

Robust and Subject-Independent Driving Manoeuvre Anticipation through Domain-Adversarial Recurrent Neural Networks

Michele Tonutti^{a,*}, Emanuele Ruffaldi^b, Alessandro Cattaneo^c,
Carlo Alberto Avizzano^c

^a*Pacmed, Amsterdam, The Netherlands*

^b*Medical Micro Instruments, Calci (PI), Italy*

^c*Teoresi Group, Assago (MI), Italy*

^d*PERCRO Laboratory, Scuola Superiore Sant'Anna, Ghezzano (PI), Italy*

Abstract

Through deep learning and computer vision techniques, driving manoeuvres can be predicted accurately a few seconds in advance. Even though adapting a learned model to new drivers and different vehicles is key for robust driver-assistance systems, this problem has received little attention so far. This work proposes to tackle this challenge through domain adaptation, a technique closely related to transfer learning. A proof of concept for the application of a Domain-Adversarial Recurrent Neural Network (DA-RNN) to multi-modal time series driving data is presented, in which domain-invariant features are learned by maximizing the loss of an auxiliary domain classifier. Our implementation is evaluated using a leave-one-driver-out approach on individual drivers from the Brain4Cars dataset, as well as using a new dataset acquired through driving simulations, yielding an average increase in performance of 30% and 114% respectively compared to no adaptation. We also show the importance of fine-tuning sections of the network to optimise the extraction of domain-independent

*Corresponding author

Email addresses: michetonu@gmail.com (Michele Tonutti),
e.ruffaldi@alumni.sssup.it (Emanuele Ruffaldi), aalessandro.cattaneo@gmail.com
(Alessandro Cattaneo), c.avizzano@santannapisa.it (Carlo Alberto Avizzano)

features. The results demonstrate the applicability of the approach to driver-assistance systems as well as training and simulation environments.

Keywords: Manoeuvre anticipation, ADAS, Deep learning, LSTM, Recurrent neural networks, Domain adaptation

1. Introduction

With 1.3 million deaths and 30 million injuries occurring yearly worldwide, road traffic accidents are the main cause of death for people aged 15-29, and represent a cost to governments of, on average, 3% of national GDPs [1]. A high proportion of those accidents occur during manoeuvres such as changing lanes and turning [2]. Advanced Driver Assistance Systems (ADAS) aim at increasing road safety by taking partial control of the car or by providing the driver with extra information when such manoeuvres could be dangerous [3]. It has been demonstrated that, thanks to recent developments in deep learning and computer vision, it is possible to predict manoeuvres a few seconds in advance and with high accuracy, by monitoring the driver's behaviour inside the vehicle and using information from the car itself (e.g. speed) and the environment (lanes configuration, presence of intersections, etc.). In particular, advances in Convolutional Neural Networks (CNN) now allow accurate extraction of head-, face-, and gaze-related features from videos [4, 5], while Recurrent Neural Networks (RNN) enable the models to take into account the temporality of an action, i.e. the order in which certain actions are performed or specific events occur [6]. However, while most recent proof-of-concept models for manoeuvre anticipation have achieved good results, little attention has been given to the problems arising from the practical implementation of such systems. One of the main concerns is their ability to generalise on subjects that were not part of the original training set. In real world applications, it is likely that ADAS installed on commercial cars will be "blind" to the driving style of new subjects. Re-training

such systems may not be feasible due to the lack of labelled examples for a new driver. While there are examples of deep neural networks trained to anticipate actions and objects in videos through unsupervised learning [7], most of the examples found in literature are not able to learn temporal relationships between features, or work exclusively with video inputs [8]. This is a problem in manoeuvre anticipation tasks, since it has been demonstrated that multi-modal inputs and the temporality of events are crucial to obtain quick and accurate predictions. Moreover, if they are not re-trained, classic deep neural networks tend to not generalize well when features in the test and training sets have different marginal distributions. Examples could include cases in which the driver has peculiar driving habits or mobility limitations, so that, for instance, they cannot turn their head fully. It is however plausible to assume that there exist common patterns and latent features in the actions of most drivers which are common regardless of the vehicle, driving style, or situation [9].

This assumption represents the basis of domain adaptation, a type of transfer learning technique which enables previously trained models to adapt to other datasets containing unlabelled observations [10]. Amongst the various approaches, domain adaptation applied to deep neural networks has been proven to be extremely effective to learn domain-invariant features from the input data even when the labels of the target distribution are unknown [11]. This enticing result is obtained by optimising a discriminative classifier while simultaneously maximising the loss of an auxiliary domain classifier. This technique has been proven to work not only with image inputs, but also with time-series observations [12]. Using this method, we hypothesize that a model can be trained to find features in sequential input data which are not only discriminative of specific manoeuvres, but also shared between the training set (large and labelled) and a fully or mostly unlabelled small test set - such as video segments of a new driver

performing unknown manoeuvres. Finding such latent features may, however, be difficult when the inputs are multi-modal time-series. This type of input is common in state-of-the-art manoeuvre anticipation models, which integrate the driver’s behaviour with information from the vehicle and the environment - such as lane configuration, car speed, and GPS data [13]. The network does not only need to learn the dependencies *within* a single time-series, but also *across* input sequences with possibly very different resolution (sparse vs. dense) or even data types (e.g. categorical vs. continuous). An effective sensory fusion approach is therefore required.

1.1. Aim

Given the problem of adaptation and generalization capabilities of multi-modal models for driver-assistance systems, we present an investigation of the application of the domain-adversarial training method to implement domain adaptation on a Recurrent Neural Network for manoeuvre anticipation. More generally, this paper proposes domain adaptation as a promising approach to improve the performance and generalization ability of machine-learning-driven ADAS. We aim at showing that domain-adversarial models are particularly beneficial in situations where either the drivers or the driving settings –or both– may differ considerably to those used to train the model.

1.2. Method and Structure

In order to evaluate our approach, we designed three experiments, which are introduced in the following paragraphs and discussed more in depth in Section 6. Before presenting the results of the experimental work, we provide a bibliographical review of the most recent published work on action prediction, maneuver anticipation, and domain adaptation. We then present a technical overview on Recurrent Neural Networks. Finally, we describe the architecture

of our models and the most salient features of both the Brain4Cars and our own dataset, including a detailed explanation of the data collection process. The experimental set-ups consist of the following:

Experiment 1 We first propose an expansion of the architecture proposed by Jain et al. [13] (Brain4Cars), which includes the driver’s gaze as an additional input, features a higher number of stacked recurrent layers, and performs enhanced sensory fusion by combining Long Short Term Memory (LSTM) and Gated Recurrent Unit (GRU) layers, amongst other improvements. We replicate the experiments presented in the original paper, training and testing the LSTM-GRU model on the Brain4Cars dataset using simple 5-fold cross-validation. This experiment is aimed at showing that the performance of our architecture is comparable to state-of-the-art models for maneuver anticipation in non-adaptive tasks. Additionally, we also test the model on a new set of driving videos obtained using an immersive virtual simulation setup, providing a performance baseline upon which to evaluate the results of the subsequent experiments.

Experiment 2 We then present a Domain-Adversarial RNN, inspired by Ganin et al. [11], in which our LSTM-GRU network from Experiment 1 serves as the feature extractor. We train and test the DA-RNN using a cross-validated, leave-one-driver-out approach on individual drivers from the Brain4Cars dataset, comparing its performance to the non-domain-adaptive LSTM-GRU trained without the target driver.

Experiment 3 Finally, we implement the same domain-adversarial approach to study how the network, trained only on the Brain4Cars data, adapts to our new dataset, in which the drivers and the driving set-up –e.g. position of the mirrors, windows, and the camera– differ from the Brain4Cars dataset.

The results confirm that, without adaptation, the model is not able to predict manoeuvres from observations in which the features have very different

marginal distributions compared to the training set. We conclude discussing the potential applications of the domain-adversarial approach to apply domain adaptation in commercial ADAS, driving training set-ups, and simulation environments.

The crucial contributions of this work can thus be summarised as follows:

- An improved LSTM-GRU Neural Network architecture for manoeuvre anticipation.
- An original dataset of observations from a driving simulation setup.
- A proof of concept for the application of a Domain-Adversarial RNN for domain adaptation on driving data, which employs the LSTM-GRU architecture to classify observations from the new dataset.

2. Related Work

The majority of studies relying on driving data do not concern themselves with the adaptation of the system to different datasets or their ability to perform well on new drivers; those who aim at anticipating manoeuvres are not an exception. Samples from all test subjects are often shuffled together before the training-test split [13]: while this enhances a classifier’s performance, it will likely cause it not to generalise and scale well in real-world applications. Once an ADAS is installed on a car, it will have been trained on a large number of drivers, but there is no guarantee that it will work well with a completely new subject without re-training. In this work we tackle this problem by applying the domain-adversarial training technique to encourage a Recurrent Neural Network to adapt to a smaller, unlabelled sets of driving data in which the features may have different marginal distributions. To obtain meaningful results we also

decided to design an improved model for manoeuvre anticipation, addressing some of the shortcomings of previous work. In this section we provide a brief survey on the most recent research on action prediction and manoeuvre anticipation, as well as on the latest techniques in the area of domain adaptation, with special focus on those applied to deep learning and sequential data. For an in-depth review of the general field of domain adaptation, we suggest the papers by Jiang [14] and Patel et al. [15]

2.1. Action Prediction

Predicting future actions differs from simple classification tasks, in that the aim is to anticipate an event with as little data as possible. One way to do this practically is by setting a threshold, processing the data at each time-step of an input sequence of observations, and only making a prediction when the probability of the corresponding output is above that threshold. To do so, a network needs to be able to remember past states and observations within a single time-series, in order to learn the temporal relations between latent features in that sequence. Most recent examples of action anticipation models do this by using LSTM- or GRU-based architectures [16, 17, 18, 19]. LSTMs and GRUs are specific types of RNN units which are able to remember and forget previous states through a number of modulating gates [20, 21]. They have been proven to solve the problem of vanishing gradients, which affects heavily "vanilla" recurrent networks, and thus are able to perform very well with long time series [22]. This aspect makes them more suitable to process sequential observations than regressive models such as Gaussian Mixtures and non-recurrent Neural Networks, or Hidden Markov Models (HMMs), which assume that each observation's probability only depends on the current state and are thus not suitable for modelling contextual effects and long sequences [8]. Despite these shortcomings, HMMs have been shown to produce promising results, especially

compared to approaches that do not employ deep learning [13]; they could therefore provide an excellent alternative in cases where the high computational requirements of RNNs cannot be met, for instance in mobile applications. A further improvement on the HMM approach is represented by Markov Decision Processes, which have been shown to provide reliable and accurate predictions for long-term driving risk inference [23].

In order to make predictions within a few time-steps, rather than simply classifying an action when the whole sequence has been processed, a recent trend has been to implement custom loss functions which exponentially increase with time. Later classification mistakes are penalised more heavily than earlier ones: the model is thus encouraged to provide a confident prediction as soon as possible. In multi-class tasks such as manoeuvre anticipation, these functions are often modifications to the standard cross-entropy loss. Aliakbarian et al. [24] included this type of time-dependent loss in a multi-stage LSTM architecture that manages to predict actions accurately with only a small percentage of video sequences, by learning context- and action-related features independently. Chan et al. [25] used the exponential loss only for positive examples in a binary classification task to predict driving accidents, also including it in a LSTM-based RNN.

2.2. Manoeuvre Anticipation

Models for manoeuvre anticipation are similar to those for action prediction; the differences are the specificity of the subject’s movements (mostly head and eyes) and the type of contextual information obtained from the environment. At the time of writing, the most recent and possibly complete approach to anticipate driving manoeuvres has been proposed by Jain et al. [13] (Brain4Cars). Their model implements an LSTM-based architecture with high-level sensory fusion to process multi-modal observations, using facial landmarks, head pose,

car speed, GPS information and lane configuration. An exponential loss function was also included, which was proven not only to successfully encourage early predictions, but also to act as a regulariser. In addition, they provided a complete dataset of driving videos, showing the driver and the outside environment in the few seconds before turns and lane changes. Because of Brain4Cars' promising results, in this work we used their architecture and dataset respectively as benchmark and for evaluation purposes.

An area of improvement identified by the Brain4Cars team lies in the addition of eye tracking information to the model's inputs. Gaze direction, in fact, has been shown to generally correlate with the direction of the subsequent movement [26, 27, 28, 29]. Including gaze direction as an additional feature vector is particularly important in the context of manoeuvre anticipation: eye movements to look in the mirrors or at objects on the road may not necessarily be accompanied by a movement of the head, but may provide information about the direction of a consequent manoeuvre. It has been argued that the choice not to implement it is justified by the difficulty of obtaining accurate measurements of gaze direction without using specialised hardware [13]. Fletcher and Zelinsky [30], for instance, successfully implemented an ADAS which analyzes driver inattentiveness using gaze tracking, and Ravichandar et al. [8] used prior probabilities based on the eye gaze to enhance the accuracy of their action-prediction model; both works were carried out using ad-hoc tracking cameras. However, recent studies have shown that it is in fact possible to estimate gaze direction accurately from videos captured by regular high definition cameras [31, 32].

While it is clear that eye tracking may provide additional benefits to infer the intentions of a driver, adding additional sensors to the model requires optimal sensory fusion in order for deep networks to perform well [33]. Indeed, a share

of the recent literature has focused on improving the integration of the multi-modal data coming from both inside and outside of the car. For instance, Doshi et al. [34] used a relevant vector machine (RVM) model to detect lane-change intent; they exploit the ability of the RVM to obtain a sparse data representation of the dataset, thus performing sensory fusion by automatically choosing discriminating features from multi-modal signals. Tawari et al. [35] developed a "Merge and Lane Change Assist" system in which sensory fusion is achieved by encoding all constraints (spatial, temporal, as well as legal) into a compact probabilistic representation [36]. Jain et al. [13] used a solution similar to those proposed by Sung et al. [37] as well as Yang and Eisenstein [38], which is not to join the features before feeding them to the network, but rather to concatenate their high-level representations through learnable layers of the neural network. This is a very simple and scalable solution that only requires simple modifications to the architecture of the network.

2.3. Domain Adaptation

Domain adaptation is a type of transfer learning where two domains (*source* and *target*) share their feature space but have different marginal distributions [12]. It attempts to solve the issue of enabling a model trained on a certain dataset (source domain) to perform well on a differently distributed dataset, of which the labels are completely or partially unknown [9]. This problem has been investigated in computer vision [39, 40, 41] and natural language processing [42] using a variety of different approaches to reduce the discrepancy between the two domains, including alignments of the subspaces [41], parameter augmentation [43], domain-invariant projection [44], and instance re-weighting [45]. Purushotham et al. [12] point out that these techniques are not able to capture the temporal dependencies in sequential data, and those which do (for instance through a Bayesian approach [46] or RNNs [47]), cannot accurately

infer non-linear relationship. Deep learning approaches, on the other hand, have proven successful in capturing time dependencies and complex, non-linear, domain-invariant relationships through domain adaptation. Examples include marginalised denoising autoencoders [48], ad-hoc CNN architectures [49], and feature embeddings [50].

Out of the deep learning approaches, one of the most elegant and easy-to-implement solutions is the Domain-Adversarial Neural Network (DANN), developed by Ganin et al. [11]. DANNs perform domain adaptation by learning domain-invariant features through a neural network architecture composed of three sections: a feature extractor, a discriminative classifier, and an adversarial domain classifier, whose loss is *maximised* through a special gradient reversal layer. The domain classifier’s role is to encourage the feature extractor to find latent representations of the features which are domain-invariant [12]. In particular, this type of domain adaptation aims at creating a common subspace for the source and target domains, so that the trained model can classify examples from the target domain without having access to the labels of the target’s training set [11]. The ability to adapt in an unsupervised manner is particularly appealing for manoeuvre anticipation models, as it would allow them to generalise well to new drivers simply by retraining them using a set of unlabelled driving videos as the target domain. While this method has been originally carried out using a CNN for the feature extraction stage, Purushotam et al. applied the technique of adversarial training on a RNN (R-DANN) and a variational RNN (VRADA), in order to capture temporal relationships from one domain to the other [12]. Our paper introduces the same concept applied to manoeuvre anticipation, expanding on the simpler R-DANN architecture and using an LSTM-GRU network as the feature extractor.

Lastly, since finding domain-invariant features in multi-modal time series can

be challenging and computationally expensive, it has been suggested that fine-tuning the feature extractor section on the source dataset can yield to an improved performance without sacrificing domain invariance [51, 11]. This has been shown to be effective even in non-adaptive cases of time-dependent features [24], and its application to the manoeuvre anticipation problem will also be demonstrated in this paper.

3. Model for Manoeuvre Anticipation

In order to apply the domain-adversarial approach to manoeuvre anticipation, we first propose an improved RNN architecture based on LSTMs and another closely-related type of gated layer, the GRU. Both prevent the gradient of the loss function from either vanishing or exploding during backpropagation thanks to the activation functions of the layers' *gates*, which are learnable and create sums of activations over which the derivatives can distribute [13, 52]. The gradient can thus propagate for a long time, allowing long time-series to be processed. Before illustrating our model, we provide a brief explanation of these two layers.

3.1. LSTM and GRU

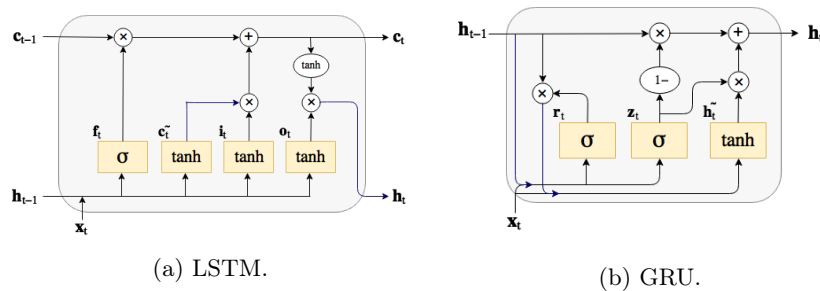


Figure 1: **Diagrams of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU).**

The structure of a typical LSTM can be seen in Fig.1a. It consists of a memory cell (\mathbf{c}), which allows information to be accumulated over long sequences, and three gates. The forget gate (\mathbf{f}) controls how the information in the memory cell is updated, deleted, or stored; the input gate (\mathbf{i}) takes in current observations and writes new values to the memory cell; and the output gate (\mathbf{o}) computes the hidden output based on the content stored in the memory cell [52]. At every time-step t of a time-series observation, the operations are computed in the following order: the activations of the input (\mathbf{i}_t) and forget gates (\mathbf{f}_t) are calculated, and the memory cell is updated (\mathbf{c}_t). Subsequently, the output of the cell is finally produced as a hidden representation (\mathbf{h}_t) depending on the activation function of the output gate (\mathbf{o}_t). The inputs into each unit are the observations (\mathbf{x}_t), the previous cell state \mathbf{c}_{t-1} , and the output \mathbf{h}_{t-1} from the LSTM at $t - 1$ [13]. The process is defined by the following equations:

$$\mathbf{i}_t = \tanh(\mathbf{W}_{xi}x_t + \mathbf{W}_{hi}h_{t-1} + \mathbf{W}_{ci}c_{t-1} + \mathbf{b}_i) \quad (1)$$

$$\mathbf{f}_t = \sigma(\mathbf{W}_{xf}x_t + \mathbf{W}_{hf}h_{t-1} + \mathbf{W}_{cf}c_{t-1} + \mathbf{b}_f) \quad (2)$$

$$\mathbf{o}_t = \tanh(\mathbf{W}_{xo}x_t + \mathbf{W}_{ho}h_{t-1} + \mathbf{W}_{co}c_{t-1} + \mathbf{b}_o) \quad (3)$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tanh(\mathbf{W}_{xc}x_t + \mathbf{W}_{hc}h_{t-1} + \mathbf{b}_c) \quad (4)$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t) \quad (5)$$

where \mathbf{W}_* are the weights and \mathbf{b}_* the biases. \odot is the Hadamart product, also known as element-wise or point-wise vector product. [53, 13].

The GRU, pictured in Fig.1b, is similar to the LSTM, but is lacking the memory

cell and the output gate [54]. It is defined by the following equations:

$$\mathbf{r}_t = \sigma(\mathbf{W}_{xr}x_t + \mathbf{W}_{hr}h_{t-1} + \mathbf{b}_r) \quad (6)$$

$$\mathbf{z}_t = \sigma(\mathbf{W}_{xz}x_t + \mathbf{W}_{hz}h_{t-1} + \mathbf{b}_z) \quad (7)$$

$$\mathbf{h}_t = \tanh(\mathbf{W}_{xh}x_t + \mathbf{W}_{hh}(\mathbf{r}_t \odot \mathbf{h}_{t-1}) + \mathbf{b}_h) \quad (8)$$

$$\tilde{\mathbf{h}}_t = \mathbf{z}_t \odot \mathbf{h}_{t-1} + (1 - \mathbf{z}_t) \odot \tilde{\mathbf{h}}_t \quad (9)$$

As it can be seen, the number of operations computed at each time-step is lower than for the LSTM. It was also found that GRUs outperform LSTMs in specific situations, for instance when no dropout is used [55, 53]. This can be explained by a lower tendency to overfit due to the reduced complexity. While the introduction of dropout makes LSTM better choices in most situations, using GRUs lowers computation and training time by reducing the number of learnable parameters. They thus represent a valuable option.

Throughout this paper, LSTM and GRU operations will be referred to as:

$$(\mathbf{h}_t, \mathbf{c}_t) = \text{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}) \quad (10)$$

$$\mathbf{h}_t = \text{GRU}(\mathbf{x}_t, \mathbf{h}_{t-1}) \quad (11)$$

3.2. Anticipation Framework

To build our model we followed the framework for manoeuvre anticipation proposed and defined by Jain et al. [13], which we summarise here. The inputs to the model at training time are represented by N time series in the form of $\{(\mathbf{x}_1, \dots, \mathbf{x}_T)_i, \mathbf{y}_i\}_{i=1}^N$, in which \mathbf{x}_t is the set of features at time-step t , and i is the index of the sequence. $\mathbf{y} = [y^1, \dots, y^J]$ is the representation of the action occurring at the end of the sequence when $t = T$, with y_j standing for the probability of the sequence leading to event j . In our case, $J = 5$: four manoeuvres (turning right, turning left, lane change right, lane change left),

plus going straight as the default action. During training, each manoeuvre is represented by a one-hot-encoded vector. At test time, an observation vector \mathbf{x}_t is received by the model at every time-step. A probability threshold p_{th} is chosen, so that when, and only when, any $y^j \geq p_{th}$, a prediction of the respective manoeuvre will be made. At $t = T$, if no manoeuvre has been predicted with sufficient confidence, the default action (going straight) will be predicted.

The inputs \mathbf{x} consist of the matrices of the head features (composed of facial features and head pose), $\phi = [\phi_1, \dots, \phi_T]$; of the gaze, $\gamma = [\gamma_1, \dots, \gamma_T]$; and of the environmental features, $\eta = [\eta_1, \dots, \eta_T]$. Each element ϕ_t , γ_t , and η_t is a vector containing the respective individual features. The feature engineering and data processing pipelines are described more in detail in Section 5.

3.3. Network Architecture

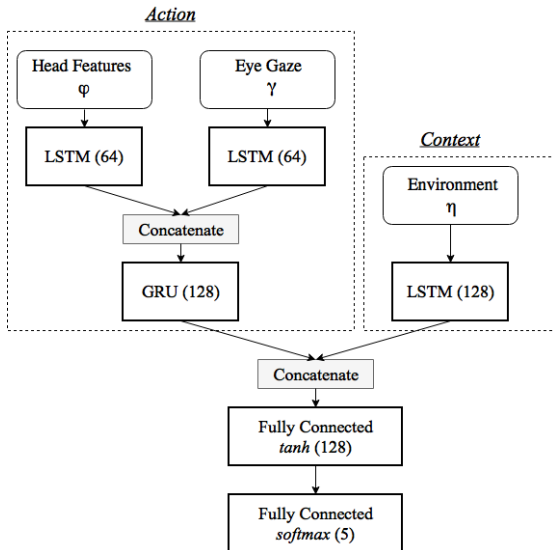


Figure 2: **Architecture of our model.** Sensory fusion is performed by learning the concatenation of higher-level representations of the features at different points in the network.

The structure of our architecture, shown in Fig.2, was inspired by the sensory fusion approach proposed in previous work, in which the inputs are not

concatenated before going through the network, but rather after the first recurrent layer(s) as high-level features [13, 37, 38]. We enhanced this idea by incorporating the concept of action- and context-dependent features proposed by Aliakbarian et al. [24]. They showed that, for a similar type of anticipation problem, this structure performs better than two parallel recurrent layers which disregard the action-context distinction. The concatenation occurs at two different places in the network: first the representations of the head features ϕ_t are concatenated with those of gaze γ_t , after each of them has gone through an LSTM layer (Eq.12 and 13). Their concatenation is then passed through a GRU layer (Eq.14). The rationale for this choice is as follows: when an event can be classified in a main type (manoeuvre vs. going straight, i.e. no manoeuvre) and a sub-type (each of the four manoeuvres), a two-layered RNN architecture enables each of the two layers to specialise in the two predictions. This idea was proposed and successfully tested by Xiao et al. [56]. In our case, the first LSTM learns the main event type, while the following GRU learns which manoeuvre is performed. A GRU layer was preferred to a second LSTM in order to limit the increase in complexity.

The output of the GRU layer is then concatenated with the hidden output of a third LSTM layer whose inputs are the environmental features η_t (Eq.15). This concatenation is then fed to a fully-connected dense layer (Eq.16) [13]. Finally, the last softmax layer provides the classification probabilities for the five classes (Eq.17).

$$(\mathbf{h}_t^\phi, \mathbf{c}_t^\phi) = \text{LSTM}_\phi(\phi_t, \mathbf{h}_{t-1}^\phi, \mathbf{c}_{t-1}^\phi) \quad (12)$$

$$(\mathbf{h}_t^\gamma, \mathbf{c}_t^\gamma) = \text{LSTM}_\gamma(\gamma_t, \mathbf{h}_{t-1}^\gamma, \mathbf{c}_{t-1}^\gamma) \quad (13)$$

$$(\mathbf{h}_t^a) = \text{GRU}_a([\mathbf{h}_t^x; \mathbf{h}_t^g], \mathbf{h}_{t-1}^a, \mathbf{c}_{t-1}^a) \quad (14)$$

$$(\mathbf{h}_t^\eta, \mathbf{c}_t^\eta) = \text{LSTM}_\eta(\eta_t, \mathbf{h}_{t-1}^\eta, \mathbf{c}_{t-1}^\eta) \quad (15)$$

$$\mathbf{z}_t = \tanh(\mathbf{W}_f[\mathbf{h}_t^a; \mathbf{h}_t^\eta] + \mathbf{b}_f) \quad (16)$$

$$\mathbf{y}_t = \text{softmax}(\mathbf{W}_y \mathbf{z}_t + \mathbf{b}_y) \quad (17)$$

Because of the high number of learnable parameters ($> 18 \times 10^4$), dropout was applied to both the recurrent and dense layers of the network. A dropout of 0.6 was set to the LSTM and GRU recurrent connections (*recurrent* dropout), while a dropout of 0.7 was set to the output of every layer [57, 58]. Additionally, the bias of the recurrent layers was initialised to 1, in order to improve the model’s performance and training [53, 22].

In order to encourage the model to anticipate early, we implemented the loss function shown in Eq.18, which weighs each term of the cross-entropy categorical loss with an exponential term, as a function of time. The loss will be greater at larger values of t , thus rewarding the network for predicting the right class as early as possible [24, 25, 13]. y_t^j represents the probability of event j computed by the model at time-step t .

$$L_y = \sum_{j=1}^N \sum_{t=1}^T -e^{-0.9(T-t)} \log(y_t^j) \quad (18)$$

The network was trained through backpropagation with gradient-based optimization; we chose the Adam optimiser [59] for its simplicity, robustness, and low computational cost [60]. The learning rate was set to 1×10^{-3} ; the other

parameters were set to the values suggested in the original paper: $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 1 \times 10^{-8}$, without temporal decay. We implemented additional regularization by using early stopping, with a patience of 80 epochs.

4. Domain-Adversarial RNN

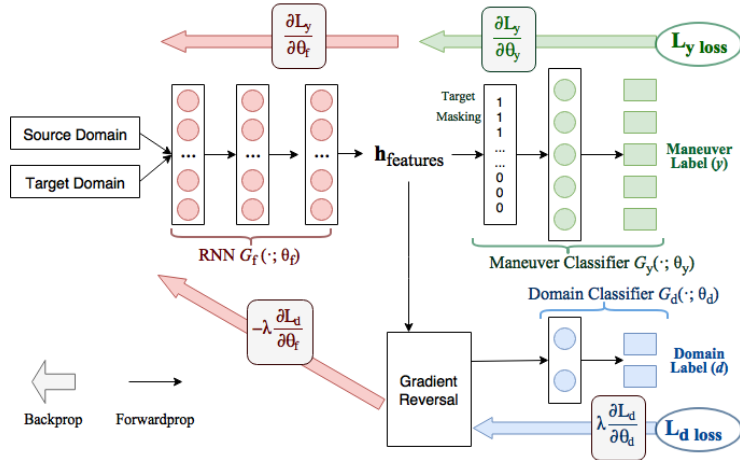


Figure 3: **DA-RNN**. The proposed architecture for domain-adversarial training of our model to achieve domain adaptation. *Left* (red): the feature extractor, consisting in the LSTM-GRU network shown in Fig.2, without the last softmax layer. *Top right* (green): manoeuvre classifier. *Bottom right* (blue): Domain classifier. Diagram style inspired by Ganin et al. [11]. (Better viewed in colour.)

Our DA-RNN, pictured in Fig.3, is made up of three sections, divided in two main branches. The first section employs the LSTM-GRU network described in Section 3.3 (except for the last softmax layer) as a feature extractor. Its role is to learn the latent relationships between the features in the observation sequences. After the latent features are extracted, their hidden representation is fed into two branches: the first one is a discriminative classifier for manoeuvre anticipation, composed of a single softmax layer; the second one is an adversarial domain classifier. The key part of the latter is the *gradient reversal* layer: during the forward pass, the input is left unchanged, while during backpropagation, the gradient is negated and multiplied by a constant, λ (lower section of Fig.3). The

loss of the domain classifier is thus maximised, thereby encouraging the feature extractor to find representations of the features which are domain-invariant [12]. Conversely, the label classifier's loss is minimised in order for the features to also be discriminative of the manoeuvres. The manoeuvre loss L_y is the time-dependent exponential function in Eq.18, while the domain loss L_d is the regular binomial cross-entropy loss function. The total loss is given by Eq.19; the effect of λ is discussed more in detail in Section 6.

$$L_{tot} = L_y - \lambda L_d \quad (19)$$

To train the model, half of each input batch is filled with samples from the source domain, and half from the target domain [11]. Since the latter, by assumption, contains fewer observations, target samples are used more than once in each epoch. This should not unfairly help the manoeuvre classifier, since L_y is only dependent on samples from the source domain: the observations belonging to the target domain are "hidden" from the manoeuvre classifier by assigning them a loss weight of 0 through the boolean mask (represented by the "target masking" block in Fig.3). The loss contribution from the samples of the source domain is left unchanged by assigning them a loss weight of 1. This approach is equivalent to computing L_y first using only transformed samples from the source domain, and then computing L_d using the combined batches. However, our method allows for the weight updates to be computed in a single forward- and back-pass, significantly cutting down training time. At test time, inference is made on samples from the target domain not included in the training set, removing the boolean mask from the model. The performance is evaluated only on the prediction of the manoeuvre classifier.

5. Data Collection and Processing

5.1. Brain4Cars Dataset

The dataset used as a benchmark for testing our anticipation architecture was developed by Brain4Cars [13]. At the time of writing and to our knowledge, it is the most complete dataset which includes synchronized recordings of the driver’s upper body and the road in front of the car. It consists of 700 observations, each including a pair of videos with a duration of 5 seconds: one showing the driver’s face inside the car, and the other one the road ahead, outside of the car. The videos are recorded at 30 frames-per-seconds, for a total 150 frames per video. Additional data is provided for each frame, including lane configuration, speed, and presence of intersections ahead of the car. Every observation is associated with a manoeuvre, performed at the end of the 5 seconds. The numbers of observations for each manoeuvre is as follows: {Going straight = 234, Changing Lane Left = 124, Changing Lane Right = 58, Turning Left = 123, Turning Right = 55}. The format of the features in the dataset is similar to that illustrated in Section 5.3 and 5.4. For a more in-depth explanation, we refer the reader to the original work by Jain et al. [13]. The dataset is publicly available on the Brain4Cars website.¹

5.2. Our Dataset

In order to build our own dataset for a more extensive evaluation of our model and the domain adaptation approach, we created a driving simulator setup following a standard structure for research [61, 62] and the same video format of the Brain4Cars data. While the Brain4Cars dataset represents an excellent resource, we believed it necessary to have access to a dataset where *both* the drivers and driving conditions are different from the Brain4Cars dataset in order

¹<https://brain4cars.com>

to evaluate the domain-adaptation capabilities of our model. In practice, the goal is to mimic a real-life situation in which a commercial product may be applied to vehicles that are substantially different in size and/or position of mirrors, conducted by unseen drivers. A driving simulator was chosen in order to collect data quickly and in a controlled environment, using a combination of highway and city driving conditions. Similar virtual simulation setups have been frequently used for validation of ADAS and statistical approaches to driving style predictions [63, 64, 65].

We used the commercial game Euro Truck Simulator 2², including an unofficial modification which allows the player to drive a car instead of a truck. This specific game has been chosen among other driving simulator software due to the realism of the physics engine, the quality of the graphical output on the main view and rear mirrors, as well as the compatibility with the instrumentation. The setup, pictured in Fig.4, consisted of a three-screen high-resolution display system (CPU Intel i7-3930K and GPU NVidia GTX 960) and driving equipment with pedals, gears, and a steering wheel with force-feedback (G27, Logitech, CH). A commercial web-cam (HD Pro C920, Logitech, CH, at resolution 1080p), placed on the middle screen, was pointed towards the face of the driver. The three-screen setup allowed to have a field-of-view of about 200°. Car data from the game engine was obtained using an unofficial telemetry server³, whose output are *JSON* blocks containing a large number of information such as orientation, speed, acceleration, steering angle, etc.

Five (5) different subjects were asked to play for a time between 30 and 50 minutes, driving both in cities and on highways. They were asked to respect road rules and to behave as if they were in a real car (looking at mirrors, turning their

²SCS Software, 2012, <https://eurotrucksimulator2.com/>

³<https://github.com/Funbit/ets2-telemetry-server>



Figure 4: **The setup for our data collection process.**

head when needed, etc.) The game screen and the driver’s face were recorded synchronously using the Open Broadcaster Studio Open Source software ⁴. We aimed at minimizing the variance in the data caused by differences in the camera placement, as well as avoiding time intervals in which the tracking does not capture the facial landmarks due to the camera being partially covered by the driver’s hands on the wheel. We therefore positioned our camera always in the same position on top of the central monitor, directly in front of the driver seat.

In order to match the Brain4Cars dataset format, we extracted 5-second videos of the desired manoeuvres using the Boris Open Source video annotation tool⁵ [66] to annotate the manoeuvres in the raw footage. A custom Python script was written to cut up the 5 seconds preceding the onset of each manoeuvre (defined as the moment when the wheel touches the lane markings or when it starts turning at the intersection [13]), using a combination of Open Source OpenCV library⁶ and FFmpeg video conversion tool⁷. The final dataset comprises 113 videos: {Going straight = 32, Changing Lane Left = 21, Changing Lane Right = 19, Turning Left = 24, Turning Right = 17}. The number of observations is smaller than those in the Brain4Cars dataset, as its aim is to provide the target samples for a domain adaptation problem; in a real-life situation, target sets

⁴<https://obsproject.com/>

⁵<http://www.boris.unito.it/>

⁶<http://opencv.org/>

⁷<https://www.ffmpeg.org/>

will always be considerably smaller than source sets.

Our dataset has been made available as Open Access on Zenodo with the following DOI: <http://dx.doi.org/10.5281/zenodo.1009540>.

5.3. Feature Extraction and Processing

To extract features from the videos of both datasets we used OpenFace, an open-source toolkit that provides facial landmark tracking, head pose estimation, and gaze tracking, from videos and images, using a combination of Conditional Local Neural Fields (CLNF) and CNNs [32]. It is capable of high performance real-time tracking using regular cameras, making it an attractive option for ADAS. Once the data was extracted from the videos, we followed the feature processing pipeline defined by Jain et al. [13], with slight modifications to binning intervals, labelling of the environmental features, and feature scaling. We were interested in obtaining the movements of the facial landmarks, the head pose, the direction of the gaze, as well as environmental informations – namely lane configuration, presence of intersections ahead in near proximity, and speed of the car. A time-series is constructed for each sample in the datasets, with each frame representing a time-step in the observation sequence \mathbf{x}_t . Data analysis and processing, feature engineering were carried out in Python 3.

5.4. Action-related Features

5.4.1. Facial Landmarks and Head Pose

We represent the movement of the driver’s head with the motion of the facial landmarks, using a binning approach. We took the velocity of each of the 68 landmarks between consecutive frames, calculating the horizontal motion as $\delta x^{face} = x_t^{face} - x_{t-1}^{face}$, in pixels and in the image space; and the angular motion in the $x - y$ plane as $\theta^{face} = \arctan2(\delta y^{face}, \delta x^{face})$, in radians. These values were binned to create histogram features. Six (6) bins were chosen for

the horizontal motion:

$$\{\delta x \leq -5; -5 < \delta x \leq -2.5; -2.5 < \delta x \leq 0; 0 < \delta x \leq 2.5; 2.5 < \delta x \leq 5; \delta x > 5\}$$

pixels, while for the angular motion we used four (4) bins:

$$\{0 < \theta \leq \pi/2; \pi/2 < \theta \leq \pi; \pi < \theta \leq 3\pi/2; 3\pi/2 < \theta \leq 2\pi\} \text{ radians.}$$

Negative values refer to motions towards the left-hand side of the image and the right-hand side of the driver.

The head pose was also included using the Euler angles representation $[R_x(\alpha_{pitch}), R_y(\alpha_{yaw}), R_z(\alpha_{roll})]$ as it was shown to improve the performance [13, 67]. Overall, the histogram features and head pose form the head features $\phi \in \mathbb{R}^{13}$.

5.4.2. Eye Gaze

Information about the driver’s gaze was also obtained through OpenFace’s output, which provides a 3D direction vector for each eye, normalised and in world coordinates. We took the average components of both eyes, and applied a Butterworth low-pass filter (4th order, sampling frequency of 30 Hz, cutoff of 1.66 Hz) in order to eliminate noise deriving from tracking inaccuracies and natural saccades. The filter introduces a minimal delay which was proven to not be detrimental to the model, and it can be applied in real time. To create the feature vector we take the x and y component of the gaze direction vector (which correspond to the horizontal and vertical direction in the image plane), scaled between -1 and 1, to create histogram features, similarly to the head features. The bins chosen are:

$$\{-1 < \delta x \leq -0.5; -0.5 < \delta x \leq 0; 0 < \delta x \leq 0.5; 0.5 < \delta x \leq 1\} \text{ pixels, in}$$

the image space. Identical bins were chosen for δy . In this case we used the direction components rather than the inter-frame velocity since we found that it correlates better with the driver’s intention. We define $\gamma \in \mathbb{R}^8$ as the gaze features.

5.5. Context-related Features (Environment)

The environmental context is expressed by $\boldsymbol{\eta} \in \mathbb{R}^4$. The first two features are boolean variables expressing the presence of a lane to the left and to the right of the car, respectively. The third feature is also boolean, and indicates the presence of an intersection ahead and in the near proximity of the car. For the purpose of this study, these three values were labelled manually for each observation. In a practical implementation, this information could be extracted automatically through lane detection algorithms [68, 69] fused with GPS data. Finally, the fourth value is the speed of the car in km/h. This value is provided in the Brain4Cars dataset, and was measured by the physics engine of the game in our dataset.

6. Experiments

In this section we present the results of the experiments conducted on the Brain4Cars’ and our dataset, using the features extracted according to the method described in Section 5. We first tested the network for manoeuvre anticipation on the two datasets separately, shuffling samples from all drivers before the training-test split, as done by Jain et al. [13]. We show that the new architecture and additional features yield to an improvement in the performance of the model. We then investigated the applications of domain adaptation to our model via domain-adversarial training, in two ways: a) employing a leave-one-out approach on each driver in the Brain4Cars dataset, in order to study the possibility of personalising a model in a partly-unsupervised manner; b) using the entirety of the Brain4Cars dataset as the source domain and our dataset as the target domain, in order to study how the approach works with different feature distributions. In both cases, we attempted to increase the performance of the models through fine-tuning –meaning initializing the weights of certain

layers with weights from a pre-trained model, and we studied the effect of varying the value of the domain loss multiplier λ .

To evaluate the results, four measures were used. Three of them provide the multi-class classification score for the predicted action, namely: precision, recall and F1 score [70]. The fourth measure is the time-to-prediction (TTP), representing how many seconds before the onset of the manoeuvre the prediction is made. When calculating precision, recall, and F1 score, the "going straight" predictions are not considered, as it is considered the baseline state. The probability threshold was set to 0.9 for all experiments, since this value was shown to yield the most confident and quick predictions [13]. This means that at each time-step t , a prediction is made only if at least one of the 5 outputs of the models is ≥ 0.9 . We trained all models with batches of size 128.

6.1. Experiment 1: Manoeuvre Anticipation

The first experiment was performed using the LSTM-GRU architecture described in Section 3.3 on the Brain4Cars dataset, without domain adaptation, in order to evaluate its performance as a feature extractor. Samples from all drivers were shuffled, and a test set was set aside taking 15% of the total observations. 5-fold cross validation was used, with the 5 validation subsets each making up a different 20% of the training set. Training and validation sets were normalised jointly before the split. Following Jain et al.'s approach, we augmented the training set by extracting subsequences of random length (T_{sub}) from the original observation, with $50 < T_{sub} < 150$. These additional samples were used as additional training examples, thereby adding redundancy and additionally decreasing the risk of over-fitting [13]. More sub-samples were taken from under-represented classes, thus balancing the class ratios. For the Brain4Cars dataset, this led to a total of 2160 training samples; 312 for ours. The results are shown in Table 1. We compare our performance with Jain et

al.’s LSTM model [13]. It can be seen that our network performs better for all manoeuvres, reaching a higher F1 score as well as a higher time-to-prediction. We theorise that this improvement is due partly to the additional information given by the gaze direction. Indeed, excluding the gaze-related features, the performance of the model was observed to drop by up to 1-1.5%. However, we attribute the improvement especially to the enhanced sensory fusion (including the action-context structure). Indeed, a parallel structure in which the hidden representations of all three inputs are concatenated after the initial recurrent layers performs marginally worse than Brain4Cars’ model. This shows that simply adding more features does not necessarily improve the performance of a model if it is not accompanied by rational modifications in the architecture.

In addition, it was noticed that the added complexity of the network caused a strong tendency to overfit, despite the high dropout, the early-stopping criteria, and the redundancy in the training data. However, it is likely that the number of samples is not large enough; bigger datasets should alleviate this problem and further enhance the model’s performance. Lastly, we report that including the car’s speed as part of the environmental features was detrimental to the performance of the model, and, for this reason, it was excluded in the subsequent experiments. This may be due to the fact that, especially in sequences leading to lane changes or driving straight, maneuvers may not be correlated with specific ranges of speed values; further analysis should be carried out on the subject. However, we observed that excluding any of the other features used by Jain et al. [13] from either dataset worsens the performance considerably. This was confirmed experimentally through 5-fold cross-validation.

The same model was also tested on our dataset, using a similar procedure. The results, illustrated in Table 2, show that the performance is lower than for the Brain4Cars dataset. This is most likely due to the lower number of samples,

	Changing lane				Turning				All manoeuvres			
	Prec. (%)	Recall (%)	F1 (%)	TTP (s)	Prec. (%)	Recall (%)	F1 (%)	TTP (s)	Prec. (%)	Recall (%)	F1 (%)	TTP (s)
B4C	95.4	85.7	88.8	3.42	68.5	78.5	72.1	3.78	82.0	82.1	82.0	3.58
B4C w/ Head pose	/	/	/	/	/	/	/	/	90.5	87.4	88.3	3.16
LSTM-GRU (ours)	96.5	90.5	93.6	3.90	91.1	90.9	91.0	4.06	92.3	90.8	91.3	3.98

Table 1: **Manoeuvre anticipation on the Brain4Cars dataset.** Results of the non-adaptive LSTM-GRU network on a test set comprised of observations from all drivers. Results reported from Brain4Cars (B4C) are taken directly from the paper [13].

and the fact that in a simulation setting, drivers tend to do less emphatic head and eye movements. Conversely, fine-tuning the network by pre-training its weights using the Brain4Cars dataset improves the performance.

	Prec. (%)	Recall (%)	F1 (%)	TTP (s)
No Pre-Training	82.0	82.1	82.0	3.9
Pre-trained on B4C	89.4	92.2	90.8	4.1

Table 2: **Manoeuvre anticipation on our dataset.** Results of the non-adaptive LSTM-GRU network on a test set comprised of observations from all drivers, without and with fine-tuning.

Overall, these results show that our model represents an improvement over the most recent LSTM-based approach in manoeuvre anticipation, and prove that our dataset can be used reliably with a comparable performance. Moreover, it was shown that fine-tuning can be a powerful approach to enhance the model’s capability if a larger, more reliable dataset is available.

6.2. Experiment 2: Domain Adaptation on Different Drivers

In order to investigate how a trained model can learn to adapt to a small set of unlabelled driving videos belonging to a subject not included in the training set, we manually separated the videos of each individual driver in the Brain4Cars dataset from each other. We then ran our DA-RNN using observations from all drivers except one as the source domain, while, as the target domain, we used the samples from the remaining driver. Once trained, the model was tested on a set of samples from the target domain not included in the training set, all from the same driver. This was done for each driver separately, and the results were

averaged. In addition, we fine-tuned the networks by initializing the weights of the LSTM-GRU feature extractor by pre-training them on the source domain. The results are shown in Table 3.

	Prec. (%)	Recall (%)	F1 (%)	TTP (s)
No adaptation	60.9	57.7	58.0	3.8
DA-RNN	71.7	66.5	68.1	3.9
DA-RNN w/ Fine Tuning	77.7	75.6	76.8	3.8

Table 3: **Performance of the DA-RNN**, using a leave-one-out approach on the Brain4Cars dataset, averaged over 6 drivers. ($\lambda = 1.10$)

From the results it is clear that when the test driver is not part of the training set, the model does not perform well without adaptation. However, when the model is trained with the domain-adversarial training, an absolute improvement of more than 10% can be observed. A further improvement is obtained by pre-training the weights of the feature extractors on the source domain, yielding an increase in performance of almost 20% compared to the non-adapting model. This is because, by initializing the weights with values that are known to produce accurate predictions, the network is already trained to find features that are discriminative of the manoeuvres.

Additionally, we found that the hyperparameter λ , which is the constant multiplier of the domain classifier’s loss during backpropagation, plays a key role in feature extraction. The higher its value, the higher the influence of the domain classifier loss, meaning a stronger push towards domain invariance in the feature extractor. The network will therefore tend to find features that are shared by the two domains, but which are not necessarily discriminative. A small λ , on the other hand, will cause the extracted features to be less domain-invariant but more effective to classify the maneuvers in the source domain samples. For our experiments, we found that a value ≈ 1 , meaning an equal weighting of the losses from the two classifiers yielded the best performance.

6.3. Experiment 3: Domain Adaptation on Our Dataset

The third experiment was performed to attempt a transfer of information between the Brain4Cars data and our dataset. Being able to adapt a model to observations in which the features are distributed very differently can be important in applications such as driver training, Virtual Reality simulations, and autonomous driving research. Fig.5 shows that the two distributions are indeed very different: the head movements made by the drivers in a real-world driving scenario have a higher within-manoeuvre variance than those driving a simulator, but a smaller variance between the different manoeuvres. The fact that the subjects were asked to perform the same movements as in a real-setting, however, makes it plausible that it is possible to find latent features which are shared by both datasets. Moreover, the fact that fine-tuning the model using one dataset improves the performance on the other provides additional support to this theory. Table 4 illustrates the results of three different approaches: 1. Training and validating the model on the Brain4Cars dataset, and testing it on our dataset (no adaptation); 2. Training the DA-RNN using the Brain4Cars dataset as source domain and our dataset as target domain; 3. Training the

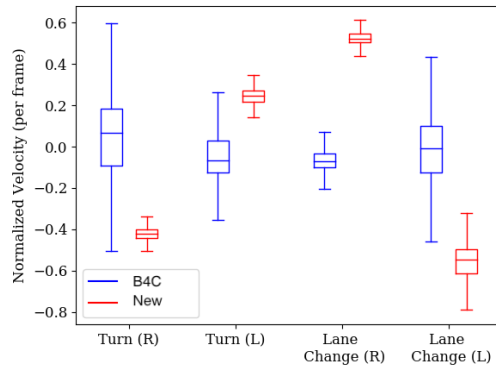


Figure 5: **Marginal distributions of the Brain4Cars and our datasets.** Box plots of the horizontal velocity of the facial landmarks for both datasets.

DA-RNN using the Brain4Cars dataset as source domain and our dataset as target domain, with fine-tuning (i.e. initializing the weights of the feature extractor by pre-training our LSTM-GRU network on the Brain4Cars dataset.)

	Prec. (%)	Recall (%)	F1 (%)	TTP (s)
No adaptation	27.3	31.5	29.0	3.8
DA-RNN	47.5	38.7	42.6	4.0
DA-RNN w/ Fine Tuning	72.6	55.1	62.7	4.0

Table 4: **Performance of the DA-RNN**, with the Brain4Cars dataset as source domain and our dataset as the target domain. ($\lambda = 1.10$)

The first striking observation is the extremely poor performance of the model without adaptation. We infer that when the marginal distribution of the features in two domains are very different, a non-adaptive model fails to generalise. The DA-RNN, in comparison, performs better, with an increase in F1 score of 12 percentage points and an increase in time-to-prediction of 0.2s. The most critical improvement was registered when the feature extractor was fine-tuned on the source domain, with an F1 score of 62.7%. The overall performance is still lower than the case of Experiment 2, when the drivers belonged to the same dataset (i.e. driving in similar conditions), but much higher than the no-adaptation case. These relative improvements in performance are comparable to the ones found in recent works performing similar domain adaptation tasks, such as the ones reported in the original DANN paper by Ganin et al. [11]. Overall, these results confirm the hypothesis that there exist latent features shared by datasets of observations in which similar driving tasks are performed, but in largely different settings; moreover, they highlight the need of an adaptive approach for the practical implementations and personalizations of ADAS.

The models were created, trained and tested using Python 3, using the deep learning framework Keras 2.0 with the TensorFlow 1.2 backend. Network

training was run on a machine with CPU Intel Core i7-7700K and GPU NVidia GTX 1080, and 32GB of memory.

The code of the architecture of the DA-RNN can be found at https://github.com/michetonu/DA-RNN_manoeuver_anticipation.

7. Conclusions

In this paper we proposed and tested a Domain-Adversarial Recurrent Neural Network for adaptive driving manoeuvre anticipation. Trained on a large source dataset of driving observations, our DA-RNN is able to adapt to smaller, unlabelled sets of observations by maximizing the loss of a domain classifier used as an auxiliary output. To extract domain-invariant features from multi-modal time-varying data, we designed a multi-stage LSTM-GRU architecture based on Ganin et al.’s DANN [11], which uses a CNN as feature extractor, and Purushotham et al.’s R-DANN [12], which instead uses vanilla RNN layers. In order to apply it to the problem of manoeuver anticipation, we expanded on the work done by Jain et al.[13], adding eye gaze direction to the set of input features used to predict driving actions. An alternative approach to carry out advanced sensory fusion is implemented by learning the concatenation of the hidden representations of the features through recurrent layers. The LSTM-GRU section of the network was proven to outperform state-of-the-art work on non-adaptive manoeuvre anticipation tasks. We also present a new dataset obtained through a driving simulation set-up, and made it available for public use.

The evaluation of the DA-RNN was carried out initially using a leave-one-out approach, in which the observations of each individual driver in the Brain4Cars dataset was left out from the training set and used as target domain. An increase of 17 percentage points in F1 score was registered when the DA-RNN’s

feature extractor was pre-trained on the source domain. The results show the potential of domain-adversarial training to adapt models to new drivers without the need to retrain them with additional labelled examples. In a real-world scenario, observations to be used as the target domain could be captured automatically during the first drive, eliminating the necessity for manual labelling. The second evaluation consisted in using the Brain4Cars dataset as the source domain and our dataset as the target domain, with the same feature space but a different feature distribution. We reported an improvement in F1 score by 114% (32.7 percentage points) compared to the non-adaptive case, when the model’s weights are fine-tuned. Overall, we demonstrate that the domain-adversarial approach represents a promising approach to increase the flexibility and generalization capabilities of commercial ADAS through domain adaptation.

We conclude that non-adaptive models are not able to generalise well in contexts a) where the target driver was not part of the training set, and b) where the features in the target domain have a very different marginal distribution, which is the case when the driving set-up in the target set differs from the one used to collect the training data. Adaptive models will therefore be necessary for ADAS installed in commercial vehicles, and will prove helpful in virtual simulations and driving training tasks. In order to further validate the approach in the context of assisted driving, additional tests should be conducted on a larger target dataset obtained in real-life conditions. Larger and more diverse driving video datasets will enable even higher performances, as one of the bottlenecks of our approach was found in the network’s tendency to overfit, due to the complexity of the architecture and the relatively small size of the datasets. This work sets the bases for further research aimed at enabling adaptive deep neural networks to reach performances comparable to fully supervised models.

8. Acknowledgments, Contributions and Funding

We would like to thank the members of the PERCRO Lab who volunteered for the data collection process, namely L. Peppoloni, G. Dabisias, L. Landolfi, and F. Brizzi. A huge thank you also to J. von Kügelgen for the technical support on domain adaptation and thorough proofreading of the manuscript. Finally, we would like to extend our appreciation towards the creator of all the software used in this work, as well as to the authors of the Brain4Cars dataset for making it available online for free. We strongly believe in open-source software and open access to data; we therefore decided to also publish and share our dataset for free use.

Study conception and design: MT, ER, CAA. Acquisition of data: MT, AC. Analysis and interpretation of data: MT, ER. Drafting of manuscript: MT, ER, CAA.

The activities related to the results in this manuscript have been supported within the program of the Excellence Department on Robotics and Artificial Intelligence. The related project is supported by the National Ministry for Education and Research (MIUR). The authors are grateful to Scuola Superiore Sant’Anna, the TeCIP institute and the Department above for the offered logistic, technical and financial support.

9. Copyright Notice

©2019. This manuscript version is made available under the CC-BY-NC-ND 4.0 license, which can be found at the following URL:
<http://creativecommons.org/licenses/by-nc-nd/4.0/>

10. Bibliography

References

- [1] A. Paul, R. Chauhan, R. Srivastava, M. Baruah, Advanced Driver Assistance Systems, SAE Technical Paper 2016-28-0223doi : 10.4271/2016-28-0223.
- [2] H. H. Hurt, J. V. Ouellet, D. R. Thom, Motorcycle accident cause factors and identification of countermeasures: Appendix, Tech. rep., National Highway Traffic Safety Administration (1981).
- [3] J. Levinson, J. Askeland, et al., Towards fully Autonomous driving: systems and algorithm, IEEE Intelligent Vehicles Symposium (2011) 163–168.
- [4] Q. V. Le, W. Y. Zou, S. Y. Yeung, A. Y. Ng, Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, Cvpr (2011) 3361–3368doi : 10.1109/CVPR.2011.5995496.
- [5] A. Karpathy, G. Toderici, et al., Large-scale video classification with convolutional neural networks, IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2014) 1725–1732doi : 10.1109/CVPR.2014.223.
- [6] M. Hermans, B. Schrauwen, Training and Analyzing Deep Recurrent Neural Networks, NIPS (2013) 190–198.
- [7] C. Vondrick, H. Pirsiavash, A. Torralba, Anticipating the future by watching unlabeled videoarXiv: 1504.08023v1.
- [8] H. C. Ravichandar, A. Kumar, A. P. Dani, K. R. Pattipati, Learning and Predicting Sequential Tasks Using Recurrent Neural Networks and Multiple Model Filtering, in: AAAI Fall Symposium, 2016, pp. 331–337.
- [9] S. Ben-David, J. Blitzer, Analysis of representations for domain adaptation, Advances in Neural Information Processing Systems 19 (2007) 137–144.
- [10] S. Ben-David, J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, J. W. Vaughan, A theory of learning from different domains, Machine Learning 79 (1-2) (2010) 151–175. doi : 10.1007/s10994-009-5152-4.
- [11] Y. Ganin, E. Ustinova, H. Ajakan, et al., Domain-adversarial training of neural networks, Journal of Machine Learning Research 17 (2016) 1–35.
- [12] S. Purushotham, W. Carvalho, T. Nilanon, Y. Liu, Variational Recurrent Adversarial Deep Domain Adaptation, International Conference on Learning Representations (2016) 1–11.
- [13] A. Jain, A. Singh, H. S. Koppula, S. Soh, A. Saxena, Recurrent neural networks for driver activity anticipation via sensory-fusion architecture, in: IEEE ICRA, 2016, pp. 3118–3125. doi : 10.1109/ICRA.2016.7487478.

- [14] J. Jiang, A literature survey on domain adaptation of statistical classifiers 2007 (March).
URL http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/da_survey.pdf
- [15] V. Patel, R. Gopalan, Visual Domain Adaptation: A survey of recent advances, *IEEE Sig Proc Mag* 02138 (2015) 1–36. doi : 10.1109/MSP.2014.2347059.
- [16] J. C. B. Gamboa, Deep Learning for Time-Series Analysis *arXiv*: 1701.01887.
- [17] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal lstm with trust gates for 3d human action recognition, in: *ECCV*, Springer, 2016, pp. 816–833.
- [18] K. Greff, R. K. Srivastava, J. Koutnik, B. R. Steunebrink, J. Schmidhuber, LSTM: A Search Space Odyssey, *IEEE TNNLS* doi : 10.1109/TNNLS.2016.2582924.
- [19] M. Baccouche, F. Mamalet, C. Wolf, Sequential deep learning for human action recognition, *Proc. Int. Conf. Human Behavior Understanding (HBU)* (2011) 29–39 doi : 10.1007/978-3-642-25446-8.
- [20] S. Hochreiter, J. Schmidhuber, Long short-term memory, *Neural Comput.* 9 (8) (1997) 1735–1780. doi : 10.1162/neco.1997.9.8.1735.
- [21] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling (2014) 1–9.
- [22] F. A. Gers, J. Schmidhuber, F. Cummins, Learning to Forget: Continual Prediction with LSTM, *Neural Computation* 12 (10) (2000) 2451–2471. doi : 10.1162/089976600300015015.
- [23] M. Shimosaka, T. Kaneko, K. Nishi, Modeling risk anticipation and defensive driving on residential roads with inverse reinforcement learning, in: *2014 IEEE 17th International Conference on Intelligent Transportation Systems (ITSC)*, 2014, pp. 1694–1700.
- [24] M. S. Aliakbarian, F. Saleh, M. Salzmann, B. Fernando, L. Petersson, L. Andersson, Encouraging LSTMs to Anticipate Actions Very Early *arXiv*: 1703.07023.
- [25] F.-H. Chan, Y.-T. Chen, Y. Xiang, M. Sun, Anticipating accidents in dash-cam videos, in: *Asian Conference on Computer Vision*, 2016, pp. 136–153.
- [26] J. R. Flanagan, R. S. Johansson, Action plans used in action observation, *Nature* 424 (6950) (2003) 769–771. doi : 10.1038/nature01861.

- [27] G. Gredebäck, T. Falck-Ytter, Eye movements during action observation, *Perspectives in Cognitive Science* 10 (5) (2015) 591–598. doi : 10.1177/1745691615589103.
- [28] Y. Matsumoto, J. Heinzmann, A. Zelinsky, The essential components of human-friendly robot systems, in: *International Conference on Field and Service Robotics*, 1999, pp. 43–51.
- [29] C. Miyajima, K. Takeda, Driver-Behavior Modeling Using On-Road Driving Data: A new application for behavior signal processing, *IEEE Sig Proc Mag* 33 (6) (2016) 14–21. doi : 10.1109/MSP.2016.2602377.
- [30] L. Fletcher, A. Zelinsky, Driver Inattention Detection based on Eye Gaze—Road Event Correlation, *IJRR* 28 (6) (2009) 774–801. doi : 10.1177/0278364908099459.
- [31] C. Gou, Y. Wu, K. Wang, K. Wang, F. Y. Wang, Q. Ji, A joint cascaded framework for simultaneous eye detection and eye state estimation, *Pattern Recognition* 67 (2017) 23–31. doi : 10.1016/j.patcog.2017.01.023.
- [32] T. Baltrušaitis, P. Robinson, L.-P. Morency, Openface: an open source facial behavior analysis toolkit, in: *IEEE WACV*, 2016, pp. 1–10.
- [33] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng, Multimodal Deep Learning, *Proceedings of the 28th International Conference on Machine Learning (ICML)* (2011) 689–696doi : 10.1145/2647868.2654931.
- [34] A. Doshi, B. T. Morris, M. M. Trivedi, On-road prediction of driver’s intent with multimodal sensory cues, *IEEE Pervasive Computing* 10 (3) (2011) 22–34. doi : 10.1109/MPRV.2011.38.
- [35] A. Tawari, S. Sivaraman, et al., Looking-in and looking-out vision for Urban Intelligent Assistance: Estimation of driver attentive state and dynamic surround for safe merging and braking, *IEEE IV (Iv)* (2014) 115–120. doi : 10.1109/IVS.2014.6856600.
- [36] S. Sivaraman, M. M. Trivedi, Dynamic probabilistic drivability maps for lane change and merge driver assistance, *IEEE TITS* 15 (5) (2014) 2063–2073. doi : 10.1109/TITS.2014.2309055.
- [37] J. Sung, S. H. Jin, A. Saxena, Robobarista: Object Part based Transfer of Manipulation Trajectories from Crowd-sourcing in 3D Pointclouds (2015) 1–16doi : 10.1007/978-3-540-48113-3.
- [38] X. Yang, P. Ramesh, R. Chitta, S. Madhvanath, E. A. Bernal, J. Luo, Deep Multimodal Representation Learning from Temporal DataarXiv: 1704.03152.
- [39] K. Saenko, B. Kulis, et al., Adapting visual category models to new domains, in: *Lecture Notes in Computer Science*, Vol. 6314 LNCS, 2010, pp. 213–226. doi : 10.1007/978-3-642-15561-1{_}16.

- [40] B. Gong, Y. Shi, F. Sha, K. Grauman, Geodesic flow kernel for unsupervised domain adaptation, in: *IEEE CVPR, 2012*, pp. 2066–2073. doi : 10.1109/CVPR.2012.6247911.
- [41] B. Fernando, A. Habrard, M. Sebban, T. Tuytelaars, Unsupervised visual domain adaptation using subspace alignment, in: *IEEE ICCV, 2013*, pp. 2960–2967. doi : 10.1109/ICCV.2013.368.
- [42] G. Foster, C. Goutte, R. Kuhn, Discriminative instance weighting for domain adaptation in statistical machine translation, in: *Proc. of Empirical Methods in Natural Language Processing, 2010*, pp. 451–459.
- [43] Y. Watanabe, K. Hashimoto, Y. Tsuruoka, Domain Adaptation for Neural Networks by Parameter Augmentation [arXiv:1607.00410](#).
- [44] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, M. Salzmann, Unsupervised domain adaptation by domain invariant projection, in: *IEEE ICCV, 2013*, pp. 769–776. doi : 10.1109/ICCV.2013.100.
- [45] J. Jiang, C. Zhai, Instance Weighting for Domain Adaptation in NLP, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics (October) (2007)* 264–271. doi : 10.1145/1273496.1273558.
- [46] F. Huang, A. Yates, Distributional representations for handling sparsity in supervised sequence-labeling, in: *Joint ACL and AFNLP, 2009*, pp. 495–503.
- [47] R. Socher, C. Lin, Parsing natural scenes and natural language with recursive neural networks, *ICML (2011)* 129–136doi : 10.1007/978-3-540-87479-9.
- [48] M. Chen, K. Q. Weinberger, F. Sha, L. Angeles, Marginalized Denoising Autoencoders for Domain Adaptation, *ICML (2012)* 767–774doi : 10.1007/s11222-007-9033-z.
- [49] E. Tzeng, J. Hoffman, T. Darrell, K. Saenko, Simultaneous deep transfer across domains and tasks, in: *IEEE ICCV, Vol. 2015 Inter, 2015*, pp. 4068–4076. doi : 10.1109/ICCV.2015.463.
- [50] Y. Yang, J. Eisenstein, Unsupervised Domain Adaptation with Feature Embeddings [arXiv:1412.4385](#).
- [51] M. Long, J. Wang, Y. Cao, J. Sun, P. S. Yu, Deep learning of transferable representation for scalable domain adaptation, *IEEE KDE 28 (8) (2016)* 2027–2040. doi : 10.1109/TKDE.2016.2554549.
- [52] I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*, MIT Press, 2016.
- [53] R. Jozefowicz, Z. Wojciech, I. Sutskever, An Empirical Exploration of Recurrent Network Architectures, in: *ICML, Vol. 37, 2015*, pp. 2342–2350.

- [54] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine TranslationarXiv: 1406. 1078.
- [55] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Gated feedback recurrent neural networks, ICML 37 (2015) 2067–2075. doi : 10. 1145/2661829. 2661935.
- [56] S. Xiao, J. Yan, M. Farajtabar, L. Song, X. Yang, H. Zha, Joint Modeling of Event Sequence and Time Series with Attentional Twin Recurrent Neural NetworksarXiv: 1703. 08524.
- [57] Y. Gal, Z. Ghahramani, A Theoretically Grounded Application of Dropout in Recurrent Neural Networks YarinarXiv: 1512. 05287v5.
- [58] T. Bluche, C. Kermorvant, J. Louradour, Where to apply dropout in recurrent neural networks for handwriting recognition?, ICDAR 2015-Novem (c) (2015) 681–685. doi : 10. 1109/ICDAR. 2015. 7333848.
- [59] D. P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization (2014) 1–15arXiv: 1412. 6980.
- [60] S. Ruder, An overview of gradient descent optimization algorithms (2016) 1–14arXiv: 1609. 04747.
- [61] W. Wang, J. Xi, X. Li, Statistical Pattern Recognition for Driving Styles Based on Bayesian Probability and Kernel Density Estimation (2016) 1–10.
- [62] M. Bergamasco, S. Perotti, et al., Fork-lift truck simulator for training in industrial environment, IEEE ETFA (2005) 689–693.
- [63] Y. Hwang, D. Yoon, H. S. Kim, K.-H. Kim, A validation study on a subjective driving workload prediction tool, IEEE Transactions on Intelligent Transportation Systems 15 (4) (2014) 1835–1843. doi : 0. 1109/TITS. 2014. 2334664.
- [64] W. Han, W. Wang, X. Li, J. Xi, Statistical-based approach for driving style recognition using bayesian probability with kernel density estimation, IET Intelligent Transport Systems 13 (1) (2019) 22–30. doi : 10. 1049/i et- i ts. 2017. 0379.
- [65] W. Wang, J. Xi, C. Liu, X. Li, Human-centered feed-forward control of a vehicle steering system based on a driver’s path-following characteristics, IEEE Transactions on Intelligent Transportation Systems 18 (6) (2017) 1440–1453. doi : 10. 1109/TITS. 2016. 2606347.
- [66] O. Friard, M. Gamba, BORIS: a free, versatile open-source event-logging software for video/audio coding and live observations, Methods in Ecology and Evolution 7 (11) (2016) 1325–1330. doi : 10. 1111/2041-210X. 12584.

- [67] M. M. Trivedi, T. Gandhi, J. McCall, Looking-in and looking-out of a vehicle: Selected investigations in computer vision based enhanced vehicle safety, in: IEEE ICVES, 2005, pp. 29–34. doi : 10.1109/ICVES.2005.1563609.
- [68] G. Kaur, D. Kumar, Lane Detection Techniques: A Review, International Journal of Computer Applications 112 (10) (2015) 975–8887.
- [69] Z. Kim, Robust lane detection and tracking in challenging scenarios, in: IEEE TITS, Vol. 9, 2008, pp. 16–26. doi : 10.1109/TITS.2007.908582.
- [70] M. Sokolova, G. Lapalme, A systematic analysis of performance measures for classification tasks, Information Processing & Management 45 (4) (2009) 427–437.