Amsterdam to Dublin Eventually Delayed?
LSTM and Transfer Learning for Predicting Delays of Low Cost Airlines

Nicholas McCarthy
Accenture Labs, Dublin, Ireland
nicholas.mccarthy@accenture.com

Mohammad Karzand
Accenture Labs, Dublin, Ireland
mohammad.karzand@accenture.com

Freddy Lecue
CortAIx Thales, Montreal, Canada
Inria, Sophia Antipolis, France
freddy.lecue@inria.fr

Abstract

Flight delays impact airlines, airports and passengers. Delay prediction is crucial during the decision-making process for all players in commercial aviation, and in particular for airlines to meet their on-time performance objectives. Although many machine learning approaches have been experimented with, they fail in (i) predicting delays in minutes with low errors (less than 15 minutes), (ii) being applied to small carriers i.e., low cost companies characterized by a small amount of data. This work presents a Long Short-Term Memory (LSTM) approach to predicting flight delay, modeled as a sequence of flights across multiple airports for a particular aircraft throughout the day. We then suggest a transfer learning approach between heterogeneous feature spaces to train a prediction model for a given smaller airline using the data from another larger airline. Our approach is demonstrated to be robust and accurate for low cost airlines in Europe.

Introduction

Delay is an important indicator of the quality of service in any transportation system. It drives the decision-making process of all players in the commercial aviation industry. It is particularly important for airlines as it is a key factor for measuring their on-time performance. Many machine learning approaches have been applied previously to the problem of predicting delay, however these generally suffer from poor performance when predicting delays in minutes with low error (less than 15 minutes), and when applied to small carriers, i.e. low-cost companies characterized by a small amount of data. Our approach has demonstrated to be robust and accurate for low cost airlines in Europe.

The commercial aviation industry defines delay as the duration of time that a flight is late or postponed, and considers any flight that arrives more than 15 minutes past its scheduled gate arrival time as delayed. In 2013 over 30% of flights were delayed by more than five minutes in Europe and by more than fifteen minutes in the US, (ANAC 2017) and (CODA Digest 2017).

The impact of flight delay on airlines is multifold, ranging from clearly defined outcomes such as compensation owed to passengers, late fines, and increased operational costs, to more intangible consequences for airline brand perception and customer loyalty. As a result of this delay, airlines suffer penalties, fines, higher rates of customer complaints, and additional operation costs such as crew and aircraft retention in airports.

Towards these issues, our contribution is twofold:

Contribution 1: In this work we propose an approach for training an LSTM network to model the flight delay. We model sequences of flights for a given aircraft over a period of time with trips to multiple airports, and aim to find a good estimator of future flight delay by learning the context of past flight delay from historic airline data.

Contribution 2: As low cost airline companies only hold a small volume of data due to the small fleet size, limited routes or crew, we suggest a transfer learning approach which uses data sourced from other airlines and domains to improve the performance of the initial LSTM model.

It appears intuitive that learning an estimator to predict the delay for multiple airlines should be easier than learning each in isolation. We show that the model trained on both data sets outperforms the model trained just on the smaller data set.

In the following section we give a summary of different flight delay prediction methods and transfer learning approaches over heterogeneous feature spaces, which are highly relevant to our application. Section 3 presents how we addressed the problem of predicting sequences of flight delays under very limited data constraints. Section 4 reports experimental results and lessons learned from low cost airline companies in Europe. Finally we draw some conclusions and discuss future work.

Related work

We review literature related to our application area, i.e., the domains of route / flight delay prediction and propagation, as well as recent work on transfer learning in heterogeneous feature spaces.

Route Delay: Flight delay prediction can be modeled in many ways, depending on the objectives of the research. The use of machine learning for analysis of flight systems has become increasingly prevalent, particularly in classification and prediction applications.

Previous methods applied to the problem of predicting airplane route delay include the k-Nearest Neighbour algorithm (Zonglei, Jiandong, and Guansheng 2008), random forests (Rebollo and Balakrishnan 2014), adaptive networks based on fuzzy inference systems (Khanmohammadi et al. 2014), and Markov decision processes incorporating a reinforcement learning strategy (Balakrishna et al. 2008). These systems report good performance when the prediction is a single instance that is close in time, but note a concurrent decrease in accuracy as the forecast horizon grows. We address this problem by considering sequence modeling of flight delays.

Delay Propagation: The primary objective in delay propagation is to understand how delay propagates along sequences of flights and through airports, based on the assumption that an initial delay has already occurred in the transportation system due to a previous flight. We consider a particular scenario in which delays are spread to other flights of the same aircraft as chain reactions, as has been previously studied (Abdelghany et al. 2004) (Wong and Tsai 2012). Although it is important to understand the stability and the reliability of the recovery of the carriers from the delay propagation as studied in (Wu 2005) and (Dück et al. 2012), we focus on the dynamics of delay propagation through sequence-to-sequence learning.

Transfer learning with heterogeneous feature spaces: Multi-view representation learning approaches aim at learning from heterogeneous "views" (feature sets) of multimodal parallel datasets. Previous work in this field include canonical correlation Analysis (CCA) based methods (Dhillon, Foster, and Ungar 2011), via auto-encoder regularization in deep networks (Wang et al. 2015), translated learning (Dai et al. 2009), Hybrid Heterogeneous Transfer Learning (HHTL) (Zhou et al. 2014), etc., all of which require source-target correspondent parallel instances meaning the feature spaces are the same.

Previously, (Seungwhan Moon 2017) and (Moon and Carbonell 2016) study a transfer learning framework where source and target datasets are heterogeneous in both feature and label spaces. As they do not assume explicit relations between source and target tasks apriori, their method provides a way to control what to and what not to transfer from source knowledge. This method, termed "Attentional Heterogeneous Transfer", is very similar to the method we use in this paper.

Flight Delay Prediction

This section presents the problem formation for flight delay prediction, sequence representation for fitting an LSTM model, and the adaption and extension for transfer learning to address the problem of small data in low-cost airlines.

Problem Formulation

Motivation: In initial settings and reviewed state-of-the-art approaches, a delay of a flight is considered as a unique and independent event. However it can be affected by many factors, and in particular temporal elements such as the late arrival of a previous flight, and in turn affects the on-time departure of succeeding flights. Such behavior, called the ripple or propagation effect in airline industry, is a strong motivation to model the aircraft delay problem as a sequential prediction problem. Thus, given a sequence of past flights for a particular aircraft, we predict the departure delay for subsequent flights in the sequence.

Recurrent neural networks (RNN) are well designed for predicting sequences over time, and are a natural fit for this problem. In particular the LSTM variant of RNNs, which have have gained traction in recent years, nicely fit the characteristics of the problem.

Problem Statement: We consider the prediction task $P = \{X, Y\}$, with the target task features $X = \{x^{(i)}\}_{i=1}^{N}$ for $x \in \mathcal{R}^M$, where (i) N is the sample size, (ii) M is the feature dimension, and (iii) $Y = \{y^{(i)}\}_{i=1}^{N}$ is the groundtruth predictions where $y \in \mathcal{R}$. We also assume a successive prediction scheme, meaning that the prediction function $f_p^0(x^{(i)}, y^{(i-1)}, y^{(i-2)}, \cdots)$ is function of current $x^{(i)}$ as well as all the previous predicted values $y^{(j)}, \quad j < i$. We might also construct series of predictors to predicts for the future values of $y^{(j)}, \quad j > i$, denoted by $f_p^{j-i}(x^{(i)}, y^{(i-1)}, y^{(i-2)}, \cdots)$.

Modeling: We could model this problem as a one shot prediction problem, feeding back the predicted delay to the model to observe the propagation of the delay. However, the benefit in using the LSTM cell is the propagation of the state of the cell which enables the model to propagate an insight into the reason behind the delay.

Flight Delay Sequencing

As flight information is stored in a transactional database, we initially required a sequencing algorithm in order to determine what sequences to train the LSTM model with.

In order to achieve this we use aircraft tail number and scheduled departure date and time to form a unique identifier for sequencing. However various sequencing approaches could be applied, strongly impacting results. We report the two best sequencing methods (based on empirical experimentation) applied below.

24-hour Sequencing: In this method we define a sequence as a 24 hour time period using a midnight cut-off point. Scheduled departure time is used as the sequence point: flights are considered part of that days sequence only if their departure time is before midnight. We make an implicit assumption that delays between consecutive days are independent of each other, and that normal flight operation happens during daytime. This assumption appears to hold if the duration of most of the journeys are relatively short and geographically clustered, which is generally true for both our datasets which are based on European low-cost airlines. However in a minority of cases there are relatively large gaps between flights in a single sequence.

Historical Pattern Sequencing: We generated sequences based on the historical pattern of turnaround time for each aircraft. With this method we assume that the journey and turn-around pattern of each aircraft does not vary significantly, i.e. an aircraft that makes several short flights per day does not then switch and make one long-haul flight per

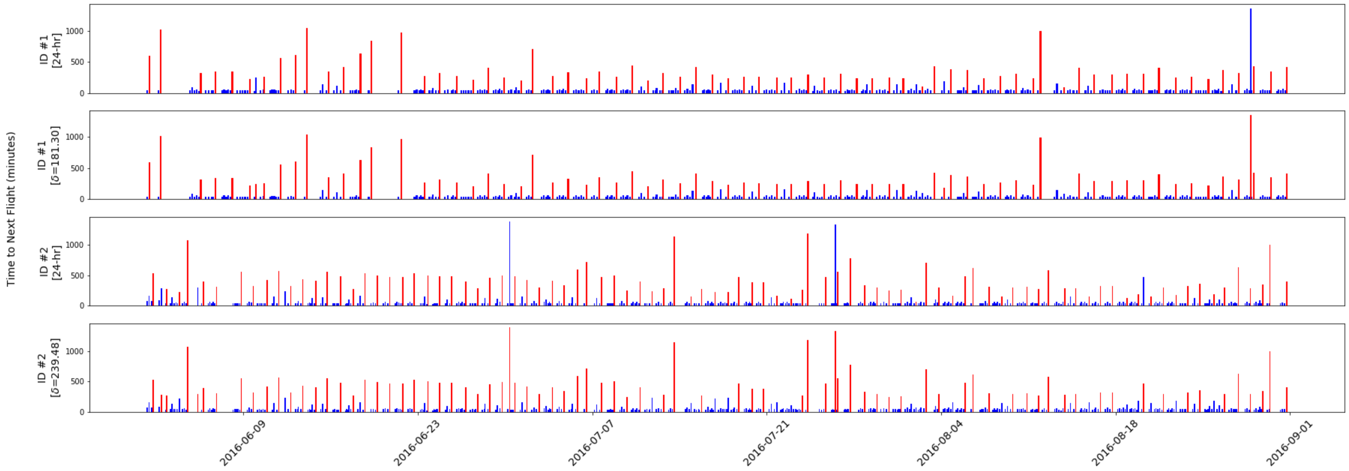


Figure 1: Comparison of sequencing methods for two airplanes for a 3 month period starting from 1st June 2016. Sequences breaks are display in blue and should delineate a natural break in flight sequences. Both methods produce largely similar sequences, however the 24-hour method (labeled [24-hr]) will occasionally include long gaps between flights in the following sequence if it crosses the midnight threshold, whereas the historical pattern sequencing (labeled with δ determined on a per airplane basis) generally performs better.

day. This assumption also appears to hold when using our European low cost airline datasets.

We compare the two sequencing methods in Figure 1, showing the time between scheduled departures for two aircraft over a 3 month period. The pattern of flights for each aircraft is periodical, with several flights with short turn-around time followed by a longer gap (generally at night-time). It is our assumption that long gaps between departure times would absorb any propagated flight delay, and are the most natural sequence breaks.

LSTM Modeling

We considered an LSTM model to capture and predict sequences of flight delays.

Background: LSTM cells were originally proposed in 1997 (Hochreiter and Schmidhuber 1997), and with several subsequent modifications (Gers, Schmidhuber, and Cummins 2000) greatly improved upon the ability of vanilla RNNs to ’remember’ long-term dependencies. LSTMs are building units for layers of a recurrent neural network, and are built from an input, output, and forget gate. Each of the three gates can be thought of as a conventional artificial neuron, that is, they compute an activation of a weighted sum. Intuitively, they can be thought as regulators of the flow of information that goes through the connections of the LSTM. Due to their structure and performance, LSTM cells have been used extensively for time series prediction, and in our case applied to address the problem of sequence-to-sequence learning for flight delays.

Customization: Figure 2 demonstrates a brief schematic overview of the architecture of a recurrent neural network model. At each stage of the prediction, the model produces an output vector which is then mapped to the estimated prediction using a perceptron layer and also passes that predicted values for the next stage of the prediction to itself.

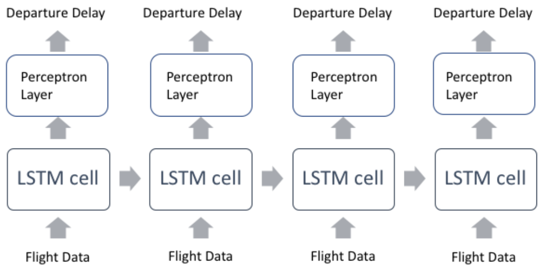
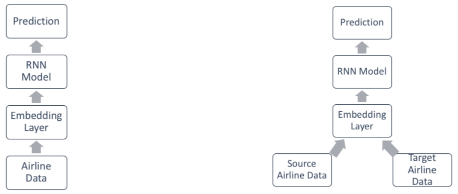


Figure 2: The LSTM Model

Transfer Learning

Motivation: Due to the difference in the number and frequency of flights, variation in reported statistics, and overall size of each airline, there is a discrepancy in the size of the data available to train the model. This is a strong motivation to apply transfer learning, aiming at improving model performance for much smaller airlines. In addition, all airlines do not fly to and from the same set of airports. Instead departure and arrival airports might be very different, as reported by our analysis i.e., 48% of airport coverage. This brought an additional motivation to formulate the problem using a framework for learning a target regression task given a source dataset with heterogeneous feature as follows.

Framework: We first define a dataset for the target task $T = \{X_T, Y_T\}$, with the target task features $X_T = \{x_T^{(i)}\}_{i=1}^{N_T}$ for $x_T \in \mathcal{R}^{M_T}$, where N_T is the target sample size and M_T is the target feature dimension, the ground-truth la-



(a) The Basic Model (b) Transfer Learning Model

Figure 3: Training procedure for the basic and transfer learning models

bels $Y_T = \{Y_T^{(i)}\}_{i=1}^{N_T}$, where $y_T \in \mathcal{R}$. For a new airline target task, we assume that we are given very few labeled instances. Similarly, we define a heterogeneous airline source dataset $S = \{X_S, Y_S\}$, with $X_S = \{x_S^{(i)}\}_{i=1}^{N_S}$ for $x_S \in \mathcal{R}^{M_S}$ for $Y_S = \{Y_S^{(i)}\}_{i=1}^{N_S}$.

Objective: The goal is to build a regression model $f : X_T \rightarrow Y_T$ that is robust for the flight delay prediction task, trained with $x_T(i), y_T(i)$ as well as knowledge transferred from $x_S(i), y_S(i)$.

Approach: Our approach aims to leverage a source data that lies in a separate feature space to the target problem. Transferring knowledge directly from heterogeneous spaces is intractable, and thus using the same approach as (Seungwhan Moon 2017), we begin by obtaining a unified vector representation of source and target categories. This process is aimed at gaining a compact representation of the source and target features that encode abstract information of the initial target and source features, which allows for more tractable transfer through affine projections. Once the label terms for the source and the target datasets are anchored in the embedding space, we learn projections into a new common latent feature space of the source and the target spaces (W_S and W_T), respectively, from which f_p maps the joint features into the label space.

We define W_S and W_T to denote the sets of learnable parameters that project source and target features into a latent joint space, where the mappings can be learned with deep neural networks.

Attention Mechanism: Our approach also used an attention mechanism, clustering the source dataset into C different clusters, $(X_S^{c_i}, Y_S^{c_i})$, to combat the effect of the negative transfer by optimizing over the importance of different clusters in the overall loss function using a learnable parameter $\mathbf{a}_i$. To learn these parameters simultaneously, we solve the following joint optimization problem with hinge rank losses for both source and target over the parameter space $(f_p, W_S, W_T, \mathbf{a})$:

$$\begin{aligned} L_{S,T} = & \sum_{j \in T} \frac{1}{|T|} L(Y_j, f_p(T(X_j, W_T))) \\ & + \sum_{i \in C} \sum_{j \in S^{c_i}} \frac{\alpha_i}{|C_i|} L(Y_{S,j}^{c_i}, f_p(T(X_{S,j}^{c_i}, W_S)) \qquad (1) \\ & + \lambda_S ||W_S|| + \lambda_S ||W_T|| + \lambda_p ||f_p||, \end{aligned}$$

where,

$$\alpha_k = \frac{exp(\mathbf{a}_k)}{\sum_i exp(\mathbf{a}_i)},$$

and $L(.,.)$ is a loss function defined based on the task, $|.|$ and $||.||$ denote the size of the dataset and an appropriate norm for the transformation respectively, and λ_S, λ_S and λ_p are regularization parameters.

Experiments

Set-up: Experiments are run on two datasets: airline S (for small) and airline B (for big), which contain 24,000 and 340,000 flights respectively. Both airlines are based in Europe, however airline B operates in 181 airports with 140 aircrafts while airline S operates in 89 airports with 44 aircrafts. Data available from S (i.e., the target domain) is nearly 15 times smaller than B, and contains only a 45% feature overlap. Both B and S share information related to flight date, scheduled and actual departure and arrival times, origin and destination airports (although only overlap in 46% of airports served by both airlines), and differ in features such as number of adult and infant passengers, total number and weight of booked bags, number of wheelchair users.

From sequencing methods in Section we can extrapolate some additional features such as turn-around time, arrival delay of previous flight, time to next scheduled departure, and sequencing.

Models: Models are trained for 700 epochs, in mini-batches of size 128, using a learning rate of 0.001 and the Adam optimizer. Sequences are created with variable lengths of 6-8 flights. We cluster the data for the airline B into two groups, the first one includes all the data points that share a departure airport common with airline S and the second one includes all the other data points. Categorical variables are encoded using one-hot encoding. The date range of the datasets did not allow us to perform a temporal train-test split (i.e. train on one year, test on the next) within the transfer learning paradigm, therefore 5-fold cross validation is used to generate the following results. We minimize effects of temporal proximity by generating training sequences such that data points do not occur in more than one sequence, and do not overlap temporally. We construct three models to evaluate performance:

- Linear Regression [LR]: We train a one shot linear model at each hop using data from Airline S.
- Basic Model [BM]: We train the basic model with the recurrent neural network described in Section 3.
- Transfer Learning Model [TL] : We train the transfer learning model using the data from Airline B.

With regard to both neural network models, the embedding layer is constructed from 512 cells with a rectified linear (ReLU) activation function, the dynamic recurrent neural network is constructed with an LSTM cell of size 64 and forget bias 1, and the output layer is constructed from 128 cell with ReLU activation mapped to single prediction using a linear transformation. All weights are initialized using the xavier initialization scheme, the bias vectors are initialized to a constant 0.1, and the initial state of the LSTM cell is initialized as zero.

Transfer Learning Training Phase:

We train the weight parameters (W_S, W_T) in (1) to produce a higher similarity between the projected source or target instance and the embedding representation of its true delay, than between the projected instance and other incorrect term embeddings. The intuition of the model is that the learned W_f is a shared and more generalized LSTM transformation model capable of mapping the joint intermediate subspace into a prediction for delay.

RMSE Results on Prediction with No Prior Knowledge:

In Table 1, we report the Root Mean Squared Error (RMSE) for the case in which we predict the departure delay at the start of the day for the following six flights assuming no prior information regarding the departure delay is known. The features used in this case are the expected number of passengers, scheduled flight times and the departure and arrival airports and at each step of the prediction, the predicted delay for the previous flight is passed to make a prediction for the next hop during the day.

		hops					
		1	2	3	4	5	6
Air. S	LR	10.5	17.7	21.5	29.3	32.3	33.1
	BM	10.2	16.7	20.4	**28.6**	31.0	**31.5**
	TL	**9.61**	**16.1**	**20.1**	**28.6**	**29.0**	32.1
Air. B	BM	19.8	21.8	23.1	29.6	33.5	44.3

Table 1: Results of RMSE for two airlines with regard to the Linear Regression(LR), Basic Model (BM) and Transfer Learning(TL) for 6 hops.

MAE Results on Prediction with No Prior Knowledge:

As we expect that very large departure delays are caused by factors not discernable from our feature set, we also report the mean absolute error (MAE), which is less sensitive and more robust with regard to large deviations than RMSE in Table 2. However, as these cases are rare, this does not affect the training process due to the batch size of 128 and small learning rate.

		1-6		2-5		3-4	
		BM	TL	BM	TL	BM	TL
S	1	6.05	**5.84**				
	2	10.01	**9.54**	9.67	**9.14**		
	3	11.38	**11.17**	11.7	**10.35**	9.41	**8.92**
	4	**16.00**	16.74	17.44	**15.89**	14.09	**12.77**
	5	**16.66**	17.01	16.94	**14.54**	15.03	**14.80**
	6	**18.01**	19.34	19.7	**17.61**	16.79	**16.73**

Table 2: Results of MAE for Airline S for Basic Model (BM) and Transfer Learning Models(TL) for 6 hops. Prediction starting at different points in the day.

MAE Results on Prediction with Prior Knowledge: In Tables 3 and 4, we report the RSME for the case that first and second flights had already departed. In this case we had access to the actual departure delays that we can feed into the prediction model to predict the departure delay for the rest of the flights of the day. Observe that as we gain more context and insight into delay profile of an aircraft during the day, the prediction becomes more accurate. In Tables 2, we report the MAE for departure delay.

		hops				
		2	3	4	5	6
Airline S	LR	17.2	18.75	40.3	28.9	30.07
	BM	16.0	21.4	42.9	29.6	34.4
	TL	**14.2**	**17.3**	**37.6**	**26.6**	**30.3**
Airline B	BM	17.1	22.1	28.7	32.5	43.2

Table 3: Results of RMSE for two airlines with regard to the Linear Regression(LR), Basic Model (BM) and Transfer Learning(TL) for 5 hops

		hops			
		3	4	5	6
Airline S	LR	18.35	31.23	33.47	31.32
	BM	17.5	28.6	33.2	28.5
	TL	**15.4**	**28.3**	**32.5**	**28.3**
Airline B	BM	18.9	25.7	30.4	42.0

Table 4: Results of RMSE for two airlines with regard to the Linear Regression(LR), Basic Model (BM) and Transfer Learning(TL) for 4 hops

Lessons Learned:

(i) On Sequencing: The first observation made was that the sequencing algorithm matters; we aim to find the best sequence cut points, such that the departure delay statistical properties of consecutive flights are independent of each other. In this regard 24 hour sequencing performed very well. However sequencing approaches would differ between airlines, depending on fleet size (network vs. low cost airlines), routes, and operated airports (EU vs. N. America).

(ii) On Transfer Learning: Despite different sets of features available with regard to two airlines, the transfer learning approach improves the performance of the model, which shows that the transfer learning model can capture the effect of the airports and time and date of the flight better the basic LSTM model due to the larger amount of data in the jointly-mapped space.

(iii) On Feature Set: Given that the feature sets used in our experiments were limited in size, we found out that increasing the size of the LSTM layer beyond 128 units does not impact the performance of the system.

(iv) On our Flight Delay Application: Recurrent neural networks using LSTM cells are powerful models capable capturing long-term dependencies in time-series data. However, the quantity and breadth of data is still the primary bottleneck when training these models. Propagated delay due to late aircraft is just one of several factors influencing flight delay, the others being roughly categorized as: operational, weather, security, carrier delay. It would be unreasonable to expect high performance over multiple hops when the current scope of our data holds such a limited perspective on the causes of flight delay. An expanded dataset that included weather conditions, airline features, airport features e.g., current occupancy, business, expected traffic, as well as contextual features such as events, national holidays, etc, would undoubtedly be able to capture in greater fidelity the delay factors and thus provide greater accuracy of predictions over flight sequences. However, the noisy 'real-world' problems that contribute directly to flight delay are still unlikely to be predicted.

(v) On Temporal Modeling for Flight Delay Application: How and what we consider a sequence of flights clearly has an impact on prediction accuracy. Given that an aircraft making a lot of short flights will affect significantly more passengers (and therefore airlines) over the course of a single day than a single long-haul flight, and with due consideration to the horizon effect introduced by longer sequences of flights, it is perhaps worth investigating the use of a temporal window around the prediction time, rather than division of flights into sequences of varying length and duration.

Conclusion and Future Work

We studied the problem of flight delay prediction in the context of a small amount of available data. Towards this challenge we presented a Long Short-Term Memory approach to predicting flight delays, modeling sequences of flights across multiple airports for a particular aircraft throughout the day. We then presented a Transfer Learning approach with heterogeneous feature space to train a system for a given smaller airline using the data from another larger airline. Our approach has demonstrated (i) to be robust and accurate for low cost airlines in Europe, and (ii) to effectively leverage knowledge from a much larger airline carrier (with different set of features) to predict flight delays in minutes.

This proved to be beneficial, and opens up future applications for smaller airlines to bootstrap the model training procedure using transfer learning, then subsequently fine-tune using their own data. Integration of this application into either airport or airline operations would have the benefit of enabling a prioritization of specific flights in order to minimize overall delay (at a later time) due to the ripple effect. However, as airport performance indicators don't take into account delays at *other* airports, there may be different incentives for airlines and airports.

In future work we will investigate (i) sequencing mechanisms on different airlines, in particular network airlines, (ii) feature space extension with integration of external data e.g., airport data, (iii) application in larger larger datasets.

References

Abdelghany, K.; Shah, S.; Raina, S.; and Abdelghany, A. 2004. A model for projecting flight delays during irregular operation conditions. *Journal of Air Transport Management* 10(6):385–394.

ANAC. 2017. http://www.anac.gov.br/.

Balakrishna, P.; Ganesan, R.; Sherry, L.; and Levy, B. S. 2008. Estimating taxi-out times with a reinforcement learning algorithm. In *2008 IEEE/AIAA 27th Digital Avionics Systems Conference*, 3.D.3–1–3.D.3–12.

CODA Digest. 2017. https://www.eurocontrol.int/articles/coda-publications.

Dai, W.; Chen, Y.; Xue, G.-r., G.; Yang, Q.; and Yu, Y. 2009. Translated learning: Transfer learning across different feature spaces. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc. 353–360.

Dück, V.; Ionescu, L.; Kliewer, N.; and Suhl, L. 2012. Increasing stability of crew and aircraft schedules. *Transportation Research Part C: Emerging Technologies* 20(1):47–61. Special issue on Optimization in Public Transport+ISTT2011.

Dhillon, P.; Foster, D. P.; and Ungar, L. H. 2011. Multi-view learning of word embeddings via cca. In Shawe-Taylor, J.; Zemel, R. S.; Bartlett, P. L.; Pereira, F.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 24*. Curran Associates, Inc. 199–207.

Gers, F. A.; Schmidhuber, J.; and Cummins, F. 2000. Learning to forget: Continual prediction with lstm. *Neural Computation* 12(10):2451–2471.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural Computer* 9(8):1735–1780.

Khanmohammadi, S.; Chou, C. A.; Lewis, H. W.; and Elias, D. 2014. A systems approach for scheduling aircraft landings in jfk airport. In *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 1578–1585.

Moon, S., and Carbonell, J. 2016. *Proactive Transfer Learning for Heterogeneous Feature and Label Spaces*. Cham: Springer International Publishing. 706–721.

Rebollo, J. J., and Balakrishnan, H. 2014. Characterization and prediction of air traffic delays. *Transportation Research Part C: Emerging Technologies* 44(Supplement C):231–241.

Seungwhan Moon, J. C. 2017. Completely heterogeneous transfer learning with attention - what and what not to transfer. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2508–2514.

Wang, W.; Arora, R.; Livescu, K.; and Bilmes, J. 2015. On deep multi-view representation learning. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 1083–1092. JMLR.org.

Wong, J.-T., and Tsai, S.-C. 2012. A survival model for flight delay propagation. *Journal of Air Transport Management* 23:5–11.

Wu, C.-L. 2005. Inherent delays and operational reliability of airline schedules. *Journal of Air Transport Management* 11(4):273–282.

Zhou, J. T.; Pan, S. J.; Tsang, I. W.; and Yan, Y. 2014. Hybrid heterogeneous transfer learning through deep learning. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*, AAAI'14, 2213–2219. AAAI Press.

Zonglei, L.; Jiandong, W.; and Guansheng, Z. 2008. A new method to alarm large scale of flights delay based on machine learning. In *2008 International Symposium on Knowledge Acquisition and Modeling*, 589–592.