Transportation Research Part C 79 (2017) 1–17

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Deep learning for short-term traffic flow prediction

Nicholas G. Polson [a], Vadim O. Sokolov [b,*]

[a] *Booth School of Business, University of Chicago, Chicago, IL 60637, USA*
[b] *Systems Engineering and Operations Research, George Mason University, Fairfax, VA 22030, USA*

ARTICLE INFO

Article history:
Received 28 April 2016
Received in revised form 25 February 2017
Accepted 28 February 2017
Available online 17 March 2017

Keywords:
Traffic Flows
Deep Learning
Trend filtering
Sparse linear models

ABSTRACT

We develop a deep learning model to predict traffic flows. The main contribution is development of an architecture that combines a linear model that is fitted using ℓ_1 regularization and a sequence of tanh layers. The challenge of predicting traffic flows are the sharp nonlinearities due to transitions between free flow, breakdown, recovery and congestion. We show that deep learning architectures can capture these nonlinear spatio-temporal effects. The first layer identifies spatio-temporal relations among predictors and other layers model nonlinear relations. We illustrate our methodology on road sensor data from Interstate I-55 and predict traffic flows during two special events; a Chicago Bears football game and an extreme snowstorm event. Both cases have sharp traffic flow regime changes, occurring very suddenly, and we show how deep learning provides precise short term traffic flow predictions.

1. Introduction

1.1. Traffic flow prediction

Real-time spatio-temporal measurements of traffic flow speed are available from in-ground loop detectors or GPS probes. Commercial traffic data providers, such as Bing maps (Microsoft Research, 2016), rely on traffic flow data, and machine learning to predict speeds for each road segment. Real-time (15–40 min) forecasting gives travelers the ability to choose better routes and authorities the ability to manage the transportation system. Deep learning is a form of machine learning which provides good short-term forecasts of traffic flows by exploiting the dependency in the high dimensional set of explanatory variables, we capture the sharp discontinuities in traffic flow that arise in large-scale networks. We provide a variable selection methodology based on sparse models and dropout.

The goal of our paper is to model the nonlinear spatio-temporal effects in recurrent and non-recurrent traffic congestion patterns. These arise due to conditions at construction zones, weather, special events, and traffic incidents. Quantifying travel time uncertainty requires real-time forecasts. Traffic managers use model-based forecasts to regulate ramp metering, apply speed harmonization, and regulate road pricing as a congestion mitigation strategy; whereas, the general public adjusts travel decisions on departure times and travel route choices, among other things.

Deep learning forecasts congestion propagation given a bottleneck location, and can provide an accurate forty minute forecasts for days with recurrent and non-recurrent traffic conditions. Deep learning can also incorporate other data sources,

* Corresponding author.
E-mail address: vsokolov@gmu.edu (V.O. Sokolov).

http://dx.doi.org/10.1016/j.trc.2017.02.024

such as weather forecasts, and police reports to produce more accurate forecasts. We illustrate our methodology on traffic flows during two special events; a Chicago Bears football game and an extreme snow storm event.

To perform variable selection, we develop a hierarchical sparse vector auto-regressive technique (Dellaportas et al., 2012; Nicholson et al., 2014) as the first deep layer. Predictor selection then proceeds in a dropout (Hinton and Salakhutdinov, 2006). Deep learning models the sharp discontinuities in traffic flow are modeled as a superposition of univariate non-linear activation functions with affine arguments. Our procedure is scalable and estimation follows traditional optimization techniques, such as stochastic gradient descent.

The rest of our paper is outlined as follows. Section 1.2 discusses connections with existing work. Section 1.3 reviews fundamentals of deep learning. Section 2 develops deep learning predictors for forecasting traffic flows. Section 3 discusses fundamental characteristics of traffic flow data and illustrates our methodology with the study of traffic flow on Chicago's I-55. Finally, Section 4 concludes with directions for future research.

1.2. Connections with existing work

Short-term traffic flow prediction has a long history in the transportation literature. Deep learning is a form of machine learning that can be viewed as a nested hierarchical model which includes traditional neural networks. Karlaftis and Vlahogianni (2011) provides an overview of traditional neural network approaches and (Kamarianakis et al., 2012) shows that model training is computationally expensive with frequent updating being prohibitive. On the other hand, deep learning with dropout can find a sparse model which can be frequently updated in real time. There are several analytical approaches to traffic flows modeling (Anacleto et al., 2013; Blandin et al., 2012; Chiou et al., 2014; Polson and Sokolov; Polson and Sokolov, 2015; Work et al., 2010). These approaches can perform very well on filtering and state estimation. The caveat is that they are hard to implement on large scale networks. Bayesian approaches have been shown to be efficient for handling large scale transportation network state estimation problems (Tebaldi and West, 1998). Westgate et al. (2013) discusses ambulance travel time reliability using noisy GPS for both path travel time and individual road segment travel time distributions. Anacleto et al. (2013) provides a dynamic Bayesian network to model external intervention techniques to accommodate situations with suddenly changing traffic variables.

Statistical and machine learning methods for traffic forecasting are compared in Smith and Demetsky (1997). Sun et al. (2006) provides a Bayes network algorithm, where the conditional probability of a traffic state on a given road, given states on topological neighbors on a road network is calculated. The resulting joint probability distribution is a mixture of Gaussians. Bayes networks for estimating travel times were suggested by Horvitz et al. which eventually became a commercial product that led to the start of Inrix, a traffic data company. Wu et al. (2004) provides a machine-learning method support vector machine (SVM) (Polson and Scott, 2011) to forecast travel times and (Quek et al., 2006) proposes a fuzzy neural-network approach to address nonlinearities in traffic data. Rice and van Zwet (2004) argues that there is a linear relation between future travel times and currently estimated conditions with a time-varying coefficients regression model to predict travel times.

Integrated auto-regressive moving average (ARIMA) and exponential smoothing (ES) for traffic forecasting are studied in Tan et al. (2009) and Van Der Voort et al. (1996). A Kohonen self-organizing map is proposed as an initial classifier. Van Lint (2008) addresses real-time parameter learning and improves the quality of forecasts using an extended Kalman filter. Ban et al. (2011) proposes a method for estimating queue lengths at controlled intersections, based on the travel time data measured by GPS probes. The method relies on detecting discontinuities and changes of slopes in travel time data. Ramezani and Geroliminis (2015) combines the traffic flow shockwave analysis with data mining techniques. Oswald et al. (2000) argues that non-parametric methods produce better forecasts than parametric models due to their ability to better capture spatial-temporal relations and non-linear effects. Vlahogianni et al. (2014) provides an extensive recent review of literature on short-term traffic predictions.

There are several issues not addressed in the current literature (Vlahogianni et al., 2014). First, predictions at a network level using data-driven approaches. There are two situations when a data-driven approach might be preferable to methodologies based on traffic flow equations. Estimating boundary conditions is a challenging task, which even in systems that rely on loop detectors as traffic sensors are typically not installed on ramps. Missing data problems are usually addressed using data imputation (Muralidharan and Horowitz, 2009) or weak formulations of boundary conditions (Strub and Bayen, 2006). Our results show that a data-driven approach can efficiently forecast flows without boundary measurements from ramps. Another challenge with physics-based approaches comes from their limited ability to model urban arterials. For example, Qiao et al. (2001) shows analytical approaches fail to provide good forecasts. Another challenge is to identify spatio-temporal relations in flow patterns, Vlahogianni et al. (2014) for further discussion. Data-driven approaches provide a flexible alternative to physical laws of traffic flows.

The challenge is to perform model selection and residual diagnostics (Vlahogianni et al., 2014). Model selection can be tackled by regularizing the loss function and using cross-validation to select the optimal penalty weight. To address this issue, when we specify our deep learning model we construct an architecture as follows. First we use is a regularized vector autoregressive model to perform predictor selection. Then, our deep learning model addresses the issue of non-linear and non-stationary relations between variables (speed measurements) using a series of activation functions.

Breiman (2003) describes the trade-off between machine learning and traditional statistical methods. Machine learning has been widely applied (Ripley, 1996) and shown to be particularly successful in traffic pattern recognition. For example,