AUTHOR(S): WIJEKOON, A., WIRATUNGA, N. and SANI, S.

TITLE: Improving human activity recognition with neural translator models.

YEAR: 2018

Improving Human Activity Recognition with Neural Translator Models

Anjana Wijekoon ✉, Nirmalie Wiratunga, and Sadiq Sani

Robert Gordon University, Aberdeen AB10 7GJ, Scotland, UK
{a.wijekoon, n.wiratunga, s.a.sani}@rgu.ac.uk

Abstract. Multiple sensor modalities provide more accurate Human Activity Recognition (HAR) compared to using a single modality, yet the latter is more convenient and less intrusive. It is advantages to create a model which learns from all available sensors; although it is challenging to deploy such model in an environment with fewer sensors, while maintaining reliable performance levels. We address this challenge with Neural Translator, capable of generating missing modalities from available modalities. These can be used to generate missing or "privileged" modalities at deployment to improve HAR. We evaluate the translator with k-NN classifiers on the SelfBACK HAR dataset and achieve up-to 4.28% performance improvements with generated modalities. This suggests that non-intrusive modalities suited for deployment benefit from translators that generate missing modalities at deployment.

Keywords: Human Activity Recognition · Machine Learning · Privileged Learning

1 Introduction

Reasoning with multi-modal sensor data is an active area of research with applications fielded in multiple domains, including Human Activity Recognition(HAR), Robotics and Interactive Natural Interfaces. Typically HAR applications are related to tracking or monitoring movements such as ambulatory activities [5], activities of daily living [1] or exercises [6]. Inertial sensors and ambient sensors are mainly used in such applications to track user activity. For HAR, having multiple modalities is advantageous as it captures contextually richer representations. However access to all sensor modalities at deployment can be restricted due to ease of use or erroneous behaviours. This poses an interesting challenge of effectively deploying reasoning models with fewer modalities, compared to the number of modalities used in training.

We address this challenge as a Privileged Learning (PL) [8] problem. PL defines an additional feature space (Privileged Information) that improves classification performance, but is only available during training. It resembles how humans learn better with a teacher. In HAR we recognise this additional feature space as "Privileged Sensor Modalities", that are available during training but not after deployment. Our initial evaluations suggested that, simply ignoring privileged modalities result in poor performance. Therefore we recognise the need for estimating privileged sensor modalities at deployment. We also learnt that there is no significant linear correlation between sensor modalities, eliminating the possibility of using a simpler estimation method such as linear regression to generate the missing data. Accordingly we implementing a generative neural networks inspired translator that can learn non-linear mapping between modalities.

Recent literature suggest the use of generative models in image/video captioning [9], language translation [7] and time-series forecasting [3] with Recurrent Neural Networks (RNNs). We did not observe any advantage of using RNN as a translator given that our data has no significant temporal dependencies. Generative Adversarial Networks (GANs) is another upcoming generative model, applied successfully in image generation from random noise [2]. Yet GANs fail to generate an output influenced by the input sensor data and the class. Our architecture closely resembles Auto-encoders, which are successfully applied in audio and video reconstruction [4]. The goal of Auto-encoders is to build an abstract feature representation of given data. In contrast we focus on learning mappings between different sensor data in order to transform one to another.

We will introduce Privileged Learning for HAR and our Neural Translator in Section 2. Successive sections will present the SelfBACK Dataset, Experiment Design and Results. Finally we will discuss future improvements in Conclusion.

2 Privileged Learning with Neural Translator

We illustrate Privileged Learning (PL) for HAR referring to the two modalities of the SelfBACK dataset [1]; Wrist (W) and Thigh (T). Let X_W and X_T represent input modalities. We select X_T as the privileged sensor modality due to its intrusiveness in real life and comparatively better performance in HAR.

Figure 1 A refers to the training stage of the classification model where both sensors' data is available. Figure 1 C illustrates deployment of the classification model where only X_W is present. If we only use X_W to recognise an activity at deployment, the model performance is highly penalised. Accordingly we train a translator which learns the mapping between two sensor data streams X_W and X_T; which is done in parallel to the training of the classification model. More generally, this mapping can be between any number of sensor modalities. The input layer consists of features representing modalities that are present at test time and the output estimates the missing modalities.

Figure 1 B illustrates the Neural Translator in detail, which uses the wrist modality X_W to estimate the thigh modality $X_{\bar{T}}$. We use a fully connected neural network to estimate privileged modality, where it learns a neural mapping between its input and output layers. A single hidden layer is introduced to learn

[1] The SelfBACK project is funded by European Union's H2020 research and innovation programme under grant agreement No. 689043. More details available: http://www.selfback.eu. The dataset is publicly accessible from https://github.com/selfback/activity-recognition

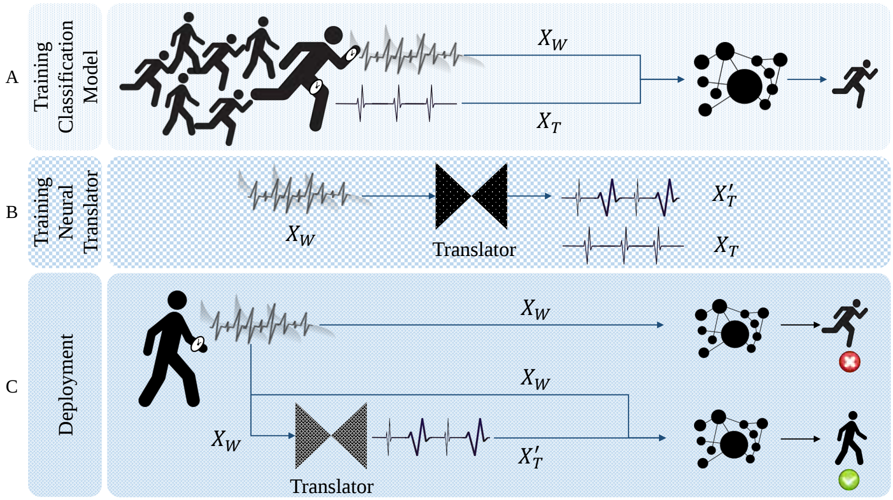

Fig. 1. Training Classification Model with Privileged Sensor Data

the feature mapping from input to the output units. During training, given an input, the network learns to generate a representation of the output modality that is as close to the actual values. This is enforced by using a loss function of Mean Squared Error (MSE) between predicted output X_{T}^{-} and expected output X_T. We will refer to the Neural Translator as T^N.

3 SelfBACK Dataset

SelfBACK Dataset is compiled with two tri-axial accelerometer data streams, belonging to 6 activity classes. Accelerometers were mounted on the right-hand wrist and thigh of each subject (thus forming 2 modalities). The data was recorded at $100Hz$ sampling rate with 34 individuals. We perform three preprocessing steps on the dataset to prepare it for the translator and the classifier. First we use a sliding window size of 3 seconds with no overlap to create instances. Next we convert three-dimensional raw data instances into single dimension Discrete Cosine Transform (DCT) feature vectors of size 180. It conveniently simplifies the task of the translator where the mapping is learnt between two abstract feature representations instead of raw data. Finally data is normalised to ensure that the k-NN classifiers are unaffected by scalar differences between different modalities.

4 Experiment Design

We use k-NN as the classifier which provides interpretable results compared to a neural network. We apply Leave-One-Person-Out (LOPO) cross validation for the classifier and the translator. We use three configurations with no privileged modalities as the baseline; and use accuracy of classification to study the contribution of the translator in the performance gains of HAR.

We experiment on different number of hidden units, while maintaining the number of hidden layers to one. We confirm that a narrow hidden layer supports learning better mappings between sensors by discarding arbitrary noise. In addition we observe that the translator is over-fitting to training data when increasing the number of hidden layers (thus increasing number of trainable parameters). Accordingly we identify the most optimal architecture for T_N as 1 hidden layer of 96 units.

We follow naming convention $f(X_i/X_j)$ to indicate a classification model trained with set of modalities X_i; and X_j are privileged modalities. Here $X_j = \in$ indicates that none of the modalities were considered as privileged after deployment. $X_j = T$ indicates that thigh is a privileged modality, and at deployment it will be estimated with Neural Translator $T_N(W/T)$ which estimates thigh data from original wrist data.

5 Results

We present baseline results of classification with no privileged modalities on Figure 2. Baseline results confirm that thigh is clearly the privileged sensor modality for the SelfBACK dataset.

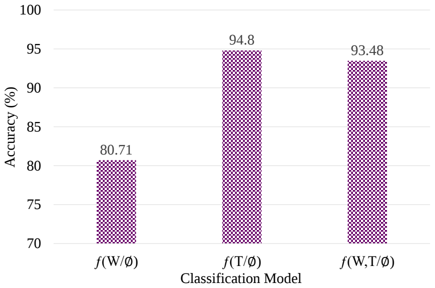

Fig. 2. Baseline classification results

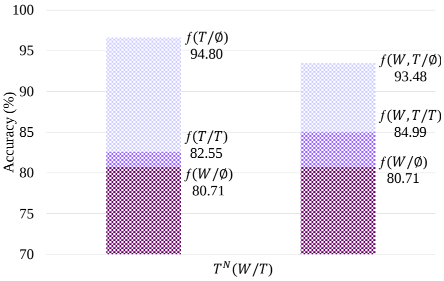

Fig. 3. Classification with T_N

Figure 3 shows classification results with $T_N(W/T)$. Here each bar shows the lower and upper bounds set by the baselines. Upper bound uses the actual data instead of the estimated after deployment; whilst the lower bound is when the privileged modality is not used for HAR. Ideally we want the translator to improve upon the lower bound to get closer to the upper.

The first bar shows that we can achieve 1.84% improvement over $f(W/\in)$ using estimated thigh on a model trained with original thigh data ($f(T/T)$). The second bar shows that the Neural Translator has significantly improved the performance by 4.28% over $f(W/\in)$ using estimated thigh data on a model trained with original wrist and thigh data ($f(W, T/T)$); brining it closer to the upper bound set by $f(W, T/\in)$.

These results suggest that a classifier trained with multiple modalities, can be used with a single or smaller subset of modalities in deployment. It is not only possible but improves performance significantly. The Neural Translator learns the non-linear correlations between input and output modalities, discarding ambiguities and noise of the source modalities. As a result the estimated modalities improve performance of the HAR classification at deployment.

6 Conclusion

We introduced the Neural Translator to improve HAR by augmenting missing modalities with estimated data. Our results show significant improvement of performance with estimated sensor data in k-NN classification. In addition to estimating privileged modalities, this versatile method can be used to augment incomplete data due to noise or technical faults. We believe there is further opportunity to improve Neural Translator with other deep learning techniques, which we plan to address in future. Finally this work demonstrates that translators can minimise sensors at deployment while improving performance which contributes towards an sustainable HAR solution.

References

1. Chavarriaga, R., Sagha, H., Calatroni, A., Digumarti, S.T., Tröster, G., Millán, J.d.R., Roggen, D.: The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. Pattern Rec. Letters **34**(15), 2033–2042 (2013)
2. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in NIPS. pp. 2672–2680 (2014)
3. Ma, X., Tao, Z., Wang, Y., Yu, H., Wang, Y.: Long short-term memory neural network for traffic speed prediction using remote microwave sensor data. Transportation Research Part C: Emerging Technologies **54**, 187–197 (2015)
4. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the ICML-11. pp. 689–696 (2011)
5. Sani, S., Massie, S., Wiratunga, N., Cooper, K.: Learning deep and shallow features for human activity recognition. In: Int. Conf. on Knowledge Science, Engineering and Management. pp. 469–482. Springer (2017)
6. Sundholm, M., Cheng, J., Zhou, B., Sethi, A., Lukowicz, P.: Smart-mat: Recognizing and counting gym exercises with low-cost resistive pressure sensing matrix. In: Proc. of the 2014 ACM Int. Joint Conf. on pervasive and ubiquitous computing. pp. 373–382. ACM (2014)
7. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: Advances in NIPS. pp. 3104–3112 (2014)
8. Vapnik, V., Vashist, A.: A new learning paradigm: Learning using privileged information. Neural networks **22**(5), 544–557 (2009)
9. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: CVPR, 2015 IEEE Conf. on. pp. 3156–3164. IEEE (2015)