University of Sheffield

Real-Time Robustness to Modality Corruption in Multimodal Machine Learning

Harry Woods

Supervisor: Valentin Radu

A report submitted in fulfilment of the requirements
for the degree of BSc in Computer Science

in the

Department of Computer Science

24th December 2020

Declaration

All sentences or passages quoted in this report from other people's work have been specifically acknowledged by clear cross-referencing to author, work and page(s). Any illustrations that are not the work of the author of this report have been used with the explicit permission of the originator and are specifically acknowledged. I understand that failure to do this amounts to plagiarism and will be considered grounds for failure in this project and the degree examination as a whole.

Name: _____

Signature: _____

Date: _____

Introduction

Machine learning has been an active area of research in recent years and has been deployed in a number of tasks. Machine learning models often use data from a single view, or modality, such as an array of sensors on a single device or the image stream of a video. However, research has shown that combining data from multiple modalities can increase accuracy [20, 2, 15]. Multimodal learning comes with a number of challenges, including deciding when to fuse data from each modality and dealing with failure in individual modalities.

Existing research and implementations often assume that the multimodal data supplied to the model is complete and clean. However, in real world conditions data from a particular modality or sensor may be corrupted or completely missing, resulting in inaccurate results from such a model. Multimodal models should either understand when they have been fed corrupted samples, suggesting the output could be inaccurate, or be trained in such a way that they are robust to them.

This project aims to maximise model performance in the presence of missing and corrupt modalities. This report will present existing methods of anomaly and distribution shift detection, data reconstruction, and model robustness, as well as other technologies that could be adapted for these purposes. It will also present possible next steps for exploration over the course of the project.

Contents

1 Background

Machine learning has been used over a wide range of tasks including computer vision, recommendation, medical diagnosis and human activity recognition. Shallow methods such as Random Forest and Support Vector Machines have been superseded by deep learning techniques as computational power availability has increased. Various additions to the basic feed-forward neural network have been made to better take advantage of spatial and temporal relationships between features in data sources such as images, audio and video streams, and real time sensor data.

A Convolutional Neural Network (CNN) [5] uses convolutional layers to identify spatial or temporal features. Each layer learns a number of kernels with small receptive field which are convolved over the input to produce an activation map. Kernels represent features in the input, e.g a 2D kernel in an image could represent a vertical line, with its activation map showing the location of all vertical lines in the image. Stacking multiple convolutional layers identifies more complex, larger scale spacial features in the input in the same way as the human eye. CNNs can also identify temporal features by stacking sequential frames of data and applying kernels over the temporal dimension.

State of the art image classifiers have achieved high accuracy with CNNs on general image recognition datasets such as ImageNet. One drawback of CNNs is the relatively large storage and computational requirements needed to use them during both training and inference.

Another method of identifying temporal features is with a Recurrent Neural Network (RNN). Recurrent layers contain cells which receive their output from the previous forward pass as an additional input. This allows information to flow through the network temporally, and features can be identified over multiple time steps.

An extension of this, as used in [12], is the Long Short Term Memory unit (LSTM) [6]. An LSTM consists of the cell itself and input, forget and output gates, controlling flow of information. These gates allow values into the cell, to remain in the cell, and to flow out of the cell via the activation function, respectively. Although basic recurrent cells can theoretically learn features over arbitrary timescales, finite-precision arithmetic used in computation makes identifying long term features difficult during training. LSTM cells do

not suffer from the same issues and are therefore better suited to learning over longer timescales.

Proposed multimodal architectures tend to differ on how early or late data from different modalities are brought together, known as sensor fusion [14]. On one extreme is Ensemble Classification, where a classification for each modality is obtained using an appropriate classifier. The results of these separate classifiers are used with a majority voting scheme for the final classification. The other extreme is Feature Concatenation, where features from all modalities are concatenated into a single vector used for classification. This has the advantage of being able to learn cross-modality relationships, but can hinder learning of intra-modality relationships.

A split multimodal network architecture [14] has been found to have better results than architectures that have early or late fusion. A small network for each modality is used to generate a set of features. These are then concatenated and fed into a further combined network. The modality specific networks allow the model to learn intra-modality relationships before being fused with the other modalities to learn cross modality relationships. Modout [9] is capable of learning when to share information between modalities during training.

Implementations often assume that the multimodal data supplied to the model is complete and clean. However, in real world conditions data from a particular modality or sensor may be corrupted or completely missing, resulting in inaccurate results from such a model. Methods of detecting or mitigating data modality corruption are discussed in the following section.

2 Research

2.1 Distribution shift detection

Canonical Correlation Analysis (CCA) [8] is a method for learning linear transformations of 2 random variables such that their representations are maximally correlated. In a multimodal setting, highly correlated representations would suggest that modalities are behaving the same way they did during training, whilst low correlation could suggest that one of the modalities is misbehaving. Clear limitations of this method, constraining it to linear transformations between 2 modalities, have been addressed separately by Deep CCA [1] which learns nonlinear transformations, and Generalised CCA [7] which can be applied to arbitrarily many modalities. Deep Generalised CCA (DGCCA) [3] addresses both simultaneously, being able to learn maximally correlated nonlinear representations between many views.

DGCCA representation are used for for K-nearest neighbour classification on XRMB and Twitter user datasets [3], obtaining better results than other CCA methods. Using a more complex classifier could improve these results. Alternatively, DGCCA could be used to identify corrupt modalities for removal or reconstruction before running a separate model on the resulting data.

Lipton et al. [10] present black box shift detection (BBSD), a method of detecting label shift. They show that, given any classifier f with invertible confusion matrix, to detect that the training distribution p differs from the real world distribution q it is sufficient to detect that $p(f(\boldsymbol{x})) \neq q(f(\boldsymbol{x}))$. Although created for label shift, studies show it can be applied to a variety of other shift types with good results, with [13] finding it performed better than other shift detection techniques in their comparison paper.

They note challenges with adapting BBSD to streaming scenarios. Accuracy increases as the number of samples considered increases, so best results would be obtained with a large window. However, this makes the estimate less fresh, limiting its utility for detecting short term shifts or anomalies in a single sample.

2.2 Modality reconstruction

A Cascading Residual Autoencoder (CRA) [19] can be used to impute data from missing modalities using those available. A Residual Autoencoder (RA) is trained to minimise the difference between a complete training sample and the same sample with corruption. Multiple RA's are stacked, with the sum of the input and output of each RA, the current reconstruction, used as the input for the next RA. The difference between the current reconstruction and complete data reduces after each RA and deeper CRAs produce better imputations.

CRA achieved good results on 4 datasets [19], outperforming all other matrix completion or autoencoder based imputation methods they test against. However, the datasets do not show much variation in modalities. For example in the HSFD dataset modalities are images of the same face taken in different spectral ranges, and in the RGB-D dataset the two modalities are RGB and Depth images of an object. It is unclear whether CRA could be effectively applied to modalities of significantly different types, such as audio, video, and caption data in movies.

Srivastava and Salakhutdinov [18] use a Deep Boltzmann Machine (DBM) to learn representations of multimodal data. A DBM is a multi-layered version of a Restricted Boltzmann Machine (RBM) [16], a generative neural network that can learn a joint probability distribution over its inputs. They independently pretrain a pathway for each modality and fuse them into a single representation with a joint layer, resulting in a network which performed better than other unimodal and multimodal models at the time. When modalities are missing the model is given the remaining modalities and the missing data can be generated by sampling the missing modality pathways. This generated data can be fed back into the network along with the available modalities for inference, giving higher accuracy than using the available modalities by themselves.

Unlike previous methods mentioned, the MIR Flickr dataset used for evaluation consists of two very different modalities, images and image tags. It was also adapted to use video and audio data from the CUAVE and AVLetters speech recognition datasets, achieving good results. This suggests a DBM could be a more generalizable method of multimodal learning than others and could be applied to more diverse modalities. Notably, the image features used were generated using PHOW, a shallow feature extractor from 2007, and modern image feature extraction methods such as CNNs could improve this performance.

Generative Adversarial Networks (GAN) have been used to reconstruct depth data from an RGB image [4]. A GAN consists of a Generator network and a Discriminator network. Given an input, in this case RGB satellite data, the generator attempts to create a depth map. Rather than training this network with a loss function directly, the discriminator attempts to decide whether the generated depth map is real or fake and this decision is used to train the

generator. Training is scheduled so that both networks get better at the same pace, resulting in a generator that can create realistic depth images from an RGB image.

This method does give modest gains in accuracy over using the corrupted modality or ignoring the modality completely. However, as with CRA the two modalities are similar, both representing spacial data with features in similar positions. Again it is unclear whether these could be applied generally to multiple different modalities.

2.3 Robustness

ModDrop [11] is a modality-scale regularisation method which makes predictions robust to missing or corrupted modalities. ModDrop works similarly to the node-scale regularisation method of dropout [17], where each node in a layer is temporarily removed from the network with a specified probability. Dropout ensures the network does not overfit the training data and reduces reliance of the network on individual connections. ModDrop removes entire modalities instead, having a similar effect of reducing reliance on individual modalities.

Networks trained with and without ModDrop are compared on 2 datasets, MNIST with each corner of an image considered a separate modality, and the ChaLearn LAP gesture recognition dataset augmented with their own audio data. Their augmented ChaLearn dataset contains RGB-D streams for each hand, an audio stream, and a motion captured articulated pose, giving 6 modalities of 3 distinct types. Removing individual modalities caused a mean accuracy reduction of 9%, compared to 23% when ModDrop was not used, suggesting it is a good method of dealing with missing modalities. They also obtained good results when individual modalities were corrupted with 50% pepper noise. However, this noise still retains much of the information contained in the original data, as suggested by the comparable scores without ModDrop, and may not be as challenging as the corruption encountered in the real world.

3 Analysis

The challenges involved in dealing with missing and corrupt modalities appear to be quite different. For starters, detection of a missing modality in a system is likely trivial as the data will be nonexistent, or the device missing. Corrupt data will be harder to detect as it will be presented to the model in the same way as a clean sample and these external indicators cannot be relied upon. Likewise, using ModDrop [11] during training appears to make a model largely robust to missing modalities, but not necessarily on those with corruption.

Supposing our model is robust to missing modalities, and possess a reliable method of detecting corruption, we may simply remove any corrupt modalities and allow our model to deal with the gaps. This would limit the effect of inaccurate information from that modality on the output and give the same performance as with missing modalities. However, this method would not be able to take advantage of any useful information that may remain in the corrupt modality.

A main focus of this project is therefore to detect any corruption in input modalities, ideally on a per-sample basis.

The BBSD method outlined in the previous section works well but requires many samples of data from a shifted distribution, in our case caused by corruption, to be accurate. This could detect longer term causes of corruption, for example a malfunctioning sensor, or any other changes that cause a modality to shift from the training data for a period of time. A key advantage is that it requires only an existing form of dimensionality reduction, which can be trained to best perform the task at hand. However, BBSD would be unable to detect corruption on a per-sample basis which could limit its applications in some environments.

CCA and its variants could be used for a novel method of corruption detection. If a CCA model is trained such that clean data samples are always transformed into maximally correlated representations, there is no guarantee that the representation of a corrupted modality will be correlated with them. Assuming this is true, modalities with low correlation can be considered corrupted. A possible advantage of this method over BBSD is that it may be able to detect corruption in single samples.

This method could be applied to detect and remove corruption before running the desired model on the remaining data. Alternatively, the representations learned by CCA, once corruption has been removed, could be used as features for inference. This may not perform as well as a purpose built model, but saves the computational cost of carrying out dimensionality reduction twice.

More exploration is needed to see if CCA can be used as proposed, for both corruption detection and informative dimensionality reduction, and this will be a major part of this project.

An alternative to inference without missing or corrupt modalities is to reconstruct them and use these instead of the missing data. Inference using the reconstruction methods outlined in the previous section have been found to increase accuracy over using missing modalities. It may be possible to incorporate corrupted data into the reconstruction process, allowing it to take advantage of any information that remains.

A more elegant solution than detecting corruption could be training the model so that it is robust to it. Further evaluation of ModDrop is necessary to see if it achieves this.

To evaluate any of these methods, it is necessary to understand what form corruption can take. [13] use a variety of techniques to alter the distribution of MNIST but many of the images are quite close to their usual form. [11] use pepper noise for testing, which could occur in some scenarios, but there are other forms of corruption to be addressed.

References

[1] Andrew, G., Arora, R., Bilmes, J., and Livescu, K. Deep canonical correlation analysis. In *International conference on machine learning* (2013), PMLR, pp. 1247–1255.

[2] Atrey, P. K., Hossain, M. A., El Saddik, A., and Kankanhalli, M. S. Multimodal fusion for multimedia analysis: a survey. *Multimedia systems 16*, 6 (2010), 345–379.

[3] Benton, A., Khayrallah, H., Gujral, B., Reisinger, D. A., Zhang, S., and Arora, R. Deep generalized canonical correlation analysis. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)* (2019), pp. 1–6.

[4] Bischke, B., Helber, P., König, F., Borth, D., and Dengel, A. Overcoming missing and incomplete modalities with generative adversarial networks for building footprint segmentation. *CoRR abs/1808.03195* (2018).

[5] Fukushima, K., and Miyake, S. Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In *Competition and cooperation in neural nets*. Springer, 1982, pp. 267–285.

[6] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[7] Horst, P. *Generalized canonical correlations and their application to experimental data.* No. 14. Journal of clinical psychology, 1961.

[8] Hotelling, H. Relations between two sets of variates. *Biometrika 28*, 3/4 (1936), 321–377.

[9] Li, F., Neverova, N., Wolf, C., and Taylor, G. Modout: Learning multi-modal architectures by stochastic regularization. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)* (2017), IEEE, pp. 422–429.

[10] Lipton, Z., Wang, Y.-X., and Smola, A. Detecting and correcting for label shift with black box predictors. In *International Conference on Machine Learning* (2018), pp. 3122–3130.

[11] Neverova, N., Wolf, C., Taylor, G., and Nebout, F. Moddrop: adaptive multi-modal gesture recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence 38*, 8 (2015), 1692–1706.

[12] Ordóñez, F. J., and Roggen, D. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors 16*, 1 (2016), 115.

[13] Rabanser, S., Günnemann, S., and Lipton, Z. Failing loudly: An empirical study of methods for detecting dataset shift. In *Advances in Neural Information Processing Systems* (2019), pp. 1396–1408.

[14] Radu, V., Tong, C., Bhattacharya, S., Lane, N., Mascolo, C., Marina, M., and Kawsar, F. Multimodal deep learning for activity and context recognition. *Proceedings of ACM on interactive, mobile, wearable and ubiquitous technologies 1*, 4 (2018), 1–27.

[15] Schuller, B., Valstar, M., Eyben, F., McKeown, G., Cowie, R., and Pantic, M. Avec 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction* (Berlin, Heidelberg, 2011), S. D'Mello, A. Graesser, B. Schuller, and J.-C. Martin, Eds., Springer Berlin Heidelberg, pp. 415–424.

[16] Smolensky, P. The mathematical role of self-consistency in parallel computation. In *Proceedings of the Sixth Annual Conference of the Cognitive Science Society* (1984), pp. 319–325.

[17] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research 15*, 1 (2014), 1929–1958.

[18] Srivastava, N., and Salakhutdinov, R. Multimodal learning with deep boltzmann machines. *The Journal of Machine Learning Research 15*, 1 (2014), 2949–2980.

[19] Tran, L., Liu, X., Zhou, J., and Jin, R. Missing modalities imputation via cascaded residual autoencoder. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 1405–1414.

[20] Yuhas, B. P., Goldstein, M. H., and Sejnowski, T. J. Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine 27*, 11 (1989), 65–71.