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:

\[ \mathcal{L}_k = \beta I(k^T, v) - I(k^T, Y) \]

\[ = \int p(k^T, v, Y) \left[ \beta \log \frac{p(k^T | v)}{p(k^T | p(v))} - \log p(k^T, Y) \right] dkdvdY \]

\[ = \int p(k^T, v, Y) \left[ \beta \log \frac{p(k^T | v)}{p(k^T | p(v))} - \log p(Y | k^T) \right] dkdvdY \]

\[ = \int p(k^T, v, X, Y) \left[ \beta \log \frac{p(k^T | v)}{q(k^T)} - \log q(Y | k^T) \right] dkdvdXdY \]

\[ = \int p(X, Y) p(k^T, v | X) \left[ \beta \log \frac{p(k^T | v)}{q(k^T)} - \log q(Y | v) \right] dkdvdXdY \]

\[ = \int p(X, Y) p(v | X) p(k^T | v) \left[ \beta \log \frac{p(k^T | v)}{q(k^T)} - \log q(Y | v) \right] dkdvdXdY \]

\[ = \mathbb{E}_{(X,Y) \sim \mathcal{D}, v \sim p(v|X)} \left[ \beta \mathbb{D}_{KL} \left[ p \left( k^T | v \right) \| q \left( k^T \right) \right] - \int p(k^T | v) \log q \left( Y | k^T \right) dk^T \right] \]

\[ = \mathcal{L}_{k1} - \mathcal{L}_{k2} \]

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

\[ q(Y | k^T) = \int q(Y, h^T | k^T) dh^T \]

\[ = \int p(h^T | k^T) q(Y | h^T) dh^T \]

\[ = \mathbb{E}_{h^T \sim p(h^T|k^T)} \left[ q(Y | h^T) \right] \]

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

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

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

\[ \mathcal{L}_{k2} = \mathbb{E}_{(X,Y) \sim \mathcal{D}, v \sim p(v|X)} \int p(k^T | v) \log q \left( Y | k^T \right) dk^T \]

\[ = \mathbb{E}_{(X,Y) \sim \mathcal{D}, v \sim p(v|X), p(kT|v)} \left[ \log q \left( Y | h^T \right) \right] \]

\[ \geq \mathbb{E}_{(X,Y) \sim \mathcal{D}, v \sim p(v|X), p(kT|v), h^T \sim p(h^T|k^T)} \left[ \log q \left( Y | h^T \right) \right] \]
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:

\[
\mathcal{L}_{k1} = \beta \mathbb{E}_{(X, Y) \sim D, \nu \sim p(v|X)} \left[ \mathcal{D}_{KL} \left[ p \left( k^T | v \right) \| q \left( k^T \right) \right] \right]
\]

\[
= \beta \mathbb{E}_{v \sim p(v)} \left[ \sum_j \left( \frac{\mu^2_{kj} + \sigma^2_{kj}}{\xi_{kj}} \right) \cdot f_{kj}(v)^2 \right]
\]

\[
- \log \left( \frac{\sigma^2_{kj} \cdot f_{kj}(v)^2}{\xi_{kj}} \right)
\]

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

\[\xi_{kj}^* = \beta \mathbb{E}_{v \sim p(v)} \left[ \left( \frac{\mu^2_{kj} + \sigma^2_{kj}}{\xi_{kj}} \right) \cdot f_{kj}(v)^2 \right]\]

Putting \( \xi_{kj}^* \) in \( \mathcal{L}_{k1} \), we get:

\[
\mathcal{L}_{k1} = \beta \sum_j \left[ \log \left( 1 + \frac{\mu^2_{kj}}{\sigma^2_{kj}} \right) \right] + \psi_{kj}
\]

\[
\geq \beta \sum_j \log \left( 1 + \frac{\mu^2_{kj}}{\sigma^2_{kj}} \right)
\]

where, \( \psi_{kj} \geq 0 \) by Jensen’s Inequality and is given by:

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

Therefore, loss function corresponding to a gate becomes:

\[
\mathcal{L}_k = \mathcal{L}_{k1} - \mathcal{L}_{k2}
\]

\[
= \beta \sum_j \log \left( 1 + \frac{\mu^2_{kj}}{\sigma^2_{kj}} \right) - \mathbb{E}_{(X, Y) \sim D, \nu \sim p(v|X), h_T \sim p(h_T|v)} \left[ \log q(Y | h^T) \right]
\]

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

\[
\mathcal{L} = \sum_{k^r \in \{1^m, \ldots, m^r, \ldots, R^r\}} \mathcal{L}_k
\]

\[
= \sum_k \beta \sum_j \left[ \log \left( 1 + \frac{\mu^2_{kj}}{\sigma^2_{kj}} \right) + \psi_{kj} \right] - 4 \mathbb{E}_{X, Y, \nu, h^T} \left[ \log q(Y | h^T) \right]
\]

C. Derivation of Eq. 11 and Eq. 12

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

\[
\mathcal{L}_v = \beta_v I(v, x) - I(v, Y)
\]

\[
= \int p(v, X, Y) \left[ \beta_v \log \frac{p(v, X)}{p(v)p(X)} - \log \frac{p(v, Y)}{p(Y)} \right] dvdXdY
\]

\[
= \int p(v, X, Y) \left[ \beta_v \log \frac{p(v | X)}{p(v)} - \log p(Y | v) \right] dvdXdY
\]

\[
= \int p(v, X, Y) \left[ \beta_v \log \frac{p(v | X)}{q(v)} - \log q(Y | v) \right] dvdXdY
\]

\[
= \mathbb{E}_{(X, Y) \sim D} \left[ \beta_v \mathcal{D}_{KL} \left[ p \left( v | x \right) \| q \left( v \right) \right] \right]
\]

\[
- \int p(v \mid X) \log q(Y \mid v) dv
\]

\[
= \mathcal{L}_{v1} - \mathcal{L}_{v2}
\]

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

\[
\mathcal{L}_{v1} = \beta_v \mathbb{E}_{(X, Y) \sim D} \left[ \mathcal{D}_{KL} \left[ p \left( v | x \right) \| q \left( v \right) \right] \right]
\]

\[
= \beta_v \mathbb{E}_{(X, Y) \sim D} \left[ \sum_j \left( \frac{\mu^2_{vj} + \sigma^2_{vj}}{\xi_{vj}} \right) \cdot f_{vj}(X)^2 \right]
\]

\[
- \log \left( \frac{\sigma^2_{vj} \cdot f_{vj}(X)^2}{\xi_{vj}} \right)
\]

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

\[\xi_{vj}^* = \beta_v \mathbb{E}_{(X, Y) \sim D} \left[ \left( \frac{\mu^2_{vj} + \sigma^2_{vj}}{\xi_{vj}} \right) \cdot f_{vj}(X)^2 \right]\]

Putting \( \xi_{vj}^* \) in \( \mathcal{L}_{v1} \), we get:

\[
\mathcal{L}_{v1} = \beta_v \sum_j \left[ \log \left( 1 + \frac{\mu^2_{vj}}{\sigma^2_{vj}} \right) + \psi_{vj} \right]
\]

\[
\geq \beta_v \sum_j \log \left( 1 + \frac{\mu^2_{vj}}{\sigma^2_{vj}} \right)
\]
Figure 1. (a) Generation of VIB mask $z \sim N(\mu, \sigma)$ where the parameters $\mu$ and $\sigma$ 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, $\psi_{vji} \geq 0$ by Jensens’s Inequality and is given by:

$$
\psi_{vji} = \log \left( \mathbb{E}_{(X, Y) \sim D} \left[ f_{kji}(X)^2 \right] \right)
- \mathbb{E}_{(X, Y) \sim D} \left[ \log \left( f_{kji}(X)^2 \right) \right]
$$

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

$$
q(Y \mid v) = \int q(Y, h^T \mid v) dh^T
= \int p(h^T \mid v) q(Y \mid h^T) dh^T
= \mathbb{E}_{h^T \sim p(h^T \mid v)} [ q(Y \mid h^T) ]
$$

By using Jensen’s inequality on the equation above and putting in $L_{v2}$, we get the simplified $L_{v2}$ as follows:

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

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 parameters 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.

<table>
<thead>
<tr>
<th>Input size</th>
<th>Hidden state size</th>
<th>LSTM parameters</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2048</td>
<td>2048</td>
<td>33.57M</td>
<td>98.6</td>
</tr>
<tr>
<td>2048</td>
<td>266</td>
<td>18.97M</td>
<td>98.53</td>
</tr>
<tr>
<td>2048</td>
<td>88</td>
<td>17.5M</td>
<td>98.53</td>
</tr>
<tr>
<td>2048</td>
<td>33</td>
<td>17.0M</td>
<td>98.2</td>
</tr>
<tr>
<td>2048</td>
<td>28</td>
<td>17.0M</td>
<td>98.2</td>
</tr>
<tr>
<td>2048</td>
<td>264</td>
<td>7.34M</td>
<td>98.65</td>
</tr>
<tr>
<td>2048</td>
<td>224</td>
<td>2.35M</td>
<td>98.65</td>
</tr>
<tr>
<td>2048</td>
<td>64</td>
<td>0.541M</td>
<td>98.53</td>
</tr>
<tr>
<td>2048</td>
<td>8</td>
<td>65856</td>
<td>98.53</td>
</tr>
<tr>
<td>2048</td>
<td>6</td>
<td>65856</td>
<td>98.53</td>
</tr>
<tr>
<td>2048</td>
<td>264</td>
<td>440832</td>
<td>98.65</td>
</tr>
<tr>
<td>2048</td>
<td>224</td>
<td>39424</td>
<td>98.65</td>
</tr>
<tr>
<td>2048</td>
<td>6</td>
<td>2256</td>
<td>98.2</td>
</tr>
<tr>
<td>2048</td>
<td>8</td>
<td>1216</td>
<td>97.1</td>
</tr>
</tbody>
</table>

Table 1. Compressed models trained with UCF11. Each row depicts a compressed model with corresponding dimensions of the LSTM matrices and validation accuracy.

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

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

<table>
<thead>
<tr>
<th>Input size</th>
<th>Hidden state size</th>
<th>LSTM parameters</th>
<th>Accuracy(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2048</td>
<td>2048</td>
<td>33.570816M</td>
<td>68.34</td>
</tr>
<tr>
<td>1024</td>
<td>2048</td>
<td>25.182208M</td>
<td>68.16</td>
</tr>
<tr>
<td>149</td>
<td>2048</td>
<td>18.014208M</td>
<td>68.16</td>
</tr>
<tr>
<td>96</td>
<td>2048</td>
<td>17.580032M</td>
<td>68.16</td>
</tr>
<tr>
<td>64</td>
<td>2048</td>
<td>17.317888M</td>
<td>67.32</td>
</tr>
<tr>
<td>149</td>
<td>512</td>
<td>1.357824M</td>
<td>68.32</td>
</tr>
<tr>
<td>96</td>
<td>512</td>
<td>1.24928M</td>
<td>68.32</td>
</tr>
<tr>
<td>96</td>
<td>382</td>
<td>0.73344M</td>
<td>68.16</td>
</tr>
<tr>
<td>96</td>
<td>277</td>
<td>0.4155M</td>
<td>68.16</td>
</tr>
<tr>
<td>64</td>
<td>512</td>
<td>1.183744M</td>
<td>67.32</td>
</tr>
<tr>
<td>64</td>
<td>382</td>
<td>0.684544M</td>
<td>67.32</td>
</tr>
</tbody>
</table>

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.

![Sample frames from videos from UCF11 dataset.](image1)

![Sample frames from videos from UCF101 dataset.](image2)

![Sample frames from videos from HMDB51 dataset.](image3)

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.