Self-Attention Architecture

Input Image (In x In) · 5x5 s=1 · convA

ReLU ConvA — 32 · In-4 · 3x3 s=1 32 · convB · In-6

ReLU ConvB — 64 · In-6 · 3x3 s=1 64 · convC · In-8

ReLU ConvC — 64 · In-8 · 3x3 s=2 64 · convD · (In-11)/2+1

ReLU ConvD — 128 · (In-11)/2+1 · (In-11)/2+1 · 128 filters · Depthwise Conv2D · 128

Split into 8D Capsules & squash (no particular order)

Primary Capsules · $W_{ij}[8 \times 16]$

Capsule Votes (u) — (each primary capsule makes as many votes as the number of DigitCaps)

DigitCaps · Self-Attention · 16 · 10 · L2 · 10

Self-Attention

For each parent capsule j=p:

Q × K → $A_{j=p}$ [16 x 16] × V → **NewEmbeddings**

Algorithm 1 · $V_{j=p}$

Algorithm 2 · $A_{j=p}$ — *Compute sum of each row. This is the similarity score.* → sc1, sc2, sc3, sc4, sc5, sc6, sc7, sc8, sc9, sc10, sc11, sc12, sc13, sc14, sc15, sc16

Find winner row - the one with the maximum sc - (after optional softmax) and use only that one to compute $v_{j=p}$.

Winner Row × V

Optional Softmax of SC across j in Ω_{L+1} so that $\Sigma_j sc_{ji} = 1$ (each part can belong to one whole).

$V_{j=p}$

$$A_{j=p} = \hat{u}_{j=p|i} \times (\hat{u}_{j=p|i})^T / \sqrt{D}$$

$$NewEmbeddings_{j=p} = A_{j=p} \times \hat{u}_{j=p|i}$$