averaged over a minibatch*, is constant.

$$\mu_{B,m} = \frac{\sum_{i\in B} f^m(x^m_i)}{|B|}$$

and $$\sigma^2_{B,m} = \frac{\sum_{i\in B} \|f^m(x^m_i) - \mu_B\|_2^2}{|B|}$$

Figure 1: Synthetic data experiments with conditionally independent views



Tensor Fusion and Degree Inflation

- **TensorFusion** To capture the the inter and intra modality relevance of features the Kronecker product of the last layer activations of modalities is taken.
- **Degree Inflation** Phenomenon of concentration of weights in higher order products.

Figure 2: Degree Inflation Example

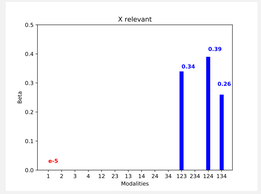
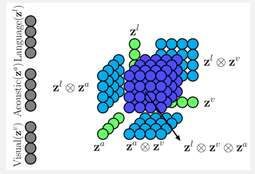


Figure 3: Tensor fusion model for three modalities



Learning of constants

- A tensor fusion model can learn constants in one modality to represent, inside tensor products between two modalities, information that effectively comes only from one modality.

Figure 4: Relevant modality - 1 with the highest eigen vector value

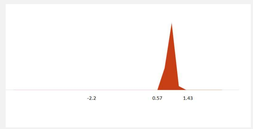
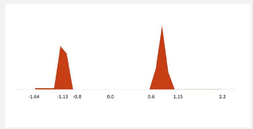


Figure 5: Irrelevant modality - 2 ⊗ 3 with the highest eigen vector value



- To remedy the problem of degree inflation we propose the *Iterative Batch Norm* procedure, a technique which makes it impossible for any combination of a set of modalities to contain information corresponding to features from a strict subset of it. When tensor products are of order 2 or less, this reduces to taking applying VBN to each modality before taking their products in the tensor fusion.

Iterative Batch Norm (IBN)

- For each $I \subset \{1, 2, \ldots, M\}$, in the centering step of our Batchnorm procedure, f^I is replaced by $\hat{f}^I$ where

$$\hat{f}^I_j = \Psi((f^m_{j_m})_{m\in I}) = \sum_{\ell=0}^{|I|} (-1)^\ell \ell! \sum_{\emptyset\neq S^1\subset S^2\subset\ldots\subset S^\ell\subset I \text{ disjoint}} \prod_{m\in I\setminus(\cup S^k)} f^m_{j_m} \prod_{k\in\{1,2,\ldots,\ell\}} \mathbb{E}\left(\prod_{m\in S^k} f^m_{j_m}\right).$$

- Replace $\hat{f}^I$ by $\tilde{f}^I$ defined by

$$\tilde{f}^I = \frac{\hat{f}^I}{\sqrt{\frac{1}{B}\sum_{i=1}^{B}\sum_j (\hat{f}^I_j(x_i))^2}}.$$

Toy Experiments

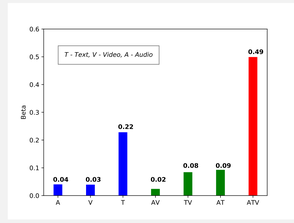
Each view is a 100-length sequence of symbols in the set $\{0, 1, 2, 3\}$ mostly generated uniformly randomly, but with some fixed information-carrying subsequences inserted at random positions.

Dataset	1	2	3	12	23	13	123
XRel							
VBN	10^{-6}	10^{-6}	-	0.99	-	-	-
DBN	0.99	10^{-6}	-	10^{-6}	-	-	-
IBN	0.99	10^{-6}	-	10^{-6}	-	-	-
XRelIrrXOR							
VBN	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}	0.99
DBN	0.24	10^{-5}	10^{-5}	10^{-5}	10^{-5}	10^{-5}	0.76
IBN	0.99	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}	10^{-6}

Multimodal Sentiment Analysis

- Multimodal Opinion-level Sentiment Intensity Corpus(MOSI)

Figure 6: Sentiment Analysis with Iterative BatchNormalisation



References

[1] Marius Kloft.
lp-Norm Multiple Kernel Learning.
PhD thesis, Berlin Institute of Technology, 2011.

[2] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency.
Tensor Fusion Network for Multimodal Sentiment Analysis.
ArXiv e-prints, July 2017.