Non-parametric Bayesian Approaches to Deep Neural Networks

Konstantinos P. Panousis*

National and Kapodistrian University of Athens
Department of Telecommunications and Informatics
kon.p.panousis@gmail.com

Abstract. Despite the recent successes in Machine Learning, there remain many open challenges. The goal of this thesis is to introduce two different design paradigms for both batch as well as sequential data. The thesis initially focuses on the batch scenario, where we consider Deep Neural Networks. We revisit the current design paradigms of DNNs, aiming to introduce a novel, principled approach for network pruning and compression based on biologically inspired Local Winner-Takes-All mechanism. To this end, we propose an inferential construction for explicitly inferring the utility of network components in the context of LWTA-based networks. We employ appropriate arguments from the solid non-parametric Bayesian framework, namely stick-breaking priors. We derive efficient training and inference procedures for our model and demonstrate the capacity of our approach in a supervised classification setting in a variety of benchmark architectures and datasets. In the second approach, we consider sequential data, that still remain one of the most challenging tasks in the Machine Learning community. This work attempts to offer a principled way of modeling complex sequential data and time-series in general. To this end, we introduce a variant of rigid HMM architectures that constitutes an hierarchical extension; we postulate an additional latent first-order Markov Chain, allowing the model to alter the effective temporal dynamics of the conventional observation emitting Markov Chain. In this way, the model can dynamically infer which past state more strongly affects the current time frame. To increase the modeling capacity and robustness of the considered approach, we employ arguments from the Variational Bayesian framework. We demonstrate the modeling capabilities of the resulting model in the Human Action Recognition task. We employ benchmark datasets and compare the model's performance to similar baseline and state-of-the-art methods, while examining its ability to model data with missing values.

1 Non-parametric Bayesian Deep Networks with Local Competition

Deep Neural Networks have been established as state-of-the-art in many applications and tasks in Machine Learning. However, the currently employed architectures entail million of parameters, many of which are redundant. Not only their

* Dissertation Advisor: Sergios Theodoridis

structure imposes significant computational costs to the considered model, but additionally contributes to their over-parametrization. This fact renders DNNs susceptible to overfitting tendencies, undermining the generalization capabilities of the resulting models. Thus, the Deep Learning community has devoted significant effort in order to address this particular facet of DNNs and alleviate the overfitting tendencies of the considered models. Popular examples of such techniques include the *weight-decay* (ℓ_2) regularization and dropout [5]. However, these approaches are of a limited scope, focusing only on the regularization aspect; hence, they effectively train and retain all the weights of the architecture, without attempting to address any redundancy in the considered representation.

To this end, there are several different approaches that try to address the overparametrization of DNNs. A popular solution consists in the so-called *student-teacher* learning[4], where a *teacher* network is used to train a smaller *student* network. It is apparent that this paradigm suffers from two main drawbacks: (i) We cannot avoid the computational complexity and overfitting tendencies of the teacher network and (ii) designing an effective student-teacher approach and distillation process requires quite the artistry from the side of the practitioners.

As an alternative, researchers have examined the network pruning paradigm based on appropriate pruning criteria; in most cases, these are imposed on top of an appropriate regularization technique. In this context, Bayesian Neural Networks (BNNs) have been proposed as a full probabilistic paradigm for formulating DNNs by imposing suitable priors over the network weights. The incorporation of the Bayesian perspective in DNNs additionally allows for reducing floating point precision, necessary for representing the network weights. Specifically, the variance of the inferred weight posterior constitutes a measure of uncertainty in their estimations; thus the higher the variance, the lower the needed floating-point precision [7].

On the other hand, even though the currently employed non-linearities, such as the Rectified Linear Units (ReLUs) constitute a flexible computational tool for efficiently training DNNs, it is well understood that they do not come with strong biological plausibility. Indeed, there is an increasing body of evidence that neurons with similar functional properties are aggregated together and local competition takes place leading to a Local Winner-Takes-All (LWTA) mechanism. It has been shown that employing this mechanism to neural networks presents some promising results as automatic gain control, noise suppression and robustness to catastrophic forgetting [17].

This paper draws from these results and attempts to offer a principled way of designing a deep neural architecture that can intelligently infer the needed network complexity while compressing its parameters. To this end, we employ arguments from the mathematically solid nonparametric Bayesian framework in the context of LWTA-based networks. We derive efficient training and inference procedures for out approach by relying on the Stochastic Gradient Variational Bayes (SGVB) method. We evaluate our paradigm using well-known benchmark datasets and architectures.

1.1 Summary

In this work, we introduce a new design paradigm for designing DNNs, where the output of each hidden layer is computed via local competition between linear units. Moreover, we employ appropriate arguments from the nonparametric Bayesian framework in order to devise a mathematically solid approach that will allow for adapting the complexity of the architecture in a data-driven way via a component omission mechanism [14].

Hidden Layers in traditional neural networks contain nonlinear units; each unit is presented with a linear combination of the inputs obtained via the inner product of the input with a weights matrix and produce the corresponding output vectors as input to the next layer. In our approach, this mechanism is replaced by the introduction of LWTA blocks, each comprising a set of competing units. In this case, in order to denote that the input is now presented to each block and each unit therein, the weights are now organized in a three dimensional matrix, $W \in \mathbb{R}^{J \times K \times U}$, where J is the input dimensionality, K are the number of blocks and U the number of competing units in each block.

Within each block, each linear unit computes its activation; then, the block selects one winner unit on the basis of a competitive random sampling procedure and sets the rest to zero. This leads us to a sparse layer output that is then passed to the next layer. Before turning to the competitive random sampling procedure, we must first introduce our novel component omission mechanism.

To allow for inferring the utility of network components, we adopt concepts from the non-parametric Bayesian framework. Specifically, we choose to focus on the utility of the layer connections. To this end, we introduce a binary matrix $Z \in \{0,1\}^{J \times K}$, where each entry therein denotes if a particular feature j of the input is presented to a specific k LWTA block. If the entry is equal to zero, the corresponding set of weights for this specific feature, LWTA block, and units therein are effectively canceled out from the model. Subsequently, we impose an Indian Buffet Process prior over the binary matrix. IBP constitutes a probability distribution over infinite binary matrices. By using it as a prior, it allows for inferring how many components are needed for modeling a given set of observations, in a way that ensures sparsity in the obtained representations.

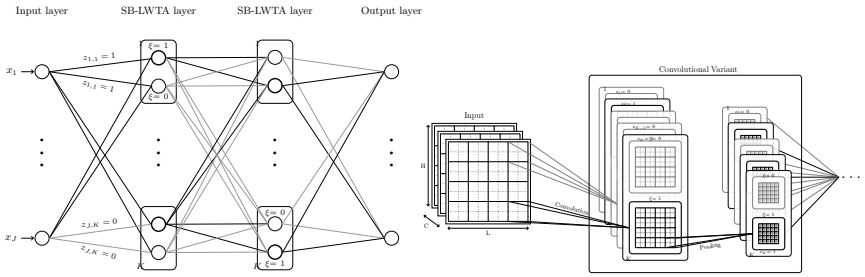
For each example n, block k and unit u therein, the expression for the output $y_n \in \mathbb{R}^{K \cdot U}$, yields:

$$[\boldsymbol{y}_n]_{ku} = [\boldsymbol{\xi}_n]_{ku} \sum_{j=1}^{J} (w_{jku} \cdot z_{jk}) \cdot [\boldsymbol{x}_n]_j \in \mathbb{R}$$

Turning to the winner sampling procedure within each LWTA block, we postulate appropriate latent variables that are driven from the layer input, and exploit the connection utility information encoded into the inferred binary matrices. The latent vectors are drawn from a Categorical distribution with a data-driven computation of the involved probabilities:

$$q([\boldsymbol{\xi}_n]_k) = \text{Categorical}\left([\boldsymbol{\xi}_n]_k \middle| \text{softmax}\left(\sum_{j=1}^{J} [w_{jku}]_{u=1}^{U} \cdot z_{jk} \cdot [\boldsymbol{x}_n]_j\right)\right)$$

We impose appropriate priors over all model parameters and seek to infer their corresponding posteriors. This concludes the formulation of a layer of the proposed Stick-Breaking LWTA (SB-LWTA) model. A graphical representation of the envisioned rationale is depicted in Fig. 1a.



(a) The proposed Approach (b) Convolutional Variant

In order to accommodate convolutional operations, we consider a variant of the previous approach. These are of importance when dealing with data of 2D structure, e.g. images. In order to perform a convolution operation over an input *tensor*, in this case, we define a set of kernels, comprising *competing feature maps*. Hence, contrary to the grouping of linear units in LWTA blocks, local competition is now perform among *feature maps*. That is, each kernel is treated as an LWTA block, and each layer comprised multiple kernels of competing feature maps (Fig. 1b).

To train the proposed model, we resort to maximization of the resulting Evidence Lower Bound (ELBO) expression. To this end, we adopt SGVB combined with (i) the standard reparameterization trick for the postulated Gaussian weights, (ii) the Gumbel-Softmax relaxation trick [6] for the introduced winning and utility latent indicator variables; and (iii) the Kumaraswamy reparameterization trick. In the inference context, we can exploit the two distinct approaches of our paradigm: (i) Component Omission Mechanism: We can exploit the introduced utility indicator variables to devise a method for assessing the utility of a network component; this way, we can infer which components are redundant and can thus be omitted from computations, and (ii) Bit-Compression: By imposing full Gaussian posteriors over the network weights, we obtain a natural way of reducing the necessary floating point precision required to represent the data. Specifically, we can utilize the posterior variances of the weights to measure their inherent uncertainty. Using this information, we can identify which bits are significant, while removing those that fluctuate under posterior uncertainty.

1.2 Results and Discussion

We perform experiments for both introduced variants, using different benchmark architectures and datasets. We assess the predictive performance metric of our

approach as well as the resulting component omission and compression capabilities, compared to state-of-the-art methods. We additionally explore the potency of the LWTA mechanism compared to the commonly employed nonlinearities.

We first consider the well-known LeNet-300-100 feedforward architecture. The corresponding comparative results are depicted in Table 1. As we observe, our method yields competitive classification accuracy, on par with the best performing alternative, while at the same retaining the least number of weights, with orders of magnitude less bit precision required to represent them. It is noteworthy that, even though the models were initialized in the same fashion, with the same number of weights and active connections, we completely outperform the competition with a greatly reduced computational footprint. Additionally, in Table 1 we introduce an additional variant of our model; we replace the LWTA blocks with ReLU units, while retaining the IBP-based mechanism; the approach is dubbed SB-ReLU. Using the aforementioned variant, we yield clearly inferior performance compared to SB-LWTA. The empirical evidence vouch for the potency of the LWTA mechanism compared to conventional nonlinearities, at least in the way that was introduced in the considered approach.

Table 1: Pruned LeNet 300-100 Architectures.

Architecture	Method	Error (%)	# Weights	Bit precision
LeNet 300-100	Original	1.6	$235K/30K/1K$	23/23/23
	StructuredBP [10]	1.7	$23,664/6,120/450$	23/23/23
	Sparse-VD [9]	1.92	$58,368/8,208/720$	8/11/14
	BC-GHS [7]	1.8	$26,746/1,204/140$	13/11/10
	SB-ReLU	1.75	13.698/6.510/730	3/4/11
	SB-LWTA (2 units)	1.7	$12,522/6,114/534$	2/3/11
	SB-LWTA (4 units)	1.75	$23,328/9,348/618$	2/3/12

We now turn to the convolutional LeNet-5-Caffe architecture. As was the case with the previous architecture, we train the network from scratch. The corresponding comparative performance is provided in Table 2. Analogously to the dense feedforward experiments, in this case, our method requires the least amount of feature maps while offering better classification accuracy accompanied by higher compression rates with respect to the best considered alternative. Moving on to a more complex dataset, CIFAR-10, and to the ConvNet convolutional architecture, we implement BC-GNJ and BC-GHS models, as described in the original paper [7]. The resulting architectures for all methods are presented in Table 2. Similar to the LeNet-5-Caffe performance, our method still retains the least number of feature maps, while providing competitive bit precision requirements, nevertheless yielding the best classification accuracy.

Our experiments have provided strong empirical evidence that the careful combination of the aforementioned approaches, allows for architectures that can greatly reduce their computational footprint, while at the same time retaining state-of-the-art predictive performance.

Table 2: Learned Convolutional Architectures.

Architecture	Method	Error (%)	# Feature Maps (Conv. Layers)	Bit precision (All Layers)
LeNet-5-Caffe	Original	0.9	25/50	23/23/23/23
	StructuredBP [10]	0.86	3/18	23/23/23/23
	VIBNet [3]	1.0	7/25	23/23/23/23
	Sparse-VD [9]	1.0	14/19	13/10/8/12
	BC-GHS [7]	1.0	5/10	10/10/14/13
	SB-ReLU	0.9	10/16	8/3/3/11
	SB-LWTA-2	0.9	6/6	6/3/3/13
	SB-LWTA-4	0.8	8/12	11/4/1/11
ConvNet	Original	17.0	64/64	23 in all layers
	BC-GNJ[7]	18.6	54/49	13/8/4/5/12
	BC-GHS[7]	17.9	42/52	12/8/5/6/10
	SB-LWTA-2	17.5	40/42	11/7/5/4/10

2 Variational Conditional Dependence Hidden Markov Models for Skeleton-based Action Recognition

There exist two major approaches for modeling sequential data: Recurrent Neural Networks (RNNs) and Hidden Markov Models (HMMs). HMMs constitute one of the most fundamental approaches, with a large history in the community. However, they have nowadays been replaced by their "deep" variants, RNNs and LSTMs, with successful application in a variety of domains. RNNs improve over the simplistic assumptions of HMMs; nevertheless, both methods exhibit several disadvantages.

On the one hand, even though the commonly considered first-order Markov Chain in a HMM allows for simplicity and low computational complexity, it introduces a significant modeling restriction to the model. More complex temporal dynamics are ignored, rendering the models practically unusable in real world scenarios. Moreover, even though the existing higher order variants alleviate this restriction, the significantly increased computational complexity, prevents their employment to complex tasks. More flexible temporal dynamics can be modeled through Hidden Semi Markov Models [21], but as is the case with HMMs, potential non-homogeneous temporal dynamics are ignored. On the other, RNNs exhibit three main drawbacks, namely: (i) They need more data to train, (ii) Exploding or Vanishing Gradients, and (iii) Training RNNs is known to be very slow, e.g., [8].

In this work, we focus on presenting a principled design paradigm for HMM methods, aiming to sidestep the simplistic or over-complicated HMM assumptions by striking a balance between flexibility and complexity. To this end, we propose a different formulation of HMMs, whereby the dependence on past frames is dynamically inferred from the data. Specifically, we introduce a hierarchical extension by postulating an additional latent variable layer; therein, the (time-varying) temporal dependence patterns are treated as latent variables over which inference is performed. We leverage solid arguments from the Variational Bayes framework and derive a tractable inference algorithm based on the forward-backward algorithm. We dub our approach Variational Conditional Dependence Hidden Markov Models (VB-CD-HMM). As we experimentally show

using benchmark datasets, our approach yields competitive recognition accuracy and can effectively handle data with missing values.

2.1 Summary

We revisit the design paradigms for modeling sequential data and introduce an hierarchical extension to HMMs that is able to capture complex temporal dependency patterns present in the data. The proposed approach comprises the postulation of an additional latent variable layer; temporal dependencies are now treated as latent variables over which inference is performed [15]. In this way, the dependence of the current frame to previous frames is inferred in a data-driven fashion. Moreover, we employ arguments from the Variational Bayes framework and introduce tractable training and inference algorithms by deriving a variant of the well-known forward-backward algorithm.

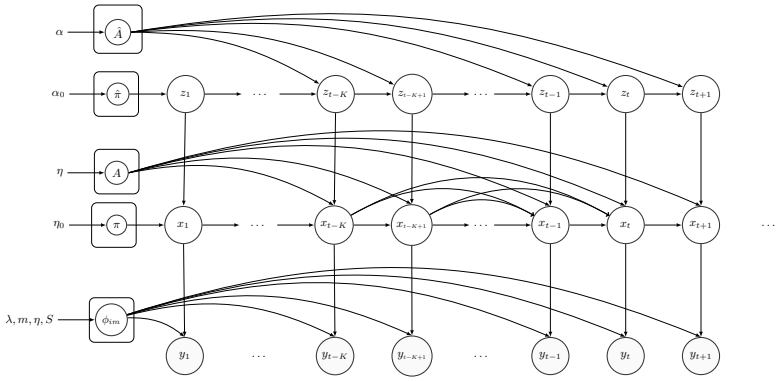


Fig. 2: The hierarchical VB-CD-HMM obtained after imposing appropriate conjugate prior distributions for all model parameters.

Let $\{\boldsymbol{y}_t\}_{t=1}^T \in \mathbb{R}^D$ be the input time-series with T frames and D features. In the following formulation, we follow the definitions of conventional HMMs. In HMMs, each observation is associated with a *discrete hidden state*, and in this work, we assume that the observation distributions are modeled via finite mixture models. Specifically, let us assume an N-state *observation-emitting* latent chain denoted as $X = \{x_t\}, x_t \in [1, \ldots, N]$, where x_t indicates the state from which, the t^{th} was emitted. For the observation model, we employ finite mixture models with M-components each, with $L = \{l_t\}_{t=1}^T,\ l_t \in [1, \ldots, M]$ denoting the *mixture component indicators*; l_t indicates which of the M components, generated the observation at time t.

The hierarchical model comprises the postulation of an additional layer, called the *dependence-generator* layer. The *latent data* of the model are now augmented by the *temporal dependence indicators* $Z = \{z_t\}_{t=1}^T,\ z_t \in [1, \ldots, K]$.

The temporal dependence indicators express the temporal dependencies between the current state at time t and the previous frames $t-1, \ldots, t-K$ in the second layer, the *observation emitting* chain; K is a pre-defined constant regarding the number of *steps-back* that the model can turn to. Thus, according to the value indicated by z_t at time t, we consider different pairwise states (x_t, x_{t-z_t}). It is apparent that, in this way, compared to a conventional first order HMM where we always consider pairwise states (x_t, x_{t-1}), the flexibility of the considered variant is greatly enhanced.

In this work, we employ arguments for the Bayesian framework in order to increase the capacity and flexibility of the model. To this end, we resort to approximate inference techniques and specifically to *Variational Inference*; we facilitate efficient training and inference procedures by deriving an appropriate variant of the forward-backward algorithm. We impose appropriate priors over all model parameters and seek to infer their corresponding posteriors. The resulting hierarchical Bayesian approach is presented in Fig. 2.

2.2 Results and Discussion

In order to test the modeling capacity of the VB-CD-HMM model, we compare our approach on action recognition benchmarks with other baseline, as well as, state-of-the-art methods. To this end, we begin our experimental approach by examining the recognition accuracy metric for each individual dataset; we then investigate the ability of the model to handle data with missing values.

For all the considered datasets, we only use skeletal data for training the models, and we follow the same training-testing splits as suggested by the authors in the original papers of each dataset. We initially focus on the recognition accuracy

Table 3: Recognition Accuracy (%) for individual dataset experiments.

Model	MSRA	UTD	G3D	Penn	Avg.
HMM	67.8	82.8	68.1	82.3	75.3
HMM^2	80.2	83.1	82.6	84.4	82.6
HSMM	66.3	82.3	77.5	78.9	76.35
LSTM	74.7	77.0	82.2	90.3	81.1
HCRF	70.7	74.2	79.0	86.3	77.6
HDM-PI	70.3	84.4	79.4	89.8	81.0
HDM-PL	80.6	90.2	87.7	91.6	87.5
HDM-BV	82.1	91.4	87.7	90.8	88.0
VB-CD-HMM	**82.5**	**92.7**	**90.6**	**92.0**	**89.45**

of VB-CD-HMM, compared to similar benchmark adaptations of conventional models such as HMMs, HSMMs, and LSTMs; in this set of experiments, we additionally consider the recently proposed Hierarchical Dynamic Model (HDM) [23], where a Bayesian hierarchical extension to Hidden Semi Markov Models is

proposed. Therein, appropriate hierarchical priors are imposed in such a way as to enable the model to capture the significant temporal and spatial variations present in the human action recognition task. In this set of experiments we focus on the recognition accuracy of the proposed model on individual datasets, as well as, the average predictive accuracy on all the considered datasets. The comparative results can be found in Table 3. We begin with the MSRA dataset, where compared to the baseline models such as HMMs, HSMMs, HCRFs and LSTMs, VB-CD-HMM outperforms them by a large margin. Specifically, over the first three considered methods, we observe an average improvement in recognition accuracy of 12.9%. The considered HDM variants consistently improve over the considered baseline models. However they fall short compared to our VB-CD-HMM model. The best performing variant is HDM-BV, where the recognition accuracy reaches 82.1%, inferior to our approach which yields 82.5%. The same behavior is consistent across all the considered datasets. Averaging the resulting classification accuracy over all the considered datasets, leads to an overall recognition accuracy of 89.45%, outperforming the baseline HMM by 14.15% and the more recent and complex HDM model by 1.45%. The obtained empirical evidence vouch for the efficacy of employing a full VB approach to the hierarchical extension, contrary to just using VB during inference. Moreover, the experimental results suggest, that the postulated first layer process can sufficiently cope with the complexity of the temporal patterns of the considered task, without requiring the introduction of any additional estimation techniques, such as *Empirical Bayes*[16]. Lastly, our proposed approach consistently and significantly outperforms a second-order HMM (HMM^2) evaluated under the same experimental and modeling setup.

To thoroughly assess the capacity of the approach, we now turn to the comparison of the resulting recognition accuracy when compared to state-of-the-art methods for each individual dataset. The corresponding results are presented in Table 4. Therein, we observe that our approach yields significant accuracy improvements over the other considered methods on the UTD dataset; the same behavior is evident in the Penn dataset. For the remaining datasets, VB-CD-HMM outperforms LRBM [12], but R3DG [18] performs better. The existent performance gap can be explained via the sophisticated feature engineering and ensemble of different complex approaches in the considered method. In contrast, in our work, we presented a simple but yet powerful hierarchical extension to the conventional HMM approaches, while at the same time utilizing very simple features, namely the joints locations and motions.

Since generative models, explicitly model the distribution of the data, they come with the additional benefit of robustness to missing values. This property is of great significance, especially in the human action recognition task, where the data may be corrupted due to hardware failure or camera occlusion. Especially, in our case, where we employ only skeletal data, robustness to missing values is extremely crucial. Since the considered model constitutes an hierarchical extension to the generative HMM model, it is itself a generative model. To assess the capacity of the proposed model, we construct an experimental setting similar

Table 4: Recognition accuracy for all the considered datasets compared to alternative state-of-the-art methods.

Dataset	Method	Acc. %
MSRA	AS[13]	83.5
	AL[19]	**88.2**
	VB-CD-HMM	82.5
UTD	Fusion [2]	79.1
	DMM [1]	84.1
	CNN [20]	85.8
	VB-CD-HMM	**92.7**
G3D	LRBM [12]	90.5
	R3DG [18]	**91.1**
	VB-CD-HMM	90.6
Penn	Actemes[22]	86.5
	AOG [11]	84.8
	VB-CD-HMM	**92.0**

Table 5: Recognition Accuracy (%) with missing values. Accuracies for R3DG [18], DLSTM [24] and HDM [23] were taken from the latter.

Dataset		UTD			MSRA			G3D		
Missing Portion		10%	30%	50%	10%	30%	50%	10%	30%	50%
Model	R3DG [18]	81.5	74.0	72.0	78.0	72.0	70.0	87.0	86.0	83.0
	DLSTM [24]	70.5	66.0	63.0	68.0	63.0	61.0	81.0	76.0	73.0
	HDM [23]	91.0	90.5	90.0	80.5	78.0	76.0	90.0	89.0	88.0
	VB-CD-HMM	**92.55**	**91.6**	**90.2**	**81.7**	**80.1**	**79.1**	**90.2**	**89.3**	**88.5**

to [23], where for three of the four datasets, we randomly omit a portion of the observations. Specifically, we utilize the UTD, MSRA and G3D datasets, and we consider three different configurations where we randomly omit 10%, 30% and 50% of the observations for both the train and test data. The recognition rates are presented in Table 5. As is clearly shown, our method clearly outperforms the R3DG [18] and DLSTM [24] methods by a large margin. Turning to the more relative HDM approach [23], we observe the same pattern. It is noteworthy that, our VB-CD-HMM model not only exhibits the higher recognition accuracy in each setting and dataset, but it additionally exhibits the *smallest decrease* in accuracy relative to the increase of missing values.

3 Conclusions

In this thesis, we considered two of the most popular paradigms in Machine Learning, Deep Neural Networks and Hidden Markov Models, aiming to introduce principled design paradigms to tackle the inherent problems of their

conventional formulations.

In the former, our work focused on one of the most significant problems of deep architectures, namely, their complexity, and specifically their overparameterization. We devised a principled mechanism to explicitly model and infer component utility in a data-driven way. Through inference, the model can learn which components are of utility to the model and which can be safely omitted from computations, intelligently adapting its structure to accommodate the complexity of the data. Moreover, our approach also examined the potency of a different activation function, assessing the overall performance of the resulting architectures. Our extensive experimental results vouch for the potency of the LWTA approach compared to currently employed activations. The introduced component omission mechanism allows for retaining the least number of weights with the least bit precision necessary to represent them, while at the time providing comparative performance, compared to related state-of-the-art approaches.

Turning to the HMM paradigm, we aimed to remove the restriction of the first-order Markovian assumption of conventional approaches, while avoiding the introduced complexity of higher-order methods. To this end, in this work, we introduced an hierarchical extension by postulating an additional latent chain that effectively determined the temporal dependencies of a conventional latent Markov Chain. To this end, we treated the temporal dependencies as random variables over which inference was performed. To facilitate efficient training and inference procedures, we derived a variant of the well-known backward algorithm used in conventional HMMs. The considered model was additionally augmented by employing arguments from the solid Bayesian framework. We evaluated our approach in one of the most challenging tasks in the Computer Vision community, namely, Human Action Recognition. The experimental results vouch for the efficacy of our approach. The model outperforms all baseline models, provides competitive recognition accuracy when compared to the state-of-the-art methods and can effectively model data with missing values.

References

1. Bulbul, M.F., Jiang, Y., Ma, J.: Dmms-based multiple features fusion for human action recognition. Int. J. Multimed. Data Eng. Manag. **6**, 23–39 (Oct 2015)
2. Chen, C., Jafari, R., Kehtarnavaz, N.: Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In: 2015 IEEE International Conference on Image Processing (ICIP) (Sep 2015)
3. Dai, B., Zhu, C., Guo, B., Wipf, D.: Compressing neural networks using the variational information bottleneck. In: Proceedings of the 35th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 80, pp. 1135–1144 (10–15 Jul 2018)
4. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015)
5. Hinton, G.E., Srivastava, N., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.R.: Improving neural networks by preventing co-adaptation of feature detectors. CoRR (2012), cite arxiv:1207.0580

6. Jang, E., Gu, S., Poole, B.: Categorical reparameterization using gumbel-softmax. In: Proc. ICLR (2017)
7. Louizos, C., Ullrich, K., Welling, M.: Bayesian compression for deep learning. In: Proc. NIPS. pp. 3290–3300 (2017)
8. Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., Khudanpur, S.: Recurrent neural network based language model. In: Kobayashi, T., Hirose, K., Nakamura, S. (eds.) INTERSPEECH. pp. 1045–1048. ISCA (2010)
9. Molchanov, D., Ashukha, A., Vetrov, D.: Variational dropout sparsifies deep neural networks. In: Proc. ICML 34. Proc. MLR, vol. 70, pp. 2498–2507. PMLR, International Convention Centre, Sydney, Australia (06–11 Aug 2017)
10. Neklyudov, K., Molchanov, D., Ashukha, A., Vetrov, D.P.: Structured bayesian pruning via log-normal multiplicative noise. In: Proc. NIPS 31, pp. 6775–6784 (2017)
11. Nie, B.X., Xiong, C., Zhu, S.: Joint action recognition and pose estimation from video. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1293–1301 (June 2015)
12. Nie, S., Wang, Z., Ji, Q.: A generative restricted boltzmann machine based method for high-dimensional motion data modeling. Comput. Vis. Image Underst. **136**, 14–22 (Jul 2015)
13. Ohn-Bar, E., Trivedi, M.M.: Joint angles similarities and hog2 for action recognition. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 465–470 (June 2013)
14. Panousis, K., Chatzis, S., Theodoridis, S.: Nonparametric Bayesian deep networks with local competition. In: Proceedings of the 36th International Conference on Machine Learning. Procs of Machine Learning Research, vol. 97. PMLR (Jun 2019)
15. Panousis, K.P., Chatzis, S., Theodoridis, S.: Variational conditional-dependence hidden markov models for human action recognition (2020)
16. Robbins, H.: An empirical bayes approach to statistics. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics. University of California Press, Berkeley, Calif. (1956)
17. Srivastava, R.K., Masci, J., Kazerounian, S., Gomez, F., Schmidhuber, J.: Compete to compute. In: Proc. NIPS 26, pp. 2310–2318. Curran Associates, Inc. (2013)
18. Vemulapalli, R., Arrate, F., Chellappa, R.: Human action recognition by representing 3d skeletons as points in a lie group. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 588–595 (June 2014)
19. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. Proc. CVPR (June 2012)
20. Wang, P., Li, Z., Hou, Y., Li, W.: Action recognition based on joint trajectory maps using convolutional neural networks. In: Procs. ICM. p. 102–106. MM '16, Association for Computing Machinery, New York, NY, USA (2016)
21. Yu, S.Z.: Hidden semi-markov models. Artif. Intell. **174** (2010)
22. Zhang, W., Zhu, M., Derpanis, K.G.: From actemes to action: A strongly-supervised representation for detailed action understanding. In: The IEEE International Conference on Computer Vision (ICCV) (Dec 2013)
23. Zhao, R., Xu, W., Su, H., Ji, Q.: Bayesian hierarchical dynamic model for human action recognition. In: Procs CVPR (June 2019)
24. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., Xie, X.: Co-occurrence feature learning for skeleton based action recognition using regularized deep lstm networks. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence. p. 3697–3703. AAAI'16, AAAI Press (2016)