Abstract

Deep neural networks (DNNs) provide state-of-the-art results for a multitude of applications, but the use of DNNs for multimodal audiovisual applications is still an unsolved problem. The current approaches that combine audiovisual information do not consider inherent uncertainty or leverage true classification confidence associated with each modality in the final decision. Our contribution in this work is to apply Bayesian variational inference to DNNs for audiovisual activity recognition and quantify model uncertainty along with principled confidence. We propose a novel approach that combines deterministic and variational layers to estimate model uncertainty and principled confidence. Our experiments with in- and out-of-distribution samples selected from a subset of the Moments-in-Time (MiT) dataset show more reliable confidence measure as compared to the non-Bayesian baseline. We also demonstrate the uncertainty estimates obtained from this framework can identify out-of-distribution data on the UCF101 and MiT datasets. In the multimodal setting, the proposed framework improved precision-recall AUC by 14.4% on the subset of MiT dataset as compared to non-Bayesian baseline.

1. Introduction

Audio and vision are complementary inputs and fusing the audiovisual modalities can greatly benefit an activity recognition application. Multimodal audiovisual activity recognition is still an unsolved problem, and deep neural network (DNN) architectures are not successful in modeling the inherent ambiguity in the correlation between the two modalities. Our evaluation of audiovisual inputs indicate that both audio and visual modalities may not be correlated for all the activity categories. One of the modalities (e.g., sneezing in audio, writing in vision) can be more certain about the activity class than the other modality. It is important to model reliable uncertainty estimates for the individual modalities to benefit from multimodal fusion.

Deep learning utilizes complex models that are trained with large datasets [1, 2, 3] and have proven to be scalable and successful in solving many perception tasks providing state-of-the-art results. However, DNNs are trained to obtain the maximum likelihood estimates and disregard uncertainty around the model parameters that eventually can lead to predictive uncertainty. Deep learning models may fail in the case of noisy or out-of-distribution data, leading to overconfident decisions that could be erroneous as SoftMax probability does not capture overall model confidence. Instead, it represents relative probability that an input is from a particular class compared to the other classes.

Probabilistic Bayesian models provide principled ways to gain insight about data and capture reliable uncertainty estimates in predictions. Bayesian deep learning [4, 5] has allowed bridging DNNs and probabilistic Bayesian theory to leverage the strengths of both methodologies. Bayesian deep learning framework with Monte Carlo (MC) dropout approximate inference [6] has been used in visual scene understanding applications including camera relocalization [7], semantic segmentation [8] and depth regression [9]. In this work, we propose a Bayesian deep learning framework applied to a multimodal activity recognition task using the variational inference [10, 11] technique.

Activity recognition is an active area of research with multiple approaches depending on the application domain and the types of sensors [12]. Human activity recognition using wearable sensors such as accelerometer/gyroscopes and heart-rate monitors is used to recognize everyday human activities that include walking, running, and swimming. Human pose-based activity recognition [13, 14] methods aggregate motion and appearance information along tracks of human body parts to recognize human activity. Multimodal methods which combine optical flow or depth information along with RGB data [15, 16] are shown to provide state-of-the-art results for generic (not just hu-
man) activity recognition tasks. Methods which combine semantic level information [17] such as pose, object/scene context and other attributes including linguistic descriptors are proposed to detect group activities and shown to be more robust to sensor noise.

Multimodal models are proposed for audiovisual analysis tasks such as emotion recognition [18], audiovisual speech recognition [19], speech localization [20, 21], cross-modal retrieval [22]. A deep Boltzmann machine (DBM) [23] based architecture has been used to learn a generative model for the joint image and text space that is useful in the information retrieval task for both unimodal and multimodal queries. The audiovisual speech recognition (AVSR) task is shown to benefit from multimodal training of the joint models. In [19], a deep autoencoder model for cross-modality feature learning is proposed, where better features for one modality can be learned if multiple modalities are present at training time. A deep audio-visual speech recognition model [24] using self-attention encoder architecture is proposed to recognize speech from talking faces using vision and audio inputs. Recent work on sound localization and separation [20, 21] has shown the benefits of a joint audiovisual representation for cross-modal self-supervised learning using only audio-visual correspondence as the objective function. These audiovisual methods apply joint modeling of the audio and vision inputs during the training phase for better generalizability of the models, but then use single modality during the inference phase. Also, these methods do not provide a quantifiable means to determine the relative importance of each modality.

In this work, we focus on multimodal audiovisual activity recognition and use Bayesian DNNs to reliably estimate uncertainty associated with the individual modalities. We propose an architecture which combines state-of-the-art DNN architectures with Bayesian variational layers to obtain reliable uncertainty estimates along with principled confidence. The proposed architecture can be extended to train an end-to-end model which can use the uncertainty estimates from individual modalities to gate the contribution from an uncertain input modality towards the classification results. To the best of our knowledge, this is the first research effort that applies a Bayesian deep learning framework with variational inference for a multimodal activity recognition task to capture reliable uncertainty measures.

Our main contributions include:

1. A Bayesian variational inference model applied to the multimodal framework. Specifically, we focus on the audiovisual activity recognition task to capture principled confidence and uncertainty estimates.

2. A Bayesian DNN architecture which combines deterministic and variational layers modeled with mean-field distributions.

3. Uncertainty estimates obtained from the proposed method to identify out-of-distribution data for activity recognition.

4. Audiovisual activity recognition on subset of MiT dataset using our proposed architecture which outperforms the non-Bayesian baseline.

The rest of the document is divided into the following sections. The background on Bayesian DNNs and audiovisual activity recognition are presented in Section 2. In Section 3, the architecture of the proposed Bayesian variational inference framework is presented. The results are presented in Section 4, followed by conclusions in Section 5.

2. Background

2.1. Bayesian deep neural networks

Bayesian DNNs provide a probabilistic interpretation of deep learning models by placing distributions over the model parameters (shown in Figure 1). Bayesian Inference can be applied to estimate the predictive distribution by propagating over the model likelihood while marginalizing over the learned posterior parameter distribution. Bayesian DNNs also help in regularization by introducing distribution over network weights, capturing the posterior uncertainty around the neural network parameters. This allows transferring inherent DNN uncertainty from the parameter space to the predictive uncertainty.

Given training dataset \( D = \{x, y\} \) with inputs \( x = x_1, \ldots, x_N \) and their corresponding outputs \( y = y_1, \ldots, y_N \), in parametric Bayesian settings we would like to infer a distribution over weights \( w \) as a function \( y = f_w(x) \) that represents the DNN model. With the posterior for model parameters inferred during Bayesian neural network training, we can predict the output for a new data point by propagating over the model likelihood \( p(y|x, w) \) while drawing samples from the learned parameter posterior \( p(w|D) \).
shows the posterior distribution of model parameters obtained from model likelihood.

\[
p(w, D) = \frac{p(y \mid x, w) p(w)}{p(y \mid x)} \tag{1}
\]

Computing the posterior distribution \( p(w \mid D) \) is often intractable, some of the previously proposed techniques to achieve analytically tractable inference include: (i) Markov chain Monte Carlo sampling based probabilistic inference [4, 25] (ii) variational inference techniques to infer the tractable approximate posterior distribution around model parameters [26, 27, 28] and (iii) Monte Carlo dropout approximate inference [6]. In our work, we use the variational inference approach to infer the approximate posterior distribution around the model parameters.

Variational inference [11] is an active area of research in Bayesian deep learning, which uses gradient based optimization. This technique approximates a complex probability distribution \( p(w \mid D) \) with a simpler distribution \( q_\theta(w) \), parameterized by variational parameters \( \theta \) while minimizing the Kullback-Leibler (KL) divergence [29]. Minimizing the KL divergence is equivalent to maximizing the log evidence lower bound [6, 29].

\[
\mathcal{L} := \int q_\theta(w) \log p(y \mid x, w) dw - KL[q_\theta(w) \mid \mid p(w)] \tag{2}
\]

Predictive distribution is obtained through multiple stochastic forward passes through the network during the prediction phase while sampling from the posterior distribution of network parameters through Monte Carlo estimators. Equation 3 shows the predictive distribution of the output \( y^* \) given new input \( x^* \):

\[
p(y^* \mid x^*, D) = \int p(y^* \mid x^*, w) q_\theta(w) dw \]

\[
p(y^* \mid x^*, D) \approx \frac{1}{T} \sum_{i=1}^{T} p(y^* \mid x^*, w_i), \quad w_i \sim q_\theta(w) \tag{3}
\]

where, \( T \) is number of Monte Carlo samples.

We evaluate the model uncertainty using Bayesian active learning by disagreement (BALD) [30] for the activity recognition task. BALD quantifies mutual information between parameter posterior distribution and predictive distribution, which captures model uncertainty, as shown in Equation 4.

\[
BALD := H(y^* \mid x^*, D) - \mathbb{E}_{p(w \mid D)}[H(y^* \mid x^*, w)] \tag{4}
\]

where, \( H(y^* \mid x^*, D) \) is the predictive entropy given by:

\[
H(y^* \mid x^*, D) = -\sum_{i=0}^{K-1} p_{i\mu} \ast \log p_{i\mu} \tag{5}
\]

and \( p_{i\mu} \) is predictive mean probability of \( i^{th} \) class from \( T \) Monte Carlo samples.

### 2.2. Audiovisual Activity Recognition

Vision and audio are the ubiquitous sensor inputs which are complementary in nature and have different representations. Audiovisual methods apply joint modeling of the audio and vision inputs [19, 24] to achieve higher accuracies for complex tasks such as action recognition.

Vision-based activity recognition techniques apply a combination of spatiotemporal models to capture pixel-level information and temporal dynamics of the scene. In recent years, visual activity recognition models often use ConvNets-based models for spatial feature extraction. The image-based models [31, 32] are pre-trained on ImageNet dataset to represent the spatial features. The temporal dynamics for activity recognition is typically modeled either by using a separate temporal sequence modeling such as variants of RNNs [33, 34] or by applying 3D ConvNets [35], which extend 2D ConvNets to the temporal dimension.

Following the successes of ConvNets on vision tasks, they are shown to provide state-of-the-art results for audio classification as well. Many of the top performing methods from recent audio classification challenges [36, 37] use CNN architectures [38, 39, 40] with convolutional layers. In [41], a model similar to the VGG architecture (VGGish model) from the vision domain was trained using log-Mel spectrogram features on the AudioSet [42] dataset. AudioSet contains over one million Youtube video samples labeled with a vocabulary of acoustic events.

In this work, we focus on audiovisual activity recognition using DNNs on the trimmed video samples. The 3D-ConvNet (C3D) architecture [43] is shown to provide generic spatiotemporal representation for multiple vision tasks. We use a variant of 3D-ConvNet ResNet-101 C3D [44] architecture for the visual representation. We use VGGish architecture [41] for audio representation, which is shown to provide generic features for audio classification tasks.

### 3. Bayesian Multimodal DNN Architecture

We present a Bayesian deep learning framework for audiovisual activity recognition to obtain principled confidence and capture predictive uncertainty. Bayesian DNN models provide an uncertainty measure which is valuable in multimodal setup to fuse different modalities. The block diagram of the proposed audiovisual activity recognition using Bayesian variational inference is shown in Figure 2. Recent approaches for audiovisual analysis tasks use DNN architectures to represent the vision and audio features. Likewise, we use the ResNet-101 C3D and VGGish architectures for visual and audio modalities, respectively.

For the Bayesian DNN multimodal framework, we replace the final fully connected layer for both visual and
audio DNN models with three fully connected variational layers followed by the categorical distribution (shown in Figure 2). The weights and bias parameters in the fully connected variational layers are modeled through mean-field normal distribution, and the network is trained using Bayesian variational inference based on KL divergence [27, 28]. We use Flipout [45], which is an efficient method that correlates the gradients within a mini-batch by implicitly sampling pseudo-independent weight perturbations for each input.

Bayesian DNN maintains a probability distribution for every parameter, which can be complex to scale for deeper models as they are compute and memory intensive. In [46], it is shown that applying approximate Bayesian inference with Monte Carlo dropout to final few layers can be effective in estimating the model uncertainty. In the proposed framework, during prediction we perform multiple forward passes for the final variational layers and the remaining deterministic layers require only one forward pass.

For the comparison with the non-Bayesian baseline, we maintain the same model depth as the Bayesian DNN model and use three deterministic fully connected final layers for the non-Bayesian DNN model. The dropout layer is used after every fully connected layer to avoid over-fitting of the model. In the rest of the document, we refer the non-Bayesian DNN model as simply the DNN model. In the following section, we present the results from our experiments showing the effectiveness of Bayesian DNN over conventional DNN models.

4. Results

We analyze the model performance on the Moments-in-Time (MiT) [3] dataset. The MiT dataset consists of 339 classes, and each video clip is 3 secs (~90 frames) in length. In this work, we considered a subset of 54 classes as in-distribution and another 54 classes as out of distribution samples which include audio samples. In order to check whether DNNs can provide a reliable confidence measure, the subset of 54 classes for each category are selected after subjective evaluation to confirm the activities fall into two distinct distribution of classes. This will allow the comparison of confidence measures between DNN and Bayesian DNN models for in- and out-of-distribution classes, and the uncertainty estimates for the Bayesian DNN models (as the DNN model does not provide uncertainty estimates).

The ResNet-101 C3D DNN model is initialized with pre-trained weights for the Kinetics dataset [1]. This model is then optimized with transfer learning by training the final fourteen layers. The VGGish model is initialized with pre-trained weights for the Audio set [42] dataset. This model is then optimized with transfer learning by training the final five layers. We used stochastic gradient descent (SGD) optimizer with an initial learning rate of 0.0001 and momentum factor of 0.9 along with rate decay when the loss is plateaued.

We trained the ResNet101-C3D vision and VGGish audio architectures using the in-distribution class samples, which include ~150K training and ~5.3K validation data. We select individual vision and audio paths from the model...
| Model                  | Top1 (%) | Top5 (%) |
|-----------------------|----------|----------|
| DNN: Vision           | 52.65    | 79.79    |
| Bayesian DNN: Vision  | 53.3     | 81.20    |
| DNN: Audio            | 34.13    | 61.68    |
| Bayesian DNN: Audio   | 35.80    | 63.40    |
| DNN: Audiovisual      | 56.61    | 79.39    |
| Bayesian DNN: Audiovisual | 58.68    | 84.00    |

Table 1: Accuracies for MiT activity recognition dataset (In-distribution classes): Comparison of accuracies for Bayesian DNN and DNN models for audio, vision and audiovisual activity recognition is presented. Audiovisual Bayesian DNN model shows an improvement of 3.6% top1 and 5.8% top5 accuracy improvement over the audiovisual DNN model.

The classification accuracy for MiT in-distribution samples is presented in Table 1. Bayesian DNN model consistently provides higher accuracies for individual and combined audio-vision modalities. The Bayesian DNN audiovisual model provides top-1 accuracy improvement of 10.1% over the Bayesian DNN vision model and 3.6% over the DNN audiovisual model results. Similarly, the audiovisual Bayesian DNN model provides top-5 accuracy improvement of 3.4% over the Bayesian DNN vision model and 5.8% over the DNN audiovisual model results. The slight decrease in the accuracy of DNN audiovisual model over the vision only model can be attributed to the overconfident SoftMax probabilities between vision and audio modalities which after mean-pooling results in lower accuracy values.

Figure 3 shows the comparison of precision-recall (top) and ROC (bottom) plots using the confidence measures for DNN and Bayesian DNN models. It is observed from the plots that Bayesian DNN model consistently outperforms the DNN model for the individual modalities and also for the combined audiovisual modalities. The Precision-Recall AUC plot for the audiovisual Bayesian-DNN model shows an improvement of 14.4% over the audiovisual DNN model and an improvement of 9.5% over the vision only Bayesian DNN model. The ROC plot for audiovisual Bayesian-DNN model shows an improvement of 2.7% over the audiovisual DNN model.

4.1. Confidence measure

In this section, we compare the confidence measure obtained from the DNN and Bayesian DNN models. The confidence measure for the conventional DNN is the SoftMax probabilities used for the predictions. The mean of the categorical predictive distribution obtained from Monte Carlo sampling provides the confidence measure for Bayesian DNNs.
Figure 4: Density histogram of confidence measures for DNN and Bayesian DNN models: A distribution skewed towards right (near 1.0 on x-axis) indicates the model has higher confidence in predictions than the distribution skewed towards left. [The density histogram is a histogram with area normalized to one. Plots are overlaid with kernel density curves for better readability.]
The density histograms for the confidence measure are plotted in Figure 4. The density histogram is a histogram with area normalized to one. The height (y-axis) of density histograms indicate the distribution of confidence measure. A distribution skewed towards the right (near 1.0 on x-axis) indicates the model has higher confidence in the predictions and the distributions skewed towards left indicate lower confidence. In Figure 4 (a) & (b), the density histograms for the DNN and Bayesian DNN vision models are presented, respectively. For true (correct) predictions both DNN and Bayesian DNN models show confidence measure density histograms peaked near 1.0, indicating higher confidence in the predictions. In the case of false (incorrect) predictions, the DNN model still shows confidence measure density histogram skewed towards lower values, indicating the reliability in the predictions.

In density histograms for the audio classification results (shown in Figure 4), the DNN model for false predictions shows a peak near higher confidence value, whereas the Bayesian DNN model shows overall lower confidence. The results from Bayesian DNN model does not show a strong peak for the true predictions, which may be attributed to lower accuracies which imply lower confidence in the model predictions.

In the case of audiovisual inputs (shown in Figure 4), the density histogram plots for true predictions indicate DNN and Bayesian DNN models peak near higher confidence values. But in the case of false predictions, DNN model confidence histograms incorrectly represent higher values while the Bayesian DNN model indicates overall lower confidence values.

We compare the confidence measure for audiovisual inputs obtained using in- and out-of-distribution classes for the subset of MiT dataset. The confidence measure density histogram plots shown in Figure 5 (a) indicate the DNN model incorrectly estimates higher confidence for out-of-distribution classes and a peak is observed near higher values. The Bayesian DNN model (shown in Figure 5 (b)) indicates a lower confidence for out-of-distribution classes and is skewed towards lower values. This signifies Bayesian DNNs are being transparent in their predictions.

These results confirm that the proposed Bayesian DNN model provides more reliable confidence measure for false predictions and out-of-distribution samples, and is applicable to multimodal audiovisual settings.

4.2. Uncertainty measure

Bayesian DNN models capture uncertainty quantification which is beneficial to identify out-of-distribution samples. Out-of-distribution samples are data points which fall far off from the training data distribution. In this section, we compare BALD uncertainty measure (details are in Section 2) using in- and out-of-distribution samples from MiT dataset. Additionally, we compare the uncertainty estimates using UCF101 [47] visual action recognition dataset and use MiT dataset as out-of-distribution samples.

In Figure 6, the density histogram of BALD uncertainty measure for the Bayesian DNN model is presented. The in-
| Vision Model   | Top1 (%) | Top5 (%) |
|---------------|----------|----------|
| DNN           | 87.95    | 97.35    |
| Bayesian DNN  | 88.06    | 98.25    |

Table 2: Accuracies for the UCF101 activity recognition dataset.

and out-of-distribution classes are selected from subset of MiT dataset. For in-distribution samples, the uncertainty estimates are skewed towards lower values. In the case of out-of-distribution samples, the density histograms are skewed towards higher uncertainty values. This indicates the BALD measure is able to capture inherent model uncertainty for the out-of-distribution classes which were not seen during the training step.

We use the UCF101 visual activity recognition dataset, which has 101 activity classes to compare with MiT dataset (vision input) as the out-of-distribution samples. The training of the UCF101 dataset for vision input is done similar to the details provided in Section 3. The DNN and Bayesian DNN accuracy for the UCF101 dataset is given in Table 2, which is comparable to other results obtained for UCF101 using ResNet-101 C3D model [44].

The comparison of uncertainty measures for UCF101 dataset as in-distribution samples and the MiT dataset as out-of-distribution samples is shown in Figure 7. BALD uncertainty measure from the plots indicate a clear separation of in- and out-of-distribution samples. This validates the benefit of Bayesian DNN model which has the potential to identify out-of-distribution samples.

5. Conclusions

Effective multimodal activity recognition requires the underlying system to intelligently decide the relative importance of each modality. Bayesian inference provides a systematic way to quantify uncertainty in the deep learning model predictions. Uncertainty estimates obtained from Bayesian DNNs can identify inherent ambiguity in individual modalities, which in turn can benefit multimodal fusion. In this work, we propose a Bayesian DNN architecture that combines deterministic and variational layers applied to multimodal settings. We evaluate the proposed approach on audiovisual activity recognition using Moments-in-Time dataset. The results indicate Bayesian DNN architecture can provide more reliable confidence measure compared to the conventional DNNs. The uncertainty estimates obtained from the proposed method have the potential to identify out-of-distribution data.

The proposed Bayesian deep learning architecture can be extended to other multimodal applications. We envision extending this architecture through a principled Bayesian multimodal fusion framework that optimizes the loss function weighted by the uncertainty estimates from each modality.

References

[1] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. arXiv preprint arXiv:1705.06950, 2017.

[2] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and
Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675, 2016.

[3] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. arXiv preprint arXiv:1801.03150, 2018.

[4] Radford M Neal. Bayesian learning for neural networks, volume 118. Springer Science & Business Media, 2012.

[5] Yarin Gal. Uncertainty in deep learning. University of Cambridge, 2016.

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

[7] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In 2016 IEEE international conference on Robotics and Automation (ICRA), pages 4762–4769. IEEE, 2016.

[8] Alex Kendall, Vijay Badrinarayanan, and Roberto Cipolla. Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. arXiv preprint arXiv:1511.02680, 2015.

[9] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. arXiv preprint arXiv:1705.07115, 3, 2017.

[10] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. Machine learning, 37(2):183–233, 1999.

[11] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. Journal of the American Statistical Association, 112(518):859–877, 2017.

[12] Oscar D Lara, Miguel A Labrador, et al. A survey on human activity recognition using wearable sensors. IEEE Communications Surveys and Tutorials, 15(3):1192–1209, 2013.

[13] Michalis Raptis and Leonid Sigal. Poselet key-framing: A model for human activity recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2650–2657, 2013.

[14] Gunnar A Sigurdsson, OlgaRussakovsky, and Abhinav Gupta. What actions are needed for understanding human actions in videos? In Computer Vision (ICCV), 2017 IEEE International Conference on, pages 2156–2165. IEEE, 2017.

[15] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems, pages 568–576, 2014.

[16] Amir Shahroudy, Tian-Tsong Ng, Yihong Gong, and Gang Wang. Deep multimodal feature analysis for action recognition in rgb+d videos. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(5):1045–1058, 2018.

[17] Maryam Ziaeeafard and Robert Bergevin. Semantic human activity recognition: a literature review. Pattern Recognition, 48(8):2329–2345, 2015.

[18] Nitish Srivastava and Ruslan Salakhutdinov. Learning representations for multimodal data with deep belief nets. In International conference on machine learning workshop, volume 79, 2012.

[19] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In Proceedings of the 28th international conference on machine learning (ICML-11), pages 689–696, 2011.

[20] Ruohan Gao, Rogerio Feres, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. arXiv preprint arXiv:1804.01665, 2018.

[21] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. arXiv preprint arXiv:1804.03641, 2018.

[22] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. arXiv preprint arXiv:1706.00932, 2017.

[23] Nitish Srivastava and Ruslan R Salakhutdinov. Multimodal learning with deep boltzmann machines. In Advances in neural information processing systems, pages 2222–2230, 2012.

[24] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. arXiv preprint arXiv:1809.02108, 2018.

[25] Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In Proceedings of the 28th International Conference on Machine Learning (ICML-11), pages 681–688, 2011.

[26] Alex Graves. Practical variational inference for neural networks. In Advances in neural information processing systems, pages 2348–2356, 2011.

[27] Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. arXiv preprint arXiv:1401.0118, 2013.

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

[29] Christopher M Bishop. Pattern recognition and machine learning (information science and statistics) springer-verlag new york. Inc. Secaucus, NJ, USA, 2006.

[30] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máthé Lengyel. Bayesian active learning for classification and preference learning. arXiv preprint arXiv:1112.5745, 2011.
[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.

[32] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.

[33] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2625–2634, 2015.

[34] Yuan Yuan, Xiaodan Liang, Xiaolong Wang, Dit-Yan Yeung, and Abhinav Gupta. Temporal dynamic graph lstm for action-driven video object detection. In ICCV, pages 1819–1828, 2017.

[35] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, pages 4724–4733. IEEE, 2017.

[36] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. A multi-device dataset for urban acoustic scene classification. arXiv preprint arXiv:1807.09840, 2018.

[37] Karol J Piczak. Esc: Dataset for environmental sound classification. In Proceedings of the 23rd ACM international conference on Multimedia, pages 1015–1018. ACM, 2015.

[38] Yuma Sakashita and Masaki Aono. Acoustic scene classification by ensemble of spectrograms based on adaptive temporal divisions. Technical report, Tech. Rep., DCASE2018 Challenge, 2018.

[39] Matthias Dorfer, Bernhard Lehner, Hamid Eghbal-zadeh, Heindl Christoph, Paischer Fabian, and Widmer Gerhard. Acoustic scene classification with fully convolutional neural networks and I-vectors. Technical report, DCASE2018 Challenge, September 2018.

[40] Hossein Zeinali, Lukas Burget, and Honza Cernocky. Convolutional neural networks and x-vector embedding for dcase2018 acoustic scene classification challenge. Technical report, DCASE2018 Challenge, September 2018.

[41] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pages 131–135. IEEE, 2017.

[42] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pages 776–780. IEEE, 2017.

[43] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision, pages 4489–4497, 2015.

[44] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6546–6555, 2018.

[45] Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. arXiv preprint arXiv:1803.04386, 2018.

[46] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. arXiv preprint arXiv:1803.08533, 2018.

[47] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.