Attention via $\log \sum \exp$ energy

Alexander Tschantz

February 1, 2025

1 General Framework

Setup. We consider a single set of nodes $\boldsymbol{v} = \{\boldsymbol{v}_i : i \in \{1, 2, \ldots, N\}\}$, where each node $\boldsymbol{v}_i \in \mathbb{R}^{d_i}$. The relationships between these nodes are defined by a set of M energy functions $\{E_m : m \in \{1, 2, \ldots, M\}\}$. Each energy function E_m defines a subset of nodes acting as *children* $C_m \subseteq \{1, 2, \ldots, normalization\}$ and a subset acting as *parents* $P_m \subseteq \{1, 2, \ldots, N\}$, which may overlap.

Energy. Each energy function E_m defines a similarity function:

$$\text{sim}(\boldsymbol{v}_c, \boldsymbol{v}_p) \quad : \quad \mathbb{R}^{d_c} \times \mathbb{R}^p \to \mathbb{R}, \tag{1}$$

which produces a scalar similarity between a child $\boldsymbol{v}_c$ and a parent $\boldsymbol{v}_p$. Let $\{\boldsymbol{v}_c\} = \{\boldsymbol{v}_c : c \in C_m\}$ and $\{\boldsymbol{v}_p\} = \{\boldsymbol{v}_p : p \in P_m\}$, the energy for E_m is defined as:

$$E_m(\{\boldsymbol{v}_c\}, \{\boldsymbol{v}_p\}) = -\sum_{c \in C_m} \ln\Big(\sum_{p \in P_m} \exp(\text{sim}(\boldsymbol{v}_c, \boldsymbol{v}_p))\Big). \tag{2}$$

The global energy sums over all energy functions:

$$E(\{\boldsymbol{v}\}) = \sum_{m=1}^{M} E_m(\{\boldsymbol{v}_c\}, \{\boldsymbol{v}_p\}). \tag{3}$$

Gradient Updates. For a single node $\boldsymbol{v}_a$, the gradient of the global energy E w.r.t. $\boldsymbol{v}_a$ decomposes into two terms. Let $\mathcal{M}_c(a) = \{m : a \in C_m\}$ denote the energy functions where $\boldsymbol{v}_a$ acts as a *child*, and $\mathcal{M}_p(a) = \{m : a \in P_m\}$ the energy functions where $\boldsymbol{v}_a$ acts as a *parent*. Then:

$$-\frac{\partial E}{\partial \boldsymbol{v}_a} = \underbrace{\sum_{m \in \mathcal{M}_c(a)} \sum_{p \in P_m} \text{softmax}_p\Big(\text{sim}(\boldsymbol{v}_a, \boldsymbol{v}_p)\Big) \frac{\partial}{\partial \boldsymbol{v}_a} \text{sim}(\boldsymbol{v}_a, \boldsymbol{v}_p)}_{\boldsymbol{v}_a \text{ acting as a child}}$$

$$+ \underbrace{\sum_{m \in \mathcal{M}_p(a)} \sum_{c \in C_m} \text{softmax}_a\Big(\text{sim}(\boldsymbol{v}_c, \boldsymbol{v}_a)\Big) \frac{\partial}{\partial \boldsymbol{v}_a} \text{sim}(\boldsymbol{v}_c, \boldsymbol{v}_a)}_{\boldsymbol{v}_a \text{ acting as a parent}}. \tag{4}$$

The first term captures contributions from $\boldsymbol{v}_a$ being explained by its parents, while the second term captures contributions from $\boldsymbol{v}_a$ explaining its children.

2 Gaussian Mixture Models

Setup. We have N data points (children) $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \in C = \{1, \ldots, N\}$, and K mixture components (parents), each with mean $\boldsymbol{\mu}_k \in \mathbb{R}^d$ and covariance $\boldsymbol{\Sigma}_k$, $k \in P = \{1, \ldots, K\}$. Let π_k be the mixing proportion.

Similarity function. We define

$$\text{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k) = \ln \pi_k - \tfrac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}_k).$$

Energy.

$$E^{\mathrm{GMM}}\big(\{\boldsymbol{x}_i\}, \{\boldsymbol{\mu}_k\}\big) \;=\; -\sum_{i=1}^{N}\ln\Big(\sum_{k=1}^{K}\exp\big(\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k)\big)\Big). \tag{5}$$

Gradients. If we differentiate w.r.t. $\boldsymbol{\mu}_k$, then

$$-\frac{\partial E^{\mathrm{GMM}}}{\partial \boldsymbol{\mu}_k} \;=\; \sum_{i=1}^{N}\mathrm{softmax}_k\big(\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k)\big)\,\boldsymbol{\Sigma}_k^{-1}\big(\boldsymbol{x}_i - \boldsymbol{\mu}_k\big).$$

Setting this gradient to zero yields the usual GMM M-step:

$$\boldsymbol{\mu}_k \;=\; \frac{\sum_{i=1}^{N}\mathrm{softmax}_k\big(\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k)\big)\,\boldsymbol{x}_i}{\sum_{i=1}^{N}\mathrm{softmax}_k\big(\mathrm{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k)\big)}.$$

3 Hopfield Networks

Setup. We have a set of *children* data vectors $\boldsymbol{x}_i \in \mathbb{R}^d$, $i \in C = \{1, \ldots, N\}$, and a set of *parent* memory vectors $\boldsymbol{m}_\mu \in \mathbb{R}^d$, $\mu \in P = \{1, \ldots, K\}$.

Similarity function.

$$\mathrm{sim}\big(\boldsymbol{x}_i, \boldsymbol{m}_\mu\big) \;=\; \boldsymbol{x}_i^\top \boldsymbol{m}_\mu.$$

Energy.

$$E^{\mathrm{Hopfield}}\big(\{\boldsymbol{x}_i\}, \{\boldsymbol{m}_\mu\}\big) \;=\; -\sum_{i=1}^{N}\ln\Big(\sum_{\mu=1}^{K}\exp\big(\boldsymbol{x}_i^\top \boldsymbol{m}_\mu\big)\Big). \tag{6}$$

Gradients.

$$-\frac{\partial E^{\mathrm{Hopfield}}}{\partial \boldsymbol{x}_i} \;=\; \sum_{\mu=1}^{K}\mathrm{softmax}_\mu\big(\boldsymbol{x}_i^\top \boldsymbol{m}_\mu\big)\,\boldsymbol{m}_\mu. \tag{7}$$

$$-\frac{\partial E^{\mathrm{Hopfield}}}{\partial \boldsymbol{m}_\mu} \;=\; \sum_{i=1}^{N}\mathrm{softmax}_\mu\big(\boldsymbol{x}_i^\top \boldsymbol{m}_\mu\big)\,\boldsymbol{x}_i. \tag{8}$$

4 Slot Attention

Setup. Let $\boldsymbol{x}_j \in \mathbb{R}^d$, $j \in C = \{1, \ldots, N\}$ be the children (tokens), and $\boldsymbol{\mu}_i \in \mathbb{R}^d$, $i \in P = \{1, \ldots, S\}$ be the parents (slots). We typically apply linear transforms $\boldsymbol{W}_K, \boldsymbol{W}_Q \in \mathbb{R}^{d \times d}$ to form

$$\mathrm{sim}\big(\boldsymbol{x}_j, \boldsymbol{\mu}_i\big) \;=\; \big(\boldsymbol{W}_K\,\boldsymbol{x}_j\big)^\top\big(\boldsymbol{W}_Q\,\boldsymbol{\mu}_i\big).$$

Energy.

$$E^{\mathrm{Slot}}\big(\{\boldsymbol{x}_j\}, \{\boldsymbol{\mu}_i\}\big) \;=\; -\sum_{j=1}^{N}\ln\Big(\sum_{i=1}^{S}\exp\big(\mathrm{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\big)\Big). \tag{9}$$

Gradients.

$$-\frac{\partial E^{\mathrm{Slot}}}{\partial \boldsymbol{\mu}_i} \;=\; \sum_{j=1}^{N}\mathrm{softmax}_i\Big(\mathrm{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\Big)\,\boldsymbol{W}_Q^\top \boldsymbol{W}_K\,\boldsymbol{x}_j. \tag{10}$$

$$-\frac{\partial E^{\mathrm{Slot}}}{\partial \boldsymbol{x}_j} \;=\; \sum_{i=1}^{S}\mathrm{softmax}_i\Big(\mathrm{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\Big)\,\boldsymbol{W}_K^\top \boldsymbol{W}_Q\,\boldsymbol{\mu}_i. \tag{11}$$

5 Self-Attention

Setup. In self-attention, every node can act as both a child (query) and a parent (key). Concretely, let us have N tokens $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$. We form

$$\boldsymbol{q}_i = \boldsymbol{W}^Q \boldsymbol{x}_i, \quad \boldsymbol{k}_i = \boldsymbol{W}^K \boldsymbol{x}_i,$$

for $i = 1, \ldots, N$. Thus the set $C = \{1, \ldots, N\}$ and $P = \{1, \ldots, N\}$ coincide, with

$$\mathrm{sim}(\boldsymbol{x}_c, \boldsymbol{x}_p) = \big(\boldsymbol{W}^Q \boldsymbol{x}_c\big)^{\top} \big(\boldsymbol{W}^K \boldsymbol{x}_p\big).$$

Energy.

$$E^{\mathrm{SA}}\big(\{\boldsymbol{x}_i\}\big) = -\sum_{c=1}^{N} \ln\Big(\sum_{p=1}^{N} \exp\big((\boldsymbol{W}^Q \boldsymbol{x}_c)^{\top}(\boldsymbol{W}^K \boldsymbol{x}_p)\big)\Big). \tag{12}$$

Gradients. Since each $\boldsymbol{x}_i$ is *both* a child and a parent, its gradient is a sum of two terms (the child side and the parent side). Writing it out explicitly:

$$-\frac{\partial E^{\mathrm{SA}}}{\partial \boldsymbol{x}_i} = \underbrace{\sum_{p=1}^{N} \mathrm{softmax}_p\Big((\boldsymbol{W}^Q \boldsymbol{x}_i)^{\top}(\boldsymbol{W}^K \boldsymbol{x}_p)\Big) \boldsymbol{W}_Q^{\top} \boldsymbol{W}_K \, \boldsymbol{x}_p}_{\text{child } i \text{ being explained by parents } p}$$
$$+ \underbrace{\sum_{c=1}^{N} \mathrm{softmax}_i\Big((\boldsymbol{W}^Q \boldsymbol{x}_c)^{\top}(\boldsymbol{W}^K \boldsymbol{x}_i)\Big) \boldsymbol{W}_K^{\top} \boldsymbol{W}_Q \, \boldsymbol{x}_c}_{\text{parent } i \text{ explaining children } c}. \tag{13}$$

6 Spatiotemporal Attention

Setup. We consider a hierarchy of L layers, each containing K_l latent variables of dimension D_l. Concretely, let $\boldsymbol{x}_{k,t}^{(l)} \in \mathbb{R}^{D_l}$ denote the k-th variable (or "slot") in layer l at time t. The lowest layer ($l = 1$) has $K_1 = N$ observed variables (e.g., pixels) and dimension D_1 (e.g., $[x, y, r, g, b]$), while higher layers have separate dimensions D_l and numbers of slots K_l. Our goal is to define an energy that couples these variables *vertically* (across layers), *concurrently* (within the same layer and time), and *temporally* (across time).

Similarity Functions. We introduce three types of similarity, each with its own projection matrices. For inter-layer connections (linking layers $l - 1$ and l), we define:

$$\mathrm{sim}_v^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \big(\boldsymbol{W}_v^{K,(l)} \boldsymbol{x}\big)^{\top} \big(\boldsymbol{W}_v^{Q,(l)} \boldsymbol{x}'\big), \tag{14}$$

For intra-layer (slots within the same layer and time):

$$\mathrm{sim}_c^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \big(\boldsymbol{W}_c^{Q,(l)} \boldsymbol{x}\big)^{\top} \big(\boldsymbol{W}_c^{K,(l)} \boldsymbol{x}'\big), \tag{15}$$

For temporal connections (the same slot across different times):

$$\mathrm{sim}_t^{(l)}(\boldsymbol{x}, \boldsymbol{x}') = \big(\boldsymbol{W}_t^{Q,(l)} \boldsymbol{x}\big)^{\top} \big(\boldsymbol{W}_t^{K,(l)} \boldsymbol{x}'\big). \tag{16}$$

Here, $\boldsymbol{W}_{\bullet}^{Q,(l)}$ and $\boldsymbol{W}_{\bullet}^{K,(l)}$ are learnable projection matrices for layer l. The intra-layer and temporal parameters parallel the key-query mechanism in self-attention, while the inter-layer parameters are analogous to the inverted attention mechanism used in slot attention.

Energy At each layer l, the total energy is split into three terms:

$$E_v^{(l)} = -\sum_{t=1}^{T}\sum_{c=1}^{K_{l-1}} \underbrace{\ln\left(\sum_{k=1}^{K_l} \exp\left(\text{sim}_v^{(l)}(\boldsymbol{x}_{c,t}^{(l-1)}, \boldsymbol{x}_{k,t}^{(l)})\right)\right)}_{\text{Inter-layer energy}},$$

$$E_c^{(l)} = -\sum_{t=1}^{T}\sum_{k=1}^{K_l} \underbrace{\ln\left(\sum_{\substack{k'=1 \\ k'\neq k}}^{K_l} \exp\left(\text{sim}_c^{(l)}(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{k',t}^{(l)})\right)\right)}_{\text{Intra-layer energy}}, \qquad (17)$$

$$E_t^{(l)} = -\sum_{k=1}^{K_l}\sum_{t=2}^{T} \underbrace{\ln\left(\sum_{t'<t} \exp\left(\text{sim}_t^{(l)}(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{k,t'}^{(l)})\right)\right)}_{\text{Temporal energy}}.$$

Summing these over all layers $l \in \{1, \ldots, L\}$ defines the total energy E. These terms correspond to inter-layer connections $(E_v^{(l)})$ and match the energy function for slot attention, intra-layer connections $(E_c^{(l)})$ which match the energy function for self attention, and temporal connections $(E_t^{(l)})$, which match the energy function for *causal* self attention (as each variable is only explained by past variables).

Gradients Consider a single variable $\boldsymbol{x}_{k,t}^{(l)}$. Its gradient with respect to the energy decomposes into four parts:

$$-\frac{\partial E}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \underbrace{-\frac{\partial E_v^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}}}_{\text{bottom-up}} + \underbrace{-\frac{\partial E_v^{(l+1)}}{\partial \boldsymbol{x}_{k,t}^{(l)}}}_{\text{top-down}} + \underbrace{-\frac{\partial E_c^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}}}_{\text{intra-layer}} + \underbrace{-\frac{\partial E_t^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}}}_{\text{temporal}}. \qquad (18)$$

Bottom-up: We define a similarity matrix $\boldsymbol{A}_{c,k}$, representing the interactions between $\boldsymbol{x}_{k,t}^{(l)}$ in layer l and $\boldsymbol{x}_{c,t}^{(l-1)}$ in the layer below.

$$\boldsymbol{A}_{c,k} = \text{sim}_v^{(l)}(\boldsymbol{x}_{c,t}^{(l-1)}, \boldsymbol{x}_{k,t}^{(l)}), \qquad (19)$$

The gradient with respect to $\boldsymbol{x}_{k,t}^{(l)}$ is:

$$-\frac{\partial E_v^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \sum_{c=1}^{K_{l-1}} \text{softmax}_k\left(\boldsymbol{A}_{c,k}\right) \boldsymbol{W}_v^{Q,(l)\top} \boldsymbol{W}_v^{K,(l)} \boldsymbol{x}_{c,t}^{(l-1)}. \qquad (20)$$

This term aggregates contributions from the children in layer $l - 1$, weighted by the attention $\text{softmax}_k(\boldsymbol{A}_{c,k})$, and is analogous to Slot Attention or Gaussian mixture models.

Top-down: We define a similarity matrix $\boldsymbol{A}_{k,p}$, representing the interactions between $\boldsymbol{x}_{k,t}^{(l)}$ in layer l and $\boldsymbol{x}_{p,t}^{(l+1)}$ in the layer above.

$$\boldsymbol{A}_{k,p} = \text{sim}_v^{(l+1)}(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{p,t}^{(l+1)}), \qquad (21)$$

The gradient with respect to $\boldsymbol{x}_{k,t}^{(l)}$ is:

$$-\frac{\partial E_v^{(l+1)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \sum_{p=1}^{K_{l+1}} \text{softmax}_p\left(\boldsymbol{A}_{k,p}\right) \boldsymbol{W}_v^{K,(l+1)\top} \boldsymbol{W}_v^{Q,(l+1)} \boldsymbol{x}_{p,t}^{(l+1)}. \qquad (22)$$

This term captures the influence of $\boldsymbol{x}_{k,t}^{(l)}$ being treated as a child, weighted by the attention $\text{softmax}_p(\boldsymbol{A}_{k,p})$, and reflects the top-down influence from layer $l + 1$, and is analogous to Hopfield attention or the parent term in self-attention.

Intra-layer: We define two similarity matrices, $\boldsymbol{A}_{k,k'}$ and $\boldsymbol{A}_{k',k}$, representing the bidirectional interactions between $\boldsymbol{x}_{k,t}^{(l)}$ and other slots $\boldsymbol{x}_{k',t}^{(l)}$:

$$
\begin{aligned}
\boldsymbol{A}_{k,k'} &= \mathrm{sim}_c^{(l)}(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{k',t}^{(l)}), \\
\boldsymbol{A}_{k',k} &= \mathrm{sim}_c^{(l)}(\boldsymbol{x}_{k',t}^{(l)}, \boldsymbol{x}_{k,t}^{(l)}),
\end{aligned}
\tag{23}
$$

where

$$
\mathrm{sim}_c^{(l)}(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{k',t}^{(l)}) = \big(\boldsymbol{W}_c^{Q,(l)} \boldsymbol{x}_{k,t}^{(l)}\big)^\top \big(\boldsymbol{W}_c^{K,(l)} \boldsymbol{x}_{k',t}^{(l)}\big).
$$

The gradient with respect to $\boldsymbol{x}_{k,t}^{(l)}$ is:

$$
-\frac{\partial E_c^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \underbrace{\sum_{k' \neq k} \mathrm{softmax}_k\big(\boldsymbol{A}_{k',k}\big) \boldsymbol{W}_c^{K,(l)\top} \boldsymbol{W}_c^{Q,(l)} \boldsymbol{x}_{k',t}^{(l)}}_{\boldsymbol{x}_{k,t}^{(l)} \text{ acting as parent (explaining others)}}
$$
$$
+ \underbrace{\sum_{k' \neq k} \mathrm{softmax}_{k'}\big(\boldsymbol{A}_{k,k'}\big) \boldsymbol{W}_c^{Q,(l)\top} \boldsymbol{W}_c^{K,(l)} \boldsymbol{x}_{k',t}^{(l)}}_{\boldsymbol{x}_{k,t}^{(l)} \text{ acting as child (being explained by others)}}.
\tag{24}
$$

The first term captures the influence of $\boldsymbol{x}_{k,t}^{(l)}$ as a parent, aggregating the contributions from other slots $\boldsymbol{x}_{k',t}^{(l)}$ it explains, weighted by $\mathrm{softmax}_{k'}(\boldsymbol{A}_{k,k'})$. The second term reflects $\boldsymbol{x}_{k,t}^{(l)}$'s role as a child, being explained by other slots $\boldsymbol{x}_{k',t}^{(l)}$, weighted by $\mathrm{softmax}_k(\boldsymbol{A}_{k',k})$.

Temporal:

Setup. We define a similarity matrix $\boldsymbol{A}_{t,t'}$ representing the interaction between a slot $\boldsymbol{x}_{k,t}^{(l)}$ at time t and its past representation $\boldsymbol{x}_{k,t'}^{(l)}$ at time t', for $t' < t$:

$$
\boldsymbol{A}_{t,t'} = \mathrm{sim}_t^{(l)}\big(\boldsymbol{x}_{k,t}^{(l)}, \boldsymbol{x}_{k,t'}^{(l)}\big) = \big(\boldsymbol{W}_t^{Q,(l)} \boldsymbol{x}_{k,t}^{(l)}\big)^\top \big(\boldsymbol{W}_t^{K,(l)} \boldsymbol{x}_{k,t'}^{(l)}\big).
\tag{25}
$$

Gradient. Assuming causal attention (i.e. no future times), the gradient w.r.t. $\boldsymbol{x}_{k,t}^{(l)}$ is:

$$
-\frac{\partial E_t^{(l)}}{\partial \boldsymbol{x}_{k,t}^{(l)}} = \sum_{t' < t} \mathrm{softmax}_{t'}\big(\boldsymbol{A}_{t,t'}\big) \boldsymbol{W}_t^{Q,(l)\top} \boldsymbol{W}_t^{K,(l)} \boldsymbol{x}_{k,t'}^{(l)}.
\tag{26}
$$

Thus $\boldsymbol{x}_{k,t}^{(l)}$ is influenced only by its own past $\boldsymbol{x}_{k,t'}^{(l)}$ ($t' < t$), as in causal self-attention.

Remark (Full Version). We can generalize to let each latent $\boldsymbol{s}_{k,t}$ attend *all* data features and *all* slot latents at *any* time. This merges the three energies into a single large log-sum-exp, imposing direct competition among data, concurrency, and temporal parents, but increases complexity to $\mathcal{O}((N+K)\,T)^2$.

7 Dynamic Mixture Attention

Setup. We consider a time series of observations $\{\boldsymbol{x}_t\}_{t=1}^T$, where each $\boldsymbol{x}_t \in \mathbb{R}^d$. The transition from $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t+1}$ is governed by K distinct dynamical modes, each parameterized by $\{\boldsymbol{A}_k, \boldsymbol{b}_k\}$ for $k \in \{1, \ldots, K\}$. Additionally, each $\boldsymbol{x}_t$ can attend to K *parents* $\{\boldsymbol{z}_k\}_{k=1}^K$, which may represent previous states $\boldsymbol{x}_{t'}$ for $t' < t$, concurrent slots at the same time step, or variables from other layers. This setup generalizes various attention mechanisms, including causal self-attention and inter-layer attention.

Similarity Functions. We define two similarity measures for each mode k. The first captures the attention-based similarity between the current state $\boldsymbol{x}_t$ and its parent $\boldsymbol{z}_k$ using learnable projection matrices $\boldsymbol{W}^Q$ and $\boldsymbol{W}^K$:

$$
\mathrm{sim}_{\mathrm{parents}}\big(\boldsymbol{x}_t, \boldsymbol{z}_k\big) = \big(\boldsymbol{W}^Q \boldsymbol{x}_t\big)^\top \big(\boldsymbol{W}^K \boldsymbol{z}_k\big).
$$

The second similarity term models the dynamical compatibility between $\boldsymbol{x}_t$ and $\boldsymbol{x}_{t+1}$ under mode k:

$$
\mathrm{sim}_{\mathrm{dynamics}}\big(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}; \boldsymbol{A}_k, \boldsymbol{b}_k\big) = -\tfrac{1}{2} \big\| \boldsymbol{x}_{t+1} - \boldsymbol{A}_k \boldsymbol{x}_t - \boldsymbol{b}_k \big\|^2.
$$

Energy. Each transition $\boldsymbol{x}_t \to \boldsymbol{x}_{t+1}$ is associated with K modes, each combining attention to a parent and a dynamical transformation. The energy is defined as:

$$E^{\text{DMA}}\Big(\{\boldsymbol{x}_t\}, \{\boldsymbol{A}_k, \boldsymbol{b}_k\}, \{\boldsymbol{z}_k\}\Big) = -\sum_{t=1}^{T-1} \ln\left(\sum_{k=1}^{K} \exp\Big(\text{sim}_{\text{parents}}(\boldsymbol{x}_t, \boldsymbol{z}_k) + \text{sim}_{\text{dynamics}}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}; \boldsymbol{A}_k, \boldsymbol{b}_k)\Big)\right).$$
(27)

This formulation allows each mode k to simultaneously attend to a parent $\boldsymbol{z}_k$ and explain the transition through $\boldsymbol{A}_k$ and $\boldsymbol{b}_k$.

Gradients. Define the attention weights as:

$$\alpha_{t,k} = \text{softmax}_k\Big(\text{sim}_{\text{parents}}(\boldsymbol{x}_t, \boldsymbol{z}_k) + \text{sim}_{\text{dynamics}}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}; \boldsymbol{A}_k, \boldsymbol{b}_k)\Big).$$

Then,

$$-\frac{\partial E^{\text{DMA}}}{\partial \boldsymbol{x}_{t+1}} = \sum_{k=1}^{K} \alpha_{t,k}\left[\underbrace{\frac{\partial}{\partial \boldsymbol{x}_{t+1}}\text{sim}_{\text{parents}}(\boldsymbol{x}_t, \boldsymbol{z}_k)}_{=0} + \underbrace{\frac{\partial}{\partial \boldsymbol{x}_{t+1}}\text{sim}_{\text{dynamics}}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}; \boldsymbol{A}_k, \boldsymbol{b}_k)}_{\boldsymbol{x}_{t+1} - \boldsymbol{A}_k\,\boldsymbol{x}_t - \boldsymbol{b}_k}\right]$$

$$= \sum_{k=1}^{K} \alpha_{t,k}\big[\boldsymbol{x}_{t+1} - \boldsymbol{A}_k\,\boldsymbol{x}_t - \boldsymbol{b}_k\big].$$

$$-\frac{\partial E^{\text{DMA}}}{\partial \boldsymbol{x}_t} = \sum_{k=1}^{K} \alpha_{t,k}\left[\underbrace{\frac{\partial}{\partial \boldsymbol{x}_t}\text{sim}_{\text{parents}}(\boldsymbol{x}_t, \boldsymbol{z}_k)}_{\boldsymbol{W}^{Q\top}\boldsymbol{W}^K\boldsymbol{z}_k} + \underbrace{\frac{\partial}{\partial \boldsymbol{x}_t}\text{sim}_{\text{dynamics}}(\boldsymbol{x}_t, \boldsymbol{x}_{t+1}; \boldsymbol{A}_k, \boldsymbol{b}_k)}_{-\boldsymbol{A}_k^\top\big(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k\,\boldsymbol{x}_t - \boldsymbol{b}_k\big)}\right]$$

$$= \sum_{k=1}^{K} \alpha_{t,k}\left[\boldsymbol{W}^{Q\top}\boldsymbol{W}^K\boldsymbol{z}_k - \boldsymbol{A}_k^\top\big(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k\,\boldsymbol{x}_t - \boldsymbol{b}_k\big)\right].$$

Additionally, gradients with respect to the parameters $\boldsymbol{A}_k$ and $\boldsymbol{b}_k$ are given by:

$$-\frac{\partial E^{\text{DMA}}}{\partial \boldsymbol{A}_k} = \sum_{t=1}^{T-1} \alpha_{t,k}\big(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k\,\boldsymbol{x}_t - \boldsymbol{b}_k\big)\boldsymbol{x}_t^\top,$$

$$-\frac{\partial E^{\text{DMA}}}{\partial \boldsymbol{b}_k} = \sum_{t=1}^{T-1} \alpha_{t,k}\big(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k\,\boldsymbol{x}_t - \boldsymbol{b}_k\big).$$

Fixed Point for $\boldsymbol{x}_{t+1}$. Setting the gradient with respect to $\boldsymbol{x}_{t+1}$ to zero yields the fixed point equation:

$$\boldsymbol{x}_{t+1} = \sum_{k=1}^{K} \alpha_{t,k}\big(\boldsymbol{A}_k\,\boldsymbol{x}_t + \boldsymbol{b}_k\big),$$

where $\alpha_{t,k}$ itself depends on $\boldsymbol{x}_{t+1}$. This equation must be solved iteratively, typically via fixed-point iteration or gradient-based optimization methods to achieve consistency.

8 Predictive coding

Setup. We again have child vectors $\{\boldsymbol{x}_i\}_{i=1}^{N}$ and a set of parent $\boldsymbol{\mu}_k \in \mathbb{R}^d$, $k = 1, \ldots, K$. However, rather than a direct difference $\boldsymbol{x}_i - \boldsymbol{\mu}_k$, let us assume the *model* maps $\boldsymbol{\mu}_k$ through some non-linear function $f_\phi(\cdot)$ before comparing to $\boldsymbol{x}_i$. For instance, f_ϕ could be a neural network.

Similarity function. Define

$$\text{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k) = -\tfrac{1}{2}\left\|\boldsymbol{x}_i - f_\phi(\boldsymbol{\mu}_k)\right\|^2.$$

Energy.

$$E^{\mathrm{PC}}\Big(\{\boldsymbol{x}_i\},\{\boldsymbol{\mu}_k\}\Big) \;=\; -\sum_{i=1}^{N}\ln\Big(\sum_{k=1}^{K}\exp\Big(-\tfrac{1}{2}\left\|\boldsymbol{x}_i - f_\phi(\boldsymbol{\mu}_k)\right\|^2\Big)\Big). \tag{28}$$

Gradients.

$$\alpha_{i,k} \;=\; \mathrm{softmax}_k\Big(-\tfrac{1}{2}\left\|\boldsymbol{x}_i - f_\phi(\boldsymbol{\mu}_k)\right\|^2\Big).$$

$$-\frac{\partial E^{\mathrm{PC}}}{\partial \boldsymbol{x}_i} \;=\; \sum_{k=1}^{K}\alpha_{i,k}\big(\boldsymbol{x}_i - f_\phi(\boldsymbol{\mu}_k)\big).$$

$$-\frac{\partial E^{\mathrm{PC}}}{\partial \boldsymbol{\mu}_k} \;=\; \sum_{i=1}^{N}\alpha_{i,k}\;\underbrace{\frac{\partial f_\phi(\boldsymbol{\mu}_k)}{\partial \boldsymbol{\mu}_k}}_{\text{Jacobian of } f_\phi}\;\big(\boldsymbol{x}_i - f_\phi(\boldsymbol{\mu}_k)\big).$$

Here, $\frac{\partial f_\phi(\boldsymbol{\mu}_k)}{\partial \boldsymbol{\mu}_k}$ is the $d \times d$ Jacobian (or more general shape if $\boldsymbol{\mu}_k$ and $f_\phi(\boldsymbol{\mu}_k)$ differ in dimension).

9   Cross Attention

Setup.   We have a set of child vectors (queries) $\boldsymbol{Q} \in \mathbb{R}^{d \times N_Q}$ and a set of parent vectors (keys) $\boldsymbol{K} \in \mathbb{R}^{d \times N_K}$. Let

$$C \;=\; \{1,\ldots,N_Q\}, \quad P \;=\; \{1,\ldots,N_K\},$$

so $\boldsymbol{v}_c = \boldsymbol{q}_c$ is the c-th query, and $\boldsymbol{v}_p = \boldsymbol{k}_p$ is the p-th key. Suppose we have learnable weight matrices $\boldsymbol{W}^Q, \boldsymbol{W}^K \in \mathbb{R}^{d \times d}$. Then

$$\boldsymbol{q}_c \;=\; \boldsymbol{W}^Q \boldsymbol{x}_c^Q, \quad \boldsymbol{k}_p \;=\; \boldsymbol{W}^K \boldsymbol{x}_p^K,$$

where $\boldsymbol{x}_c^Q$ is the raw c-th query token and $\boldsymbol{x}_p^K$ the raw p-th key token.

Similarity function.

$$\mathrm{sim}\big(\boldsymbol{q}_c, \boldsymbol{k}_p\big) \;=\; \boldsymbol{q}_c^\top \boldsymbol{k}_p.$$

Energy.

$$E^{\mathrm{Cross}}\Big(\{\boldsymbol{q}_c\},\{\boldsymbol{k}_p\}\Big) \;=\; -\sum_{c=1}^{N_Q}\ln\Big(\sum_{p=1}^{N_K}\exp\big(\boldsymbol{q}_c^\top \boldsymbol{k}_p\big)\Big). \tag{29}$$

Gradients.

$$-\frac{\partial E^{\mathrm{Cross}}}{\partial \boldsymbol{q}_c} \;=\; \sum_{p=1}^{N_K}\mathrm{softmax}_p\Big(\boldsymbol{q}_c^\top \boldsymbol{k}_p\Big)\,\boldsymbol{k}_p. \tag{30}$$

$$-\frac{\partial E^{\mathrm{Cross}}}{\partial \boldsymbol{k}_p} \;=\; \sum_{c=1}^{N_Q}\mathrm{softmax}_p\Big(\boldsymbol{q}_c^\top \boldsymbol{k}_p\Big)\,\boldsymbol{q}_c. \tag{31}$$

When mapping back to the raw tokens $\boldsymbol{x}_c^Q$ or $\boldsymbol{x}_p^K$, chain-rule multiplies by $\boldsymbol{W}^Q$ or $\boldsymbol{W}^K$, respectively.

10   Switching Linear Dynamical System

Setup.   Consider a time series of observations $\{\boldsymbol{x}_t\}_{t=1}^{T}$, where each $\boldsymbol{x}_t \in \mathbb{R}^d$. We assume there are K distinct (linear) dynamical modes, each with parameters $\{\boldsymbol{A}_k, \boldsymbol{b}_k, \boldsymbol{\Sigma}_k\}$. Let π_k be the mixing weight of mode k. A typical *switching linear dynamical system* (SLDS) posits:

$$\boldsymbol{x}_{t+1} \;\approx\; \boldsymbol{A}_k\,\boldsymbol{x}_t \;+\; \boldsymbol{b}_k, \quad k \in \{1,\ldots,K\},$$

with Gaussian noise $\boldsymbol{\Sigma}_k$. We treat $\boldsymbol{x}_{t+1}$ as a *child* and the $\{\boldsymbol{A}_k, \boldsymbol{b}_k\}$ (together with $\boldsymbol{x}_t$) as *parents* in a mixture-of-linear-dynamics fashion.

Similarity function. Define, for each mode k,

$$\text{sim}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t; \boldsymbol{A}_k, \boldsymbol{b}_k) = \ln \pi_k - \tfrac{1}{2}(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \boldsymbol{x}_t - \boldsymbol{b}_k)^\top \Sigma_k^{-1}(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \boldsymbol{x}_t - \boldsymbol{b}_k).$$

Energy. Summing over all time steps $t = 1, \ldots, T-1$, we write

$$E^{\text{SLDS}}\big(\{\boldsymbol{x}_t\}, \{\boldsymbol{A}_k, \boldsymbol{b}_k\}\big) = -\sum_{t=1}^{T-1} \ln\Big(\sum_{k=1}^{K} \exp\big(\text{sim}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t; \boldsymbol{A}_k, \boldsymbol{b}_k)\big)\Big). \tag{32}$$

Gradients. Let

$$\alpha_{t,k} = \text{softmax}_k\Big(\text{sim}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t; \boldsymbol{A}_k, \boldsymbol{b}_k)\Big),$$

i.e. the normalized exponent for mode k.

$$-\frac{\partial E^{\text{SLDS}}}{\partial \boldsymbol{x}_{t+1}} = \sum_{k=1}^{K} \alpha_{t,k} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \boldsymbol{x}_t - \boldsymbol{b}_k).$$

$$-\frac{\partial E^{\text{SLDS}}}{\partial \boldsymbol{x}_t} = \sum_{k=1}^{K} \alpha_{t,k}\big(-\boldsymbol{A}_k^\top \boldsymbol{\Sigma}_k^{-1}\big)(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \boldsymbol{x}_t - \boldsymbol{b}_k) \quad (t = 1, \ldots, T-1).$$

$$-\frac{\partial E^{\text{SLDS}}}{\partial \boldsymbol{A}_k} = \sum_{t=1}^{T-1} \alpha_{t,k} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \boldsymbol{x}_t - \boldsymbol{b}_k) \boldsymbol{x}_t^\top.$$

$$-\frac{\partial E^{\text{SLDS}}}{\partial \boldsymbol{b}_k} = \sum_{t=1}^{T-1} \alpha_{t,k} \boldsymbol{\Sigma}_k^{-1}(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \boldsymbol{x}_t - \boldsymbol{b}_k).$$

11 Layer Normalization

Setup. We consider a batch of input vectors $\{\boldsymbol{x}_i\}_{i=1}^{B}$, where each vector $\boldsymbol{x}_i \in \mathbb{R}^D$. Each vector is normalized by subtracting the mean and dividing by the standard deviation, with learnable scaling and bias parameters $\gamma, \delta \in \mathbb{R}^D$. For numerical stability, a small constant $\epsilon > 0$ is added to the variance.

Energy. The energy for layer normalization is given by:

$$E^{\text{LN}}(\{\boldsymbol{x}_i\}) = \sum_{i=1}^{B} \left[\gamma \sqrt{\frac{1}{D} \sum_{j=1}^{D}(x_{ij} - \bar{\boldsymbol{x}}_i)^2 + \epsilon} + \sum_{j=1}^{D} \delta_j x_{ij} \right],$$

where:

$$\bar{\boldsymbol{x}}_i = \frac{1}{D} \sum_{j=1}^{D} x_{ij}.$$

Derivative. The normalized outputs are obtained as the derivative of the energy with respect to the inputs:

$$\frac{\partial E^{\text{LN}}}{\partial x_{ij}} = \gamma_j \frac{x_{ij} - \bar{\boldsymbol{x}}_i}{\sqrt{\frac{1}{D} \sum_{k=1}^{D}(x_{ik} - \bar{\boldsymbol{x}}_i)^2 + \epsilon}} + \delta_j.$$

$$\bar{\boldsymbol{x}}_i = \frac{1}{D} \sum_{k=1}^{D} x_{ik}.$$

12 Coordinate ascent

Consider an energy of the form

$$E(\{\boldsymbol{x}_j\}, \{\boldsymbol{\mu}_i\}; \theta) \;=\; -\sum_{j=1}^{N} \ln\Big(\sum_{i=1}^{S} \exp\big(\mathrm{sim}_\theta(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\big)\Big),$$

with θ denoting the parameters of the similarity function. In an EM procedure, we alternate:

- **E-step**: Update latent variables using fixed θ.
- **M-step**: Update parameters θ in closed form given fixed latent assignments.

For *slot attention*, we have:

$$\theta = \{\boldsymbol{W}_K, \boldsymbol{W}_Q\},$$
$$\mathrm{sim}_\theta(\boldsymbol{x}_j, \boldsymbol{\mu}_i) \;=\; (\boldsymbol{W}_K \boldsymbol{x}_j)^\top (\boldsymbol{W}_Q \boldsymbol{\mu}_i).$$
$$\mathbf{A} \quad \text{with entries} \quad A_{ji} \;=\; \mathrm{softmax}_i\Big((\boldsymbol{W}_K \boldsymbol{x}_j)^\top (\boldsymbol{W}_Q \boldsymbol{\mu}_i)\Big).$$

E-step: Update Slots

Fix $\boldsymbol{W}_K, \boldsymbol{W}_Q$. The gradient for each slot $\boldsymbol{\mu}_i$ is

$$-\frac{\partial E^{\mathrm{Slot}}}{\partial \boldsymbol{\mu}_i} = \sum_{j=1}^{N} A_{ji}\, \boldsymbol{W}_Q^\top \boldsymbol{W}_K \boldsymbol{x}_j.$$

Use this gradient to iteratively update $\{\boldsymbol{\mu}_i\}$ until convergence.

M-step: Update Parameters

Given fixed slots $\{\boldsymbol{\mu}_i\}$ and attention matrix $\mathbf{A}$, we aim to update the parameters $\boldsymbol{W}_K, \boldsymbol{W}_Q$. This is motivated by setting the gradient of the energy with respect to these parameters to zero:

$$-\frac{\partial E^{\mathrm{Slot}}}{\partial \boldsymbol{W}_K} = 0, \qquad -\frac{\partial E^{\mathrm{Slot}}}{\partial \boldsymbol{W}_Q} = 0.$$

Under the fixed assignments provided by $\mathbf{A}$ and slots $\{\boldsymbol{\mu}_i\}$, these conditions are equivalent to solving a weighted least-squares problem. Specifically, we consider minimizing the objective

$$\min_{\boldsymbol{W}_K, \boldsymbol{W}_Q} \sum_{j=1}^{N} \sum_{i=1}^{S} A_{ji} \left\| \boldsymbol{W}_K \boldsymbol{x}_j - \boldsymbol{W}_Q \boldsymbol{\mu}_i \right\|^2,$$

since the stationary point of this quadratic form corresponds to zero gradients with respect to $\boldsymbol{W}_K, \boldsymbol{W}_Q$.

Update $\boldsymbol{W}_Q$:   Differentiate w.r.t. $\boldsymbol{W}_Q$, set to zero:

$$\sum_{j,i} A_{ji} \left(\boldsymbol{W}_K \boldsymbol{x}_j - \boldsymbol{W}_Q \boldsymbol{\mu}_i \right) \boldsymbol{\mu}_i^\top \;=\; 0,$$

yielding

$$\boldsymbol{W}_Q \;=\; \left(\sum_{j,i} A_{ji}\, \boldsymbol{W}_K \boldsymbol{x}_j\, \boldsymbol{\mu}_i^\top \right) \left(\sum_{j,i} A_{ji}\, \boldsymbol{\mu}_i \boldsymbol{\mu}_i^\top \right)^{-1}.$$

Update $\boldsymbol{W}_K$:   Similarly, differentiate w.r.t. $\boldsymbol{W}_K$, set to zero:

$$\sum_{j,i} A_{ji} \left(\boldsymbol{W}_K \boldsymbol{x}_j - \boldsymbol{W}_Q \boldsymbol{\mu}_i \right) \boldsymbol{x}_j^\top \;=\; 0,$$

yielding

$$\boldsymbol{W}_K \;=\; \left(\sum_{j,i} A_{ji}\, \boldsymbol{W}_Q \boldsymbol{\mu}_i\, \boldsymbol{x}_j^\top \right) \left(\sum_{j,i} A_{ji}\, \boldsymbol{x}_j \boldsymbol{x}_j^\top \right)^{-1}.$$

Iterate: Alternate between the E-step (updating $\{\boldsymbol{\mu}_i\}$) and the M-step (updating $\boldsymbol{W}_K, \boldsymbol{W}_Q$) until convergence.

13 Capsule Networks via Energy-based Hierarchical Routing

In this section, we extend the general energy-based attention framework to formulate a capsule network architecture. In our approach, the dynamic routing between lower-level (child) and higher-level (parent) capsules is derived as gradient updates that minimize an energy function defined via a log-sum-exp over similarity scores. This provides a principled probabilistic interpretation for routing and enables an EM-like procedure for updating both the capsule representations and the associated transformation parameters.

Setup

Let the lower-level capsules (or feature detectors) be denoted by

$$\boldsymbol{x}_j \in \mathbb{R}^{d_x}, \quad j \in \{1, \dots, N\},$$

and the higher-level capsules (or entity representations) by

$$\boldsymbol{\mu}_i \in \mathbb{R}^{d_\mu}, \quad i \in \{1, \dots, S\}.$$

Each lower-level capsule is projected into a set of *vote vectors* for the higher-level capsules. For simplicity, we introduce learnable linear maps $\boldsymbol{W}_K \in \mathbb{R}^{d' \times d_x}$ for the children and $\boldsymbol{W}_Q \in \mathbb{R}^{d' \times d_\mu}$ for the parents, and define the similarity between a child and a parent as

$$\text{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i) = (\boldsymbol{W}_K \boldsymbol{x}_j)^\top (\boldsymbol{W}_Q \boldsymbol{\mu}_i). \tag{33}$$

This similarity serves as a measure of compatibility between the vote of a lower-level capsule and the state of a higher-level capsule.

Energy Formulation

Following the general framework, we define an energy function for the routing process:

$$E^{\text{Caps}}(\{\boldsymbol{x}_j\}, \{\boldsymbol{\mu}_i\}) = -\sum_{j=1}^{N} \ln\left(\sum_{i=1}^{S} \exp\left(\text{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\right)\right). \tag{34}$$

Minimizing E^{Caps} encourages every lower-level capsule to be well explained by at least one higher-level capsule, thereby establishing a soft assignment between children and parents.

Dynamic Routing via Energy Minimization

To route information from lower-level capsules to higher-level capsules, we minimize the energy with respect to the parent representations $\{\boldsymbol{\mu}_i\}$. The gradient with respect to a parent capsule is

$$-\frac{\partial E^{\text{Caps}}}{\partial \boldsymbol{\mu}_i} = \sum_{j=1}^{N} \text{softmax}_i\left(\text{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i)\right) \frac{\partial}{\partial \boldsymbol{\mu}_i} \text{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i). \tag{35}$$

Given our similarity function,

$$\frac{\partial}{\partial \boldsymbol{\mu}_i} \text{sim}(\boldsymbol{x}_j, \boldsymbol{\mu}_i) = \boldsymbol{W}_Q^\top (\boldsymbol{W}_K \boldsymbol{x}_j),$$

so that the update for each capsule becomes

$$\boldsymbol{\mu}_i \leftarrow \boldsymbol{\mu}_i + \eta \sum_{j=1}^{N} \underbrace{\text{softmax}_i\left((\boldsymbol{W}_K \boldsymbol{x}_j)^\top (\boldsymbol{W}_Q \boldsymbol{\mu}_i)\right)}_{c_{ij}} \boldsymbol{W}_Q^\top (\boldsymbol{W}_K \boldsymbol{x}_j), \tag{36}$$

where η is a step-size parameter and

$$c_{ij} = \text{softmax}_i\left((\boldsymbol{W}_K \boldsymbol{x}_j)^\top (\boldsymbol{W}_Q \boldsymbol{\mu}_i)\right)$$

plays the role of the routing coefficient, assigning a weight to the contribution of child j for parent i.

Interpretation as Capsule Routing

The routing coefficients c_{ij} determine how strongly each lower-level capsule $\boldsymbol{x}_j$ votes for a particular higher-level capsule $\boldsymbol{\mu}_i$. In our framework, these coefficients emerge naturally from the softmax over the energy-based similarity, and the capsule update is given by

$$\boldsymbol{\mu}_i \leftarrow \sum_{j=1}^{N} c_{ij}\, \tilde{\boldsymbol{v}}_{ij}, \quad \text{with} \quad \tilde{\boldsymbol{v}}_{ij} = \boldsymbol{W}_Q^\top \big(\boldsymbol{W}_K\, \boldsymbol{x}_j\big). \tag{37}$$

This update mirrors the dynamic routing mechanism in capsule networks, where each parent capsule aggregates "votes" from the children weighted by their routing coefficients.

EM-like Parameter Updates

In addition to updating the capsule representations, we can update the transformation matrices $\boldsymbol{W}_K$ and $\boldsymbol{W}_Q$ in an M-step. Given fixed capsules and routing coefficients, we minimize the weighted least-squares objective

$$\min_{\boldsymbol{W}_K, \boldsymbol{W}_Q} \sum_{j=1}^{N} \sum_{i=1}^{S} c_{ij} \left\| \boldsymbol{W}_K\, \boldsymbol{x}_j - \boldsymbol{W}_Q\, \boldsymbol{\mu}_i \right\|^2. \tag{38}$$

The stationary conditions for this quadratic form yield closed-form updates:

$$\boldsymbol{W}_Q = \left(\sum_{j,i} c_{ij}\, \boldsymbol{W}_K\, \boldsymbol{x}_j\, \boldsymbol{\mu}_i^\top \right) \left(\sum_{j,i} c_{ij}\, \boldsymbol{\mu}_i\, \boldsymbol{\mu}_i^\top \right)^{-1},$$

$$\boldsymbol{W}_K = \left(\sum_{j,i} c_{ij}\, \boldsymbol{W}_Q\, \boldsymbol{\mu}_i\, \boldsymbol{x}_j^\top \right) \left(\sum_{j,i} c_{ij}\, \boldsymbol{x}_j\, \boldsymbol{x}_j^\top \right)^{-1}.$$

By alternating between the E-step (updating $\{\boldsymbol{\mu}_i\}$ via gradient descent on E^{Caps}) and the M-step (updating $\boldsymbol{W}_K$ and $\boldsymbol{W}_Q$), we obtain a complete EM-like algorithm that jointly optimizes the capsule representations and the transformation parameters.

Summary

This section has developed a capsule network model within our unified energy-based attention framework. The routing of information from lower-level capsules to higher-level capsules is interpreted as minimizing an energy function defined by log-sum-exp similarities. The softmax over similarities naturally produces routing coefficients, and iterative gradient updates correspond to the dynamic routing procedure of capsule networks. An EM-like parameter update further refines the transformation matrices, providing a principled and interpretable approach to building capsule networks.

A   Appendix: Derivation of Gradient Updates

Let us consider a generic term:

$$-\ln\Big(\sum_{m=1}^{M}\exp\big(f_m(\boldsymbol{x})\big)\Big),$$

where $\boldsymbol{x}\in\mathbb{R}^d$ is some variable, and each f_m is a scalar function. We compute its gradient:

$$\frac{\partial}{\partial\boldsymbol{x}}\left[-\ln\Big(\sum_{m=1}^{M}\exp\big(f_m(\boldsymbol{x})\big)\Big)\right] = -\frac{1}{\sum_{m'}\exp\big(f_{m'}(\boldsymbol{x})\big)}\sum_{m=1}^{M}\exp\big(f_m(\boldsymbol{x})\big)\frac{\partial f_m(\boldsymbol{x})}{\partial\boldsymbol{x}}$$

$$= -\sum_{m=1}^{M}\left[\frac{\exp\big(f_m(\boldsymbol{x})\big)}{\sum_{m'}\exp\big(f_{m'}(\boldsymbol{x})\big)}\right]\frac{\partial f_m(\boldsymbol{x})}{\partial\boldsymbol{x}}.$$

Defining $\mathrm{softmax}_m\big(f(\boldsymbol{x})\big) = \frac{\exp(f_m(\boldsymbol{x}))}{\sum_{m'}\exp(f_{m'}(\boldsymbol{x}))}$, this is

$$-\sum_{m=1}^{M}\mathrm{softmax}_m\big(f(\boldsymbol{x})\big)\frac{\partial f_m(\boldsymbol{x})}{\partial\boldsymbol{x}},$$

which matches the softmax-weighted gradient structure.

B   Appendix

Here, we collect the explicit partial derivatives of sim for each model discussed.

Gaussian Mixture Models

$$\mathrm{sim}\big(\boldsymbol{x}_i,\boldsymbol{\mu}_k\big) = \ln\pi_k - \tfrac{1}{2}\big(\boldsymbol{x}_i-\boldsymbol{\mu}_k\big)^{\top}\boldsymbol{\Sigma}_k^{-1}\big(\boldsymbol{x}_i-\boldsymbol{\mu}_k\big).$$

$$\frac{\partial}{\partial\boldsymbol{\mu}_k}\mathrm{sim}(\boldsymbol{x}_i,\boldsymbol{\mu}_k) = \boldsymbol{\Sigma}_k^{-1}\big(\boldsymbol{x}_i-\boldsymbol{\mu}_k\big).$$

Cross Attention

$$\mathrm{sim}\big(\boldsymbol{q}_c,\boldsymbol{k}_p\big) = \boldsymbol{q}_c^{\top}\boldsymbol{k}_p.$$

$$\frac{\partial}{\partial\boldsymbol{q}_c}\mathrm{sim}(\boldsymbol{q}_c,\boldsymbol{k}_p) = \boldsymbol{k}_p, \quad \frac{\partial}{\partial\boldsymbol{k}_p}\mathrm{sim}(\boldsymbol{q}_c,\boldsymbol{k}_p) = \boldsymbol{q}_c.$$

Hopfield Networks

$$\mathrm{sim}\big(\boldsymbol{x}_i,\boldsymbol{m}_\mu\big) = \boldsymbol{x}_i^{\top}\boldsymbol{m}_\mu.$$

$$\frac{\partial}{\partial\boldsymbol{x}_i}\mathrm{sim}(\boldsymbol{x}_i,\boldsymbol{m}_\mu) = \boldsymbol{m}_\mu, \quad \frac{\partial}{\partial\boldsymbol{m}_\mu}\mathrm{sim}(\boldsymbol{x}_i,\boldsymbol{m}_\mu) = \boldsymbol{x}_i.$$

Slot Attention

$$\mathrm{sim}\big(\boldsymbol{x}_j,\boldsymbol{\mu}_i\big) = \big(\boldsymbol{W}_K\,\boldsymbol{x}_j\big)^{\top}\big(\boldsymbol{W}_Q\,\boldsymbol{\mu}_i\big).$$

$$\frac{\partial}{\partial\boldsymbol{x}_j}\mathrm{sim}(\boldsymbol{x}_j,\boldsymbol{\mu}_i) = \boldsymbol{W}_K^{\top}\,\boldsymbol{W}_Q\,\boldsymbol{\mu}_i, \quad \frac{\partial}{\partial\boldsymbol{\mu}_i}\mathrm{sim}(\boldsymbol{x}_j,\boldsymbol{\mu}_i) = \boldsymbol{W}_Q^{\top}\,\boldsymbol{W}_K\,\boldsymbol{x}_j.$$

Self-Attention

$$\mathrm{sim}\big(\boldsymbol{x}_c,\boldsymbol{x}_p\big) = \big(\boldsymbol{W}^Q\boldsymbol{x}_c\big)^{\top}\big(\boldsymbol{W}^K\boldsymbol{x}_p\big).$$

$$\frac{\partial}{\partial\boldsymbol{x}_c}\mathrm{sim}(\boldsymbol{x}_c,\boldsymbol{x}_p) = \boldsymbol{W}^{Q\top}\,\boldsymbol{W}^K\,\boldsymbol{x}_p, \quad \frac{\partial}{\partial\boldsymbol{x}_p}\mathrm{sim}(\boldsymbol{x}_c,\boldsymbol{x}_p) = \boldsymbol{W}^{K\top}\,\boldsymbol{W}^Q\,\boldsymbol{x}_c.$$

Switching Linear Dynamical System

$$\text{sim}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t; \boldsymbol{A}_k, \boldsymbol{b}_k) \;=\; \ln \pi_k \;-\; \tfrac{1}{2} \left(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \, \boldsymbol{x}_t - \boldsymbol{b}_k\right)^{\top} \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \, \boldsymbol{x}_t - \boldsymbol{b}_k\right).$$

$$\frac{\partial}{\partial \boldsymbol{x}_{t+1}} \text{sim}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t; \boldsymbol{A}_k, \boldsymbol{b}_k) \;=\; \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \, \boldsymbol{x}_t - \boldsymbol{b}_k\right),$$

$$\frac{\partial}{\partial \boldsymbol{x}_t} \text{sim}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t; \boldsymbol{A}_k, \boldsymbol{b}_k) \;=\; -\boldsymbol{A}_k^{\top} \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \, \boldsymbol{x}_t - \boldsymbol{b}_k\right),$$

$$\frac{\partial}{\partial \boldsymbol{A}_k} \text{sim}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t; \boldsymbol{A}_k, \boldsymbol{b}_k) \;=\; \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \, \boldsymbol{x}_t - \boldsymbol{b}_k\right) \boldsymbol{x}_t^{\top},$$

$$\frac{\partial}{\partial \boldsymbol{b}_k} \text{sim}(\boldsymbol{x}_{t+1}, \boldsymbol{x}_t; \boldsymbol{A}_k, \boldsymbol{b}_k) \;=\; \boldsymbol{\Sigma}_k^{-1} \left(\boldsymbol{x}_{t+1} - \boldsymbol{A}_k \, \boldsymbol{x}_t - \boldsymbol{b}_k\right).$$

Predictive Coding

$$\text{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k) \;=\; -\tfrac{1}{2} \left\| \boldsymbol{x}_i \;-\; f_\phi(\boldsymbol{\mu}_k) \right\|^2.$$

$$\frac{\partial}{\partial \boldsymbol{x}_i} \text{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k) \;=\; \boldsymbol{x}_i - f_\phi(\boldsymbol{\mu}_k),$$

$$\frac{\partial}{\partial \boldsymbol{\mu}_k} \text{sim}(\boldsymbol{x}_i, \boldsymbol{\mu}_k) \;=\; \frac{\partial f_\phi(\boldsymbol{\mu}_k)}{\partial \boldsymbol{\mu}_k} \left(\boldsymbol{x}_i - f_\phi(\boldsymbol{\mu}_k)\right).$$