MODELLING SAMPLE INFORMATIVENESS FOR DEEP AFFECTIVE COMPUTING

Georgios Rizos[1], *Björn Schuller*[1,2]

[1]Group on Language, Audio and Music, Imperial College London, UK
[2]ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany
georgios.rizos12@imperial.ac.uk

ABSTRACT

Using data with high quality annotation is crucial in emotion recognition applications, especially because the task is subjective and the raters may exhibit disagreement with respect to each sample. In this paper, we propose a meta-learning methodology that can reason about the training data and detect potentially less informative instances in order to reduce their impact in the training process. The way we inform the meta-learner on the importance of each sample is by utilising recent advances in uncertainty modelling with Bayesian neural networks that can decompose predictive uncertainty into: a) *model uncertainty* that is due to a lack of observations and b) *label uncertainty* that is due to inherent randomness in the data labelling, which we adapt for affective computing. Our proposed method for soft data selection exhibits a 6% absolute improvement in Concordance Correlation Coefficient with respect to the baseline in a two-dimensional continuous affect recognition task.

Index Terms— soft data selection, annotation quality, meta-learning, Bayesian neural networks, affect recognition

1. INTRODUCTION

Variance in the quality of annotation is currently one of the most fundamental challenges in affective computing and automatic speech analysis (ASA) in general. Even in the optimistic scenario where there is an abundance of speech samples, their acoustic quality may be corrupted by environmental or recording noise, a fact that combined with the high workload of raters can lead to lower than ideal quality in annotation. This can be especially problematic in emotion recognition applications, where the subjectiveness of the annotation task is usually addressed by aggregating the opinion of multiple raters whose inputs may not always agree. *How, then, can we teach a model to reason about whether a sample is labelled accurately? Furthermore, how can we utilise this knowledge to improve the model's results on a later iteration?*

To this end, there has been a re-surge of developments in applying Bayesian principles on neural networks [1, 2, 3] such that they are able to learn a predictive distribution for a test sample given the training data. More specifically, there has been interest [4, 5] in learning how to decompose the variance of the predictive distribution into two factors: a) *model or epistemic uncertainty*, i.e., the kind due to model parameter stochasticity which can be reduced by observing more data in that region in data space and b) *label or aleatory uncertainty*, i.e., the kind inherent in the data given our sensing capabilities which can be improved by the utilisation of better sensors of the observed variable. *We hypothesize that in the emotion recognition application, this can be realised either by adding additional raters for a sample or by removing potentially unreliable samples.*

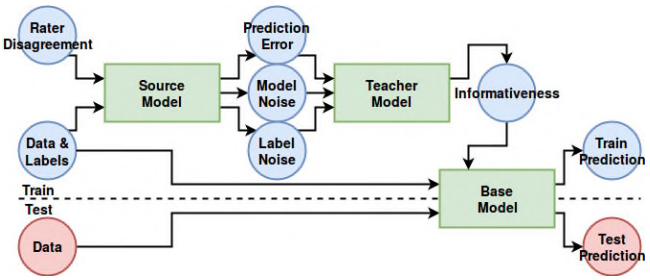

Fig. 1. Overview of the meta-learning framework.

Here, we provide a *soft sample selection scheme* integrated in the network's training process (see Fig. 1), aimed at improving performance in a test set. We achieve this by adopting a meta-learning framework, in which we first train an uncertainty decomposing Bayesian neural network that is able to model both the model and label variances associated with a prediction. We use these values along with the prediction error and the true annotator disagreement as inputs to a function that outputs *informativeness scores that appropriately weigh each sample* in the loss value calculation. This is the first application of Bayesian deep networks for *data selection* and also of uncertainty decomposing Bayesian deep models in affective computing. Our proposed meta-learning framework can be used to extend many "base" approaches, e.g., end-to-end deep learning on audio, and learning on domain knowledge features. Our experiments are performed on the RECOLA [6] database subset that was used in the AVEC 2016-2018 [7] emotion recognition sub-challenges.

2. RELATED WORK

An approach that has traditionally been used for improving the quality of speech datasets in the presence of unreliable annotations [8, 9], data redundancy [10, 11] or class imbalance [9] is *data selection* such that anomalous, mislabelled or superfluous data are removed. The problem with the aforementioned approaches, however, is that they were not designed to be used with computationally heavy, deep models for computer audition and especially end-to-end solutions [12, 13] that have proven to be an attractive alternative to the more traditional approaches that utilise specialised domain knowledge and feature engineering [14, 15]. In contrast, *we leverage deep networks that can reason on whether a sample is informative in the training process*.

A measure of *rater disagreement* has been used as an additional feature in [16] and in [17] to penalise the impact of a data sample in a support vector machine model, with positive outcomes. The authors of [9] propose to discard samples that exhibit high values of a rater disagreement measure in a categorical speech affect recognition study. Although they achieved significantly improved performance by discarding 70% of the initial dataset, this approach characterises the samples as mislabelled *based only on the rater disagreement*. The authors of a recent study on emotion recognition from video [14] avoid making 'hard' predictions by using a deep multitask network to jointly model both the aggregated emotion consensus and the rater label variance. Learning to predict the disagreement level of a sample has also been attempted in an active learning setting [18, 19] in order to query for more informative samples. *We argue that the rater disagreement of a single sample may be explained away by the presence of properly annotated samples in the neighbourhood and as such, we opt to work with informativeness estimates, based on predictive model and label noise, as well as rater disagreement.*

3. IDENTIFYING INFORMATIVE SAMPLES

In our approach, we adopt the *Monte Carlo (MC) Dropout* Bayesian deep learning approach as per [3] mainly due to the fact that it can straightforwardly be extended to convolutional [20] and recurrent neural networks [21] (henceforth CNNs and RNNs). Furthermore, an extension to this work performed in [4] yielded a methodology for decomposing variance into model and label uncertainty factors. We are going to give a succinct background review of this methodology in subsection 3.1 and then we are going to proceed with our contributions for *identification of informative samples in affective computing* in the following subsections 3.2-3.3.

3.1. Bayesian deep uncertainty decomposition

A Bayesian neural network (BNN) is defined by assuming a probabilistic prior over the weights, e.g., $p(\boldsymbol{\omega}) = \mathcal{N}(\mathbf{0}, \mathbf{1})$ and using Bayes' rule to calculate the Bayesian weight posterior $p(\boldsymbol{\omega} \mid \boldsymbol{X}, \boldsymbol{Y})$ given the data $\boldsymbol{X}, \boldsymbol{Y}$. This is an intractable operation and as such is approached via variational approximation: we sample stochastic parameters $\hat{\boldsymbol{\omega}}$ from a variational distribution ($\hat{\boldsymbol{\omega}} \sim q_{\boldsymbol{\theta}}$) that approximates the Bayesian weight posterior. This approximation is achieved by minimising the Kullback-Leibler divergence of the former from the latter, i.e., the negative Evidence Lower Bound (ELBO). In the case of MC Dropout BNNs this minimisation objective is further approximated using MC integration, which along with the choice of an appropriate variational distribution as explained in [3] implies that training with dropout [22] gives us an approximate Bayesian inference method and allows for sampling outputs from the predictive distribution $p(f_{\hat{\omega}}(\boldsymbol{x}_*) \mid \boldsymbol{x}_*, \boldsymbol{X}, \boldsymbol{Y})$ for a test sample $\boldsymbol{x}_*$.

Imagine now that our data labels are corrupted by *heteroscedastic, data-dependent* output noise. We split our BNN in two "heads" such that it predicts the parameters of a single dimensional Gaussian distribution, $[\hat{\mu}_*, \hat{\sigma}_*] = f_{\hat{\boldsymbol{\omega}}}(\boldsymbol{x}_*)$ for a test input $\boldsymbol{x}_*$. By extending the approximate Bayesian inference framework from the previous paragraph using a Gaussian heteroscedastic likelihood function as described in [4], we can train our uncertainty decomposing BNN by minimising a heteroscedastic noise mean square error (MSE) loss:

$$\min \hat{\mathcal{L}}_{VI} = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{\|y_i - \hat{\mu}_i\|_2}{\hat{\sigma}_i^2} + log\hat{\sigma}_i^2 \right) + reg. \quad (1)$$

In Eq. 1, the *reg.* factor corresponds to weight regularisation that is due to the ELBO term that encourages the variational distribution to be close to the weight prior. *We also see that if the network learns to output high uncertainty $\hat{\sigma}$ in cases it expects to get a high prediction error, this error is attenuated and as such will be backpropagated at a reduced degree*. Finally, the term $log\hat{\sigma}_i^2$ makes sure to penalise the network if it outputs high variance, since the increase of variance everywhere is a trivial means of minimising this loss function.

Finally, we can calculate the expected prediction output and the total variance of the predictive distribution as follows via Monte Carlo integration, by having the test sample perform T passes from the network, using different dropout masks for the weights every time:

$$\mathrm{E}\{\mu_*\} = \frac{1}{T} \sum_{t=1}^{T} \hat{\mu}_{*t}, \quad (2a)$$

$$\mathrm{Var}\{\mu_*\} = \frac{1}{T} \sum_{t=1}^{T} \hat{\sigma}_{*t}^2 + \frac{1}{T} \sum_{t=1}^{T} (\hat{\mu}_{*t} - \mathrm{E}\{\mu_*\})^2. \quad (2b)$$

In Eq. 2b, the first term is a proxy of learnt heteroscedastic aleatory/label uncertainty, whereas the latter of epistemic/model uncertainty, which will decrease as more data are observed near $\boldsymbol{x}_*$.

3.2. Uncertainty separation in affective computing

The authors of [4] used an extension of MSE for learning heteroscedastic data-dependent aleatory variance, but due to the recent success in using the Concordance Correlation Coefficient (CCC) ρ_c in emotion recognition as a loss function in deep learning [12, 23, 14], we need to also derive something suitable for this domain. The following Eqs. 3a-3c are the definitions of weighted mean, covariance and CCC respectively:

$$\mathrm{E}_w\{u\} = \frac{\sum u_i w_i}{\sum w_i}, \tag{3a}$$

$$\mathrm{Cov}_w\{u, v\} = \frac{\sum w_i (u_i - \mathrm{E}_w\{u\})(v_i - \mathrm{E}_w\{v\})}{\sum w_i}, \tag{3b}$$

$$\rho_{c,w}(u, v) = \frac{2\mathrm{Cov}_w\{u, v\}}{\sum\limits_{a \in \{u,v\}} \mathrm{Cov}_w\{a, a\} + (\mathrm{E}_w\{u\} - \mathrm{E}_w\{v\})^2}. \tag{3c}$$

The above means that the more confident the network is about a prediction, the more this prediction should be taken into account when calculating $\rho_{c,w}$. In this study, we propose two possible candidates as reasonable choices for the weights: a) a learnt heteroscedastic precision score from the second head of the neural network as described in Sub-sec. 3.1 (i.e., based on the first part of Eq. 2b) and b) a learnt signal output from a meta-learning model, which we describe in Subsec. 3.3. Finally, as the reader can see, *we do not use the rater disagreement in our loss function*.

3.3. Meta-learning for soft label selection

We now propose a meta-learning framework that identifies informative training samples using at its core an uncertainty decomposing BNN. Our framework utilises the following models: a) *the source meta-model*, b) *the teacher meta-model* and finally c) *the base model*. A general overview of their interaction is depicted in Fig.1. The source meta-model is a BNN trained with weighted $\rho_{c,w}$, where the w precision vector is produced by the second "head" of the neural network as described in subsections 3.1-3.2 and is capable of providing us with model and label uncertainty estimates for any given test sample. We then jointly train the base model and the teacher meta-model: the former has the same architecture as the source meta-model and the only difference is that this time the w vector is the informativeness output of the teacher model. The latter is a simpler model that takes as an input the two predictive uncertainties for each sample, the rater disagreement, as well as the prediction error of the source model. The training signal used to optimise the teacher model is the difference between the loss of the student model as calculated by using a uniform weighting for all samples and the custom weighting provided by the teacher model. The predictions of the base model are the ones used for the actual evaluation.

4. EXPERIMENTAL SETUP

Let there be a sequence of input data $\boldsymbol{X} = [\boldsymbol{x_i}]$ with corresponding labels $\boldsymbol{Y} = [\boldsymbol{y_i}]$. Any of the models we use will provide a sequence of predictions $\hat{\boldsymbol{Y}} = [\hat{\boldsymbol{y}}_{\boldsymbol{i}}]$. Our goal is to increase the Concordance Correlation Coefficient measure in the validation split for hyperparameter/architecture optimisation and in the test split for final measure reporting.

4.1. RECOLA database

We applied our framework to a two-dimensional, continuous affect recognition problem. We use the REmote COLlaborative and Affective (RECOLA) [6] database subset used in the Audio-Visual Emotion Challenge and Workshop (AVEC) 2016 [7]. This subset consists of a five-minute utterance per subject from a total of 27 subjects in a 9-9-9 train-validation-test split. The annotation period is 0.04 seconds with six annotators for each sample. We work with the speech modality and we will be using the raw audio signal as the model input, which we standardise with respect to each subject. We then augment the input by adding normally distributed noise with variance equal to 10^{-1}.

4.2. Model architectures & training

We now describe the architectures we used in our meta-learning framework used in our experiments.

Source & Base models: We utilise the model [12], which we will refer to as the **End2End** model. A sequence of raw audio signal is fed into a three-layer, single dimensional deep CNN and the latter's output into a two layer long short-term memory (LSTM) RNN. The number of filters of the convolutional layers are 64-128-256 and the widths are 8-6-6, and each one is followed by a max pooling layer that undersample at a rate of 10-8-8, respectively. The hidden units of the RNN layers are 256-256. These are trained by optimising the weighted CCC (see Eq. 3c).

Teacher model: We utilise a single layer CNN without padding (i.e., we get a single output value after convolving the input sample with each filter) and we feed that into a single layer bidirectional LSTM network with 2 output units for the two emotion targets. For the CNN we used 24 filters.

We run all competing methods for 100 epochs and keep the best model based on performance on the validation set. We also run only the source model in the case of the meta framework for the fixed amount of 50 epochs. We use 5-500 batch and sequence sizes in training mode and 1-7500 in validation and testing. We used the Adam optimiser [24] with an initial learning rate of 10^{-4}. Finally, all model predictions undergo a post-processing step that includes median filtering, centring, scaling and shifting as described in [13].

Method	Arousal	Valence	Avg
End2End (n/a)	.607	.322	.465
End2End (epi)	.691	.324	.508
End2End (both)	.680	.319	.499
End2End (meta)	***.693***	***.352***	***.523***

Table 1. Results on the AVEC-2016 RECOLA database subset **test** set. We report CCC.

5. RESULTS & DISCUSSION

We perform a comparative study by extending the *base End2End model* such that it either captures only model uncertainty, both model and label uncertainty or none, which we denote by {**epi, both, n/a**}, respectively. Finally, our implementation of a meta framework that additionally includes source and teacher models for learning sample informativeness is denoted by **meta**.

Our results are summarised in Tab. 1. The first thing we notice is that by simply extending the network such that it is Bayesian, we get a performance improvement. (*End2End (epi)* shows a .43 absolute improvement in terms of ρ_c over *End2End (n/a)*). This is probably explained by the fact that the Bayesian extension of a neural network performs much needed regularisation on the CNN and the RNN. We furthermore find that modelling model noise is better than modelling both model and label noise, possibly due to the additional complexity brought by modelling label noise in the same model. The reason the values for *End2End (n/a)* are different from the ones reported in the original paper [12] is that the authors of that paper utilise the full RECOLA database, not the AVEC-2016 subset. *This is noteworthy because it indicates that the so far best performing End2End model for speech-based affective computing requires either a large training database in order to learn the temporal patterns of the audio signal, or careful regularisation of the parameters.*

Most importantly, we see that using a meta-model to learn an informativeness measure as a means of performing soft label selection is the best performing approach. The *End2End (meta)* method exhibits a .015 absolute improvement over *End2End (epi)* for a total of .058 improvement over the baseline. We note that this improvement of the meta approach over the competing methods *holds for both emotion targets*. Given that valence is the more difficult to predict emotion for the speech modality, we may assume that the direct modelling of informativeness in such cases where there is noisy mapping between input and output can lead to a consistent performance improvement. As a final note, the fact that *End2End (meta)* performs better than *End2End (both)* indicates that a custom informativeness measure is significantly more useful for weighing samples in training (see Eq. 3c) than simply a measure of label uncertainty [4].

Finally, we show in Fig. 2 a superposition of the arousal emotion ground truth (green) and the log-informativeness measure (magenta) that is output by the teacher model for two 20s segments. *The teacher discourages the base model from focusing on temporal signal patterns that signify further from neutral affect, possibly because it is easier to capture.*

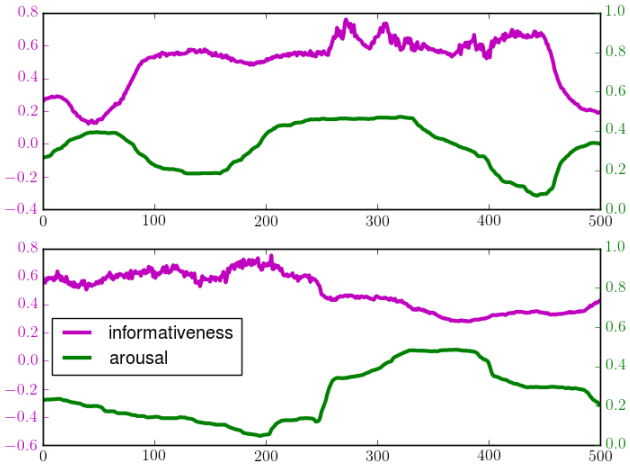

Fig. 2. Study of informativeness in speech affect. Green denotes arousal and magenta denotes informativeness.

6. CONCLUSIONS & FUTURE WORK

In this paper, we propose a meta-learning framework for soft data selection such that greater focus is placed on more informative training samples. By fusing information on annotator disagreement with model and label uncertainty estimates we can get meaningful *informativeness* estimates that can lead to impressive performance improvements in continuous affect recognition. We have shown that our framework is very effective in end-to-end affective computing, which has so far been dependent on large amounts of data. In the future we would like to generalise our results to multi-modal inputs and extend our method to categorical affective computing tasks. Given that our framework can be used to extend any deep affective computing model, we would also like to reproduce more competing methods and examine its generalisation properties.

We provide implementations of the proposed informativeness learning method and the competing methods included in the experimental comparison in the project's GitHub page (https://github.com/glam-imperial/informativeness).

7. ACKNOWLEDGMENTS

This work was supported by the UK Economic & Social Research Council through the research Grant No. HJ-253479 (ACLEW). Georgios Rizos was funded by the Imperial College President's PhD Scholarship scheme.

8. REFERENCES

[1] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra, "Weight uncertainty in neural networks," *arXiv preprint arXiv:1505.05424*, 2015.

[2] José Miguel Hernández-Lobato and Ryan Adams, "Probabilistic backpropagation for scalable learning of bayesian neural networks," in *International Conference on Machine Learning*, 2015, pp. 1861–1869.

[3] Yarin Gal and Zoubin Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[4] Alex Kendall and Yarin Gal, "What uncertainties do we need in bayesian deep learning for computer vision?," *arXiv preprint arXiv:1703.04977*, 2017.

[5] Stefan Depeweg, José Miguel Hernández-Lobato, Finale Doshi-Velez, and Steffen Udluft, "Decomposition of uncertainty for active learning and reliable reinforcement learning in stochastic systems," *arXiv preprint arXiv:1710.07283*, 2017.

[6] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne, "Introducing the recola multimodal corpus of remote collaborative and affective interactions," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–8.

[7] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016, pp. 3–10.

[8] Cigdem Eroglu Erdem, Elif Bozkurt, Engin Erzin, and A Tanju Erdem, "Ransac-based training data selection for emotion recognition from spontaneous speech," in *Proceedings of the 3rd international workshop on Affective interaction in natural environments*. ACM, 2010, pp. 9–14.

[9] Zixing Zhang, Florian Eyben, Jun Deng, and Björn Schuller, "An agreement and sparseness-based learning instance selection and its application to subjective speech phenomena," in *Proc. 5th Int. Workshop Emotion Social Signals, Sentiment, Linked Open Data, Satellite of LREC 2014*, 2014, pp. 21–26.

[10] AB Nagorski, LWJ Boves, and Herman Steeneken, "Optimal selection of speech data for automatic speech recognition systems," 2002.

[11] Yi Wu, Rong Zhang, and Alexander Rudnicky, "Data selection for speech recognition," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 562–565.

[12] Panagiotis Tzirakis, Jiehao Zhang, and Björn Schuller, "End-to-end speech emotion recognition using deep neural networks," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5089–5093.

[13] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 5200–5204.

[14] Jing Han, Zixing Zhang, Maximilian Schmitt, Maja Pantic, and Björn Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," 2017.

[15] Soheil Khorram, Zakaria Aldeneh, Dimitrios Dimitriadis, Melvin McInnis, and Emily Mower Provost, "Capturing long-term temporal dependencies with convolutional networks for continuous emotion recognition," *arXiv preprint arXiv:1708.07050*, 2017.

[16] Ingo Siegert, Ronald Böck, and Andreas Wendemuth, "Inter-rater reliability for emotion annotation in human–computer interaction: comparison and methodological improvements," *Journal on Multimodal User Interfaces*, vol. 8, no. 1, pp. 17–28, 2014.

[17] Yelin Kim and Emily Mower Provost, "Leveraging inter-rater agreement for audio-visual emotion recognition," in *Affective Computing and Intelligent Interaction (ACII), 2015 International Conference on*. IEEE, 2015, pp. 553–559.

[18] Zixing Zhang, Jun Deng, Erik Marchi, and Björn W Schuller, "Active learning by label uncertainty for acoustic emotion recognition.," in *INTERSPEECH*, 2013, pp. 2856–2860.

[19] Yue Zhang, Eduardo Coutinho, Zixing Zhang, Caijiao Quan, and Björn Schuller, "Dynamic active learning based on agreement and applied to emotion recognition in spoken interactions," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 275–278.

[20] Yarin Gal and Zoubin Ghahramani, "Bayesian convolutional neural networks with bernoulli approximate variational inference," *arXiv preprint arXiv:1506.02158*, 2015.

[21] Yarin Gal and Zoubin Ghahramani, "A theoretically grounded application of dropout in recurrent neural networks," in *Advances in neural information processing systems*, 2016, pp. 1019–1027.

[22] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[23] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *arXiv preprint arXiv:1704.08619*, 2017.

[24] Diederik P Kingma and Jimmy Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.