Supplementary Material

PaperID 9208

Appx. A: Proof of Equation 6

Given that there are K iterations and the classification loss of is $\mathcal{L}(\boldsymbol{y}, \boldsymbol{t})$, where $\boldsymbol{y} = (\|\boldsymbol{v}_1^{(K)}\|, \cdots, \|\boldsymbol{v}_M^{(K)}\|)$ is the prediction and $\boldsymbol{t}$ the target, the gradients through the routing procedure are

$$\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{u}}_{m|i}} = \frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_m^{(K)}} \frac{\partial \boldsymbol{v}_m^{(K)}}{\partial \boldsymbol{s}_m^{(K)}} c_{im}^{(K)} + \sum_{j=1}^{M} \frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_j^{(K)}} \frac{\partial \boldsymbol{v}_j^{(K)}}{\partial \boldsymbol{s}_j^{(K)}} \hat{\boldsymbol{u}}_{j|i} \frac{\partial c_{ij}^{(K)}}{\partial \hat{\boldsymbol{u}}_{m|i}} \tag{1}$$

As described in the paper, the coupling coefficients of the Digit Layer are computed as

$$c_{ij}^{(t+1)} = \frac{exp(B_{ij} + \sum_{r=1}^{t} \boldsymbol{v}_j^{(r)} \hat{\boldsymbol{u}}_{j|i})}{\sum_k exp(b_{ik} + \sum_{r=1}^{t} \boldsymbol{v}_k^{(r)} \hat{\boldsymbol{u}}_{k|i})} = \frac{exp(B_{ij})}{\sum_k exp(B_{ik})} \tag{2}$$

where the superscript t is the index of an iteration, and $B_{ik} = b_{ik} + \sum_{r=1}^{t} \boldsymbol{v}_k^{(r)} \hat{\boldsymbol{u}}_{k|i}$.

When unrolling the routing procedure (a factor of the second term in Equation 1), we have

$$\begin{aligned} \frac{\partial c_{ij}^{(K)}}{\partial \hat{\boldsymbol{u}}_{m|i}} = c_{ij}^{(K)}(1 - c_{ij}^{(K)}) \frac{\partial B_{ij}^{(K-1)}}{\partial \hat{\boldsymbol{u}}_{m|i}} \\ + \sum_{k=1 \& k \neq j}^{M} c_{ij}^{(K)} c_{ik}^{(K)} \frac{\partial B_{ij}^{(K-1)}}{\partial \hat{\boldsymbol{u}}_{m|i}} \end{aligned} \tag{3}$$

Since $c_{ik} \in (0, 1)$, by unrolling the above formulation further, we have $\frac{\partial c_{ij}^{(K)}}{\partial \hat{\boldsymbol{u}}_{m|i}} \approx 0$.

At end of the training process, the coupling coefficients are polarized. There are close either to 1 or to 0. When c_{im} is close to 1, the second term in Equation 1 can be ignored. We have

$$\frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{u}}_{m|i}} \approx \frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_m^{(K)}} \frac{\partial \boldsymbol{v}_m^{(K)}}{\partial \boldsymbol{s}_m^{(K)}} c_{im}^{(K)} \tag{4}$$

When c_{im} is close to 0, We have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \hat{\boldsymbol{u}}_{m|i}} \approx & \frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_m^{(K)}} \frac{\partial \boldsymbol{v}_m^{(K)}}{\partial \boldsymbol{s}_m^{(K)}} c_{im}^{(K)} \\ & + \frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_m^{(K)}} \frac{\partial \boldsymbol{v}_m^{(K)}}{\partial \boldsymbol{s}_m^{(K)}} \hat{\boldsymbol{u}}_{m|i} c_{im}^{(K)} (1 - c_{im}^{(K)}) \frac{\partial B_{im}^{(K-1)}}{\partial \hat{\boldsymbol{u}}_{m|i}} \\ = & \frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_m^{(K)}} \frac{\partial \boldsymbol{v}_m^{(K)}}{\partial \boldsymbol{s}_m^{(K)}} c_{im}^{(K)} (1 + \hat{\boldsymbol{u}}_{m|i} (1 - c_{im}^{(K)}) \frac{\partial B_{im}^{(K-1)}}{\partial \hat{\boldsymbol{u}}_{m|i}}) \\ = & C \cdot \frac{\partial \mathcal{L}}{\partial \boldsymbol{v}_m^{(K)}} \frac{\partial \boldsymbol{v}_m^{(K)}}{\partial \boldsymbol{s}_m^{(K)}} c_{im}^{(K)} \end{aligned} \tag{5}$$

where C is a constant for the given $\hat{\boldsymbol{u}}_{m|i}$. The constant can be absorbed into the learning rate when propagated back to scale the gradients of network parameters.

Appx. B: Experimental Setting of CapsNet

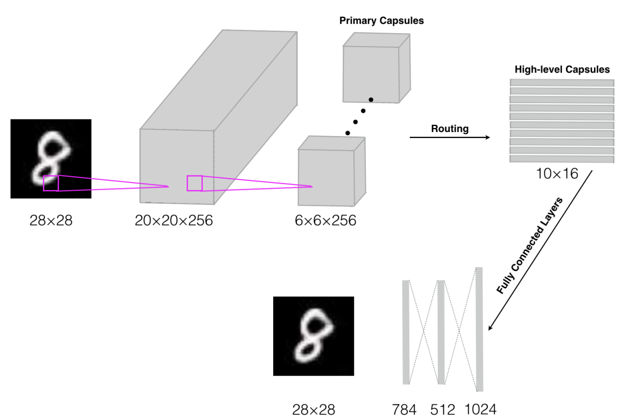

Figure 1. The Architecture of CapsNet used in the Experiments.

Training batch size	128
Training epochs	100
Learning rate	0.001
Routing iterations	3
Reconstruction weight	0.0005
Optimizer	Adam

Table 1. The Hyper-parameters of the Training Process.

Appx. C: Visualizing Computational Graphs

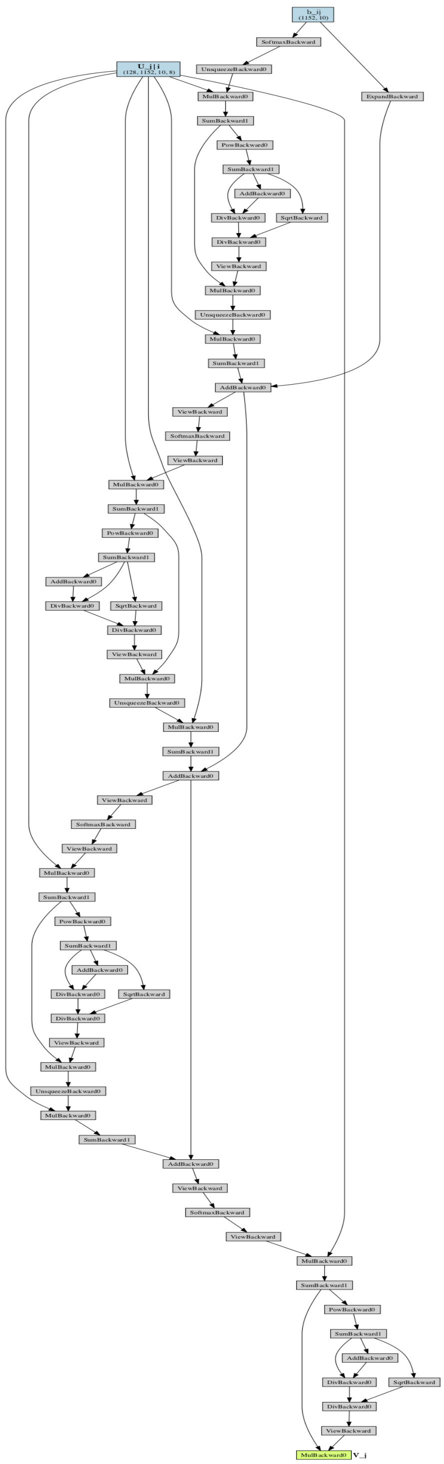

Figure 2. The computational graph of computing $\frac{\partial \hat{v}_j}{\partial \hat{\boldsymbol{u}}_{j|i}}$ where the coupling coefficients are treated as a function value of $\hat{\boldsymbol{u}}_{j|i}$. The gradients are propagated through the iterative routing iterations.

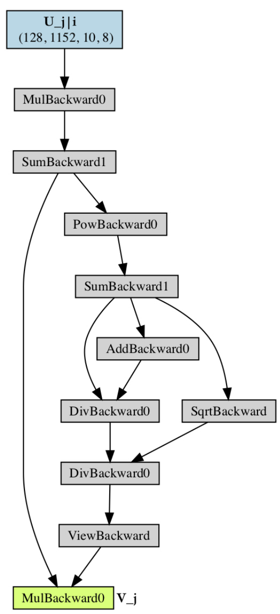

Figure 3. The computational graph of computing $\frac{\partial \hat{v}_j}{\partial \hat{\boldsymbol{u}}_{j|i}}$ where the coupling coefficients are treated as constants in gradent backpropagation.