Appendix : A Variational Information Bottleneck Based Method to Compress Sequential Networks for Human Action Recognition

A. Derivation of Eq. 2

The Objective function in Eq. 1 can be broken down into four parts, each corresponding to a specific LSTM gate as follows:

$$\mathcal{L} = \mathcal{L}_i + \mathcal{L}_f + \mathcal{L}_o + \mathcal{L}_g$$

All the gates primarily differ only in the associated parameters of the corresponding LSTM equations. Thus, the loss functions corresponding to these gates take similar forms. Thus, consider expanding the loss function corresponding to one of the LSTM gates as follows:

$$\begin{aligned}
\mathcal{L}_k &= \beta I(\mathbf{k}^T, \mathbf{v}) - I(\mathbf{k}^T, \mathbf{Y}) \\
&= \int p(\mathbf{k}^T, \mathbf{v}, \mathbf{Y}) \left[\beta \log \frac{p(\mathbf{k}^T, \mathbf{v})}{p(\mathbf{k}^T) p(\mathbf{v})} - \log \frac{p(\mathbf{k}^T, \mathbf{Y})}{p(\mathbf{k}^T) p(\mathbf{Y})} \right] d\mathbf{k}^T d\mathbf{v} d\mathbf{Y} \\
&= \int p(\mathbf{k}^T, \mathbf{v}, \mathbf{Y}) \left[\beta \log \frac{p(\mathbf{k}^T \mid \mathbf{v})}{p(\mathbf{k}^T)} - \log p(\mathbf{Y} \mid \mathbf{k}^T) \right] d\mathbf{k}^T d\mathbf{v} d\mathbf{Y} \\
&= \int p(\mathbf{k}^T, \mathbf{v}, \mathbf{Y}) \left[\beta \log \frac{p(\mathbf{k}^T \mid \mathbf{v})}{q(\mathbf{k}^T)} - \log q(\mathbf{Y} \mid \mathbf{k}^T) \right] d\mathbf{k}^T d\mathbf{v} d\mathbf{Y} \\
&= \int p(\mathbf{k}^T, \mathbf{v}, \mathbf{X}, \mathbf{Y}) \left[\beta \log \frac{p(\mathbf{k}^T \mid \mathbf{v})}{q(\mathbf{k}^T)} - \log q(\mathbf{Y} \mid \mathbf{k}^T) \right] d\mathbf{k}^T d\mathbf{v} d\mathbf{X} d\mathbf{Y} \\
&= \int p(\mathbf{X}, \mathbf{Y}) p(\mathbf{k}^T, \mathbf{v} \mid \mathbf{X}) \left[\beta \log \frac{p(\mathbf{k}^T \mid \mathbf{v})}{q(\mathbf{k}^T)} - \log q(\mathbf{Y} \mid \mathbf{v}) \right] d\mathbf{k}^T d\mathbf{v} d\mathbf{X} d\mathbf{Y} \\
&= \int p(\mathbf{X}, \mathbf{Y}) p(\mathbf{v} \mid \mathbf{X}) p(\mathbf{k}^T \mid \mathbf{v}) \left[\beta \log \frac{p(\mathbf{k}^T \mid \mathbf{v})}{q(\mathbf{k}^T)} - \log q(\mathbf{Y} \mid \mathbf{v}) \right] d\mathbf{k}^T d\mathbf{v} d\mathbf{X} d\mathbf{Y} \\
&= \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\} \sim \mathcal{D}, \mathbf{v} \sim p(\mathbf{v}|\mathbf{X})} \left[\beta \mathbb{D}_{KL} \left[p \left(\mathbf{k}^T \mid \mathbf{v} \right) \| q \left(\mathbf{k}^T \right) \right] - \int p(\mathbf{k}^T \mid \mathbf{v}) \log q \left(\mathbf{Y} \mid \mathbf{k}^T \right) d\mathbf{k}_T \right] \\
&= \mathcal{L}_{k1} - \mathcal{L}_{k2}
\end{aligned}$$

Now, to simplify $\mathcal{L}_{k2}$, we marginalize $q(\mathbf{Y} \mid \mathbf{k}^T)$ as follows:

$$\begin{aligned}
q(\mathbf{Y} \mid \mathbf{k}^T) &= \int q(\mathbf{Y}, \mathbf{h}^T \mid \mathbf{k}^T) dh^T \\
&= \int p(\mathbf{h}^T \mid \mathbf{k}^T) q(\mathbf{Y} \mid \mathbf{h}^T) dh^T \\
&= \mathbb{E}_{\mathbf{h}_T \sim p(\mathbf{h}^T|\mathbf{k}^T)} \left[q(\mathbf{Y} \mid \mathbf{h}^T) \right]
\end{aligned}$$

Taking log on both sides and using Jensen's Inequality, we get :

$$\log q(\mathbf{Y} \mid \mathbf{k}^T) \geq \mathbb{E}_{\mathbf{h}_T \sim p(\mathbf{h}^T|\mathbf{k}^T)} \left[\log q(\mathbf{Y} \mid \mathbf{h}^T) \right]$$

Using above equation in $\mathcal{L}_{k2}$ we get:

$$\begin{aligned}
\mathcal{L}_{k2} &= \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\} \sim \mathcal{D}, \mathbf{v} \sim p(\mathbf{v}|\mathbf{X})} \int p(\mathbf{k}^T \mid \mathbf{v}) \log q \left(\mathbf{Y} \mid \mathbf{k}^T \right) d\mathbf{k}^T \\
&= \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\} \sim \mathcal{D}, \mathbf{v} \sim p(\mathbf{v}|\mathbf{X}), p(\mathbf{k}_T|\mathbf{v})} \left[\log q(\mathbf{Y} \mid \mathbf{h}^T) \right] \\
&\geq \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\} \sim \mathcal{D}, \mathbf{v} \sim p(\mathbf{v}|\mathbf{X}), p(\mathbf{k}_T|\mathbf{v}), \mathbf{h}_T \sim p(\mathbf{h}^T|\mathbf{k}^T)} \left[\log q(\mathbf{Y} \mid \mathbf{h}^T) \right]
\end{aligned}$$

B. Derivation of Eq. 6

The KL term $\mathcal{L}_{v2}$ can be simplified using gaussian distributional forms specified in Eq. 4 and Eq. 5 as follows:

$$
\begin{aligned}
\mathcal{L}_{k1} &= \beta \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}\sim\mathcal{D},\mathbf{v}\sim p(\mathbf{v}|\mathbf{X})} \left[\mathbb{D}_{KL} \left[p\left(\mathbf{k}^T \mid \mathbf{v}\right) \| q\left(\mathbf{k}^T\right) \right] \right] \\
&= \beta \mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})} \Bigg[\sum_j \frac{(\mu_{kj}^2 + \sigma_{kj}^2)\cdot f_{vj}(\mathbf{v})^2}{\xi_{kj}} \\
&\qquad\qquad - \log\left(\frac{\sigma_{kj}^2 \cdot f_{kj}(\mathbf{v})^2}{\xi_{kj}} \right) \Bigg]
\end{aligned}
$$

Assuming ξ_{kj} is optimally learnt from the data, we can find optimal value of ξ_{kj} by taking gradient of above equation with respect to ξ_{kj} and equating to zero. The optimal value is given by:

$$
\xi_{kj}^* = \beta \mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})} \left[\left(\mu_{kj}^2 + \sigma_{kj}^2\right)\cdot f_{kj}(\mathbf{v})^2 \right]
$$

Putting $\xi_k j^*$ in $\mathcal{L}_{k1}$, we get:

$$
\begin{aligned}
\mathcal{L}_{k1} &= \beta \sum_j \left[\log\left(1 + \frac{\mu_{kj}^2}{\sigma_{kj}^2}\right) + \psi_{kj} \right] \\
&\geq \beta \sum_j \log\left(1 + \frac{\mu_{kj}^2}{\sigma_{kj}^2}\right)
\end{aligned}
$$

where, $\psi_{\mathbf{k}j} \geq 0$ by Jensen's Inequality and is given by:

$$
\psi_{\mathbf{k}j} = \log\left(\mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})} \left[f_{\mathbf{k}j}(\mathbf{v})^2 \right] \right) - \mathbb{E}_{\mathbf{v}\sim p(\mathbf{v})} \left[\log\left(f_{\mathbf{k}j}(\mathbf{v})^2 \right) \right]
$$

Therefore, loss function corresponding to a gate becomes:

$$
\begin{aligned}
\mathcal{L}_k &= \mathcal{L}_{k1} - \mathcal{L}_{k2} \\
&= \beta \sum_j \log\left(1 + \frac{\mu_{kj}^2}{\sigma_{kj}^2}\right) - \\
&\qquad \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}\sim\mathcal{D},\mathbf{v}\sim p(\mathbf{v}|\mathbf{X}),\mathbf{h}_T\sim p(\mathbf{h}^T|\mathbf{v})} \left[\log q(\mathbf{Y} \mid \mathbf{h}^T) \right]
\end{aligned}
$$

The overall loss function for VIB-LSTM can be obtained by summing up the losses for individual gates as follows:

$$
\begin{aligned}
\tilde{\mathcal{L}} &= \sum_{\mathbf{k}^T \in \{\mathbf{i}^T,\mathbf{f}^T,\mathbf{o}^T,\mathbf{g}^T\}} \mathcal{L}_k \\
&= \sum_{\mathbf{k}} \beta \sum_{j=1}^{l} \left[\log\left(1 + \frac{\mu_{\mathbf{k}j}^2}{\sigma_{\mathbf{k}j}^2}\right) + \psi_{\mathbf{k}j} \right] \\
&\qquad - 4\mathbb{E}_{\mathbf{X},\mathbf{Y},\mathbf{v},\mathbf{h}^T} \left[\log q\left(\mathbf{Y} \mid \mathbf{h}^T\right) \right]
\end{aligned}
$$

C. Derivation of Eq. 11 and Eq. 12

The Objective function in Eq. 10 can be simplified as follows:

$$
\begin{aligned}
\mathcal{L}_v &= \beta_v I(\mathbf{v},\mathbf{x}) - I(\mathbf{v},\mathbf{Y}) \\
&= \int p(\mathbf{v},\mathbf{X},\mathbf{Y}) \left[\beta_v \log\frac{p(\mathbf{v},\mathbf{X})}{p(\mathbf{v})p(\mathbf{x})} - \log\frac{p(\mathbf{v},\mathbf{Y})}{p(\mathbf{v})p(\mathbf{Y})} \right] d\mathbf{v}d\mathbf{X}d\mathbf{Y} \\
&= \int p(\mathbf{v},\mathbf{X},\mathbf{Y}) \left[\beta_v \log\frac{p(\mathbf{v}\mid\mathbf{X})}{p(\mathbf{v})} - \log\frac{p(\mathbf{Y}\mid\mathbf{v})}{p(\mathbf{Y})} \right] d\mathbf{v}d\mathbf{X}d\mathbf{Y} \\
&= \int p(\mathbf{v},\mathbf{X},\mathbf{Y}) \left[\beta_v \log\frac{p(\mathbf{v}\mid\mathbf{X})}{p(\mathbf{v})} - \log p(\mathbf{Y}\mid\mathbf{v}) \right] d\mathbf{v}d\mathbf{X}d\mathbf{Y} \\
&= \int p(\mathbf{v},\mathbf{X},\mathbf{Y}) \left[\beta_v \log\frac{p(\mathbf{v}\mid\mathbf{X})}{q(\mathbf{v})} - \log q(\mathbf{Y}\mid\mathbf{v}) \right] d\mathbf{v}d\mathbf{X}d\mathbf{Y} \\
&= \int p(\mathbf{X},\mathbf{Y})p(\mathbf{v}\mid\mathbf{X}) \left[\beta_v \log\frac{p(\mathbf{v}\mid\mathbf{X})}{q(\mathbf{v})} - \log q(\mathbf{Y}\mid\mathbf{v}) \right] d\mathbf{v}d\mathbf{X}d\mathbf{Y} \\
&= \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}\sim\mathcal{D}} \Bigg[\beta_v \mathbb{D}_{KL} \left[p\left(\mathbf{v}\mid\mathbf{x}\right) \| q\left(\mathbf{v}\right) \right] \\
&\qquad - \int p(\mathbf{v}\mid\mathbf{X}) \log q\left(\mathbf{Y}\mid\mathbf{v}\right) d\mathbf{v} \Bigg] \\
&= \mathcal{L}_{v1} - \mathcal{L}_{v2}
\end{aligned}
$$

We first simplify the KL term using gaussian distributional forms specified in Eq. 12 as follows:

$$
\begin{aligned}
\mathcal{L}_{v1} &= \beta_v \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}\sim\mathcal{D}} \left[\mathbb{D}_{KL} \left[p\left(\mathbf{v}\mid\mathbf{x}\right) \| q\left(\mathbf{v}\right) \right] \right] \\
&= \beta_v \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}\sim\mathcal{D}} \Bigg[\sum_j \frac{(\mu_{vj}^2 + \sigma_{vj}^2)\cdot f_{vj}(\mathbf{X})^2}{\xi_{vj}} \\
&\qquad\qquad - \log\left(\frac{\sigma_{vj}^2 \cdot f_{vj}(\mathbf{X})^2}{\xi_{vj}} \right) \Bigg]
\end{aligned}
$$

Assuming ξ_{vj} is optimally learnt from the data, we can find optimal value of ξ_{vj} by taking gradient of above equation with respect to ξ_{vj} and equating to zero. The optimal value is given by:

$$
\xi_{vj}^* = \beta_v \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}\sim\mathcal{D}} \left[\left(\mu_{vj}^2 + \sigma_{vj}^2\right)\cdot f_{vj}(\mathbf{X})^2 \right]
$$

Putting ξ_{vj}^* in $\mathcal{L}_{v1}$, we get:

$$
\begin{aligned}
\mathcal{L}_{v1} &= \beta_v \sum_j \left[\log\left(1 + \frac{\mu_{vj}^2}{\sigma_{vj}^2}\right) + \psi_{vj} \right] \\
&\geq \beta_v \sum_j \log\left(1 + \frac{\mu_{vj}^2}{\sigma_{vj}^2}\right)
\end{aligned}
$$

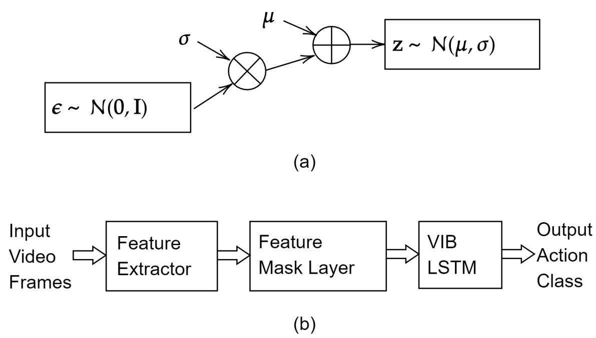

Figure 1. (a) Generation of VIB mask $\mathbf{z} \sim \mathcal{N}(\mu, \sigma)$ where the parameters μ and σ are trainable during learning of the mask but are not required during inference, (b) Basic ConvLSTM architecture with VIB layers used in our experiments. Fully connected layers at the end are not shown.

where, $\psi_{\mathbf{v}j} \geq 0$ by Jensens's Inequality and is given by:

$$\psi_{\mathbf{v}j} = \log\left(\mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}\sim\mathcal{D}}\left[f_{\mathbf{k}j}(\mathbf{X})^2\right]\right) - \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}\sim\mathcal{D}}\left[\log\left(f_{\mathbf{k}j}(\mathbf{X})^2\right)\right]$$

Now, to simplify $\mathcal{L}_{v2}$, we marginalize $q(\mathbf{Y} \mid \mathbf{v})$ as follows:

$$\begin{aligned} q(\mathbf{Y} \mid \mathbf{v}) &= \int q(\mathbf{Y}, \mathbf{h}^T \mid \mathbf{v}) dh^T \\ &= \int p(\mathbf{h}^T \mid \mathbf{v}) q(\mathbf{Y} \mid \mathbf{h}^T) dh^T \\ &= \mathbb{E}_{\mathbf{h}_T \sim p(\mathbf{h}^T \mid \mathbf{v})}\left[q(\mathbf{Y} \mid \mathbf{h}^T)\right] \end{aligned}$$

By using Jensen's inequality on the equation above and putting in $\mathcal{L}_{v2}$, we get the simplified $\mathcal{L}_{v2}$ as follows:

$$\mathcal{L}_{v2} = \mathbb{E}_{\{\mathbf{X},\mathbf{Y}\}\sim\mathcal{D}, \mathbf{v}\sim p(\mathbf{v}|\mathbf{X}), \mathbf{h}_T \sim p(\mathbf{h}^T|\mathbf{v})}\left[\log q(\mathbf{Y} \mid \mathbf{h}^T)\right]$$

D. Architecture

This section contains figures for better visualization of the approach and the pruning strategy used.

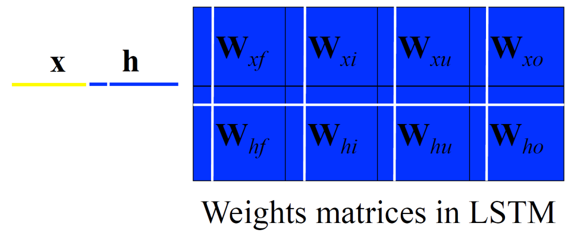

Figure 2. Figure shows the intrinsic sparse structure of LSTM parameter matrix (Wen *et al*. [24]). A single redundant unit in the LSTM hidden vector is associated with a significant number of redundant paramters in the LSTM parameter matrix.

E. Experiment results with CNN-LSTM architecture for all datasets

This section contains the tables showing detailed experimental results obtained using VIB-LSTM.

Input size	Hidden state size	LSTM parameters	Accuracy(%)
2048	2048	33.57M	98.6
266	2048	18.97M	98.53
88	2048	17.5M	98.53
33	2048	17.0M	98.2
28	2048	17.0M	97.5
2048	674	7.34M	98.65
2048	224	2.35M	98.65
2048	64	0.541M	98.53
2048	8	65856	98.53
2048	6	65856	96.53
266	224	440832	98.65
88	64	39424	98.65
33	12	2256	98.2
28	8	1216	97.1

Table 1. Compressed models trained with UCF11. Each row depicts a compressed model with corresponding details.

Input size	Hidden state size	LSTM parameters	Accuracy(%)
2048	512	5.24M	93.15
1024	512	3.14M4	93.15
256	512	1.57M	93.15
96	512	1.24M	93.15
64	512	1.18M	92.62
46	512	1.14M	92.28
31	512	1.11M	91.59
96	400	0.796M	93.15
96	297	0.469M	93.15
96	198	0.234M	92.04
96	154	0.15M	92.04
96	88	0.065M	91.5

Table 2. Compressed models trained with UCF101. Each row depicts a certain compressed model with corresponding dimensions of the LSTM matrices and validation accuracy.

Input size	Hidden state size	LSTM parameters	Accuracy(%)
2048	2048	33.570816M	68.34
1024	2048	25.182208M	68.16
149	2048	18.014208M	68.16
96	2048	17.580032M	68.16
64	2048	17.317888M	67.32
149	512	1.357824M	68.32
96	512	1.24928M	68.32
96	382	0.73344M	68.16
96	277	0.4155M	68.16
96	140	0.13328M	65.2
64	512	1.183744M	67.32
64	382	0.684544M	67.32

Table 3. Compressed models-TS-VIB-LSTM trained with HMDB51. Each row depicts a certain compressed model with corresponding dimensions of the LSTM matrices and validation accuracy.

F. Datasets

This section contains samples from all the three datasets used. Variations in various parameters like object appearance, camera position, background and object scale can be seen from the figures which makes these datasets challenging to work on.

Figure 3. Sample frames from videos from UCF11 dataset.

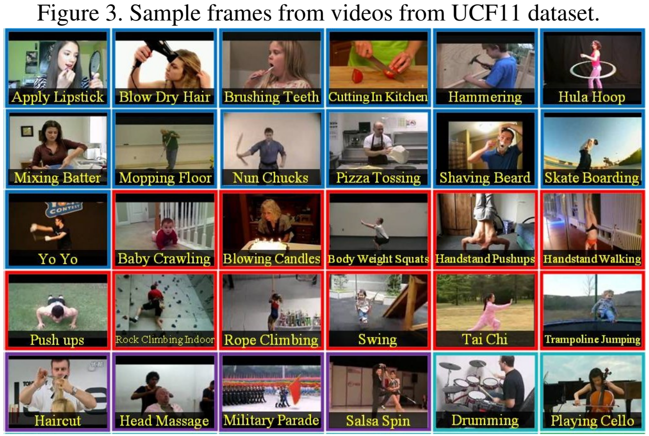

Figure 4. Sample frames from videos from UCF101 dataset.

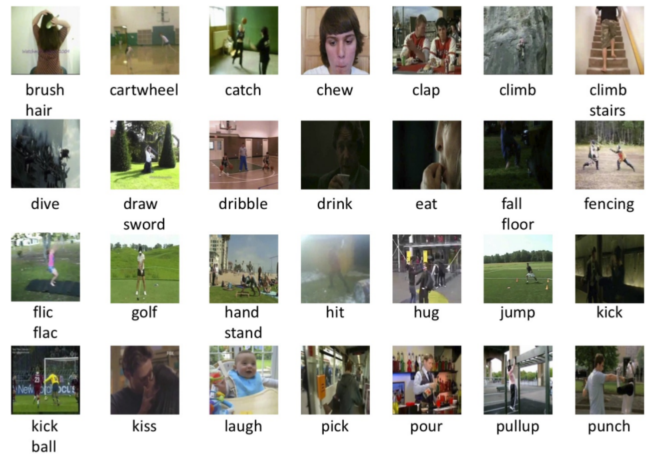

Figure 5. Sample frames from videos from HMDB51 dataset.

We have used standard datasets of which UCF101 and HMDB51 have typical train/test splits. For UCF11, we used 60:40 train/test splits with classes uniformly distributed. The datasets' sources are referred to in the main text.