Deep Recurrent Relation Transformer

Belanger Nzakimuena, C. M.

January 2022

A mathematical description of the Deep Recurrent Relation Transformer (DRRT) is provided. DRRT consists of a Deep Relation Transformer (DRT) as implemented by Song et al (2021), enhanced with local recurrence.

1 Feature Extraction

$A = \{a_1, a_2, a_3, ..., a_n\}$ is the features corresponding to the first input x_1.
$B = \{b_1, b_2, b_3, ..., b_m\}$ is the features corresponding to the second input x_2.

Feature extraction stage local recurrence feed-forward equations for A' are given by,

$$h_t^A = \sigma_h(i_t) = \sigma_h(U_h A_t + V_h h_{t-1}^A + b_h) \tag{1}$$

$$A'_t = \sigma_{A'}(j_t) = \sigma_{A'}(W_{A'} h_t^A + b_h) \tag{2}$$

where

A_t is the local RNN input vector
h_t^A is the hidden layer output vector
A'_t is the local RNN output vector
b_h is the bias vector
$U_h, W_{A'}$ are parameter matrices
V_h is a parameter matrice
$\sigma_h, \sigma_{A'}$ are activation functions

Feature extraction stage local recurrence feed-forward ouput B' is calculated in a similar way.

2 Relation Modules

2.1 Global Relation Module

Global relation maps (GRM) for second input features α^g and for first input features β^g are calculated by,

$$\alpha^g = Softmax\left(\frac{\Psi^g(A', B')}{\sqrt{d_{B'}}}\right) \tag{3}$$

$$\beta^g = Softmax\left(\frac{\Psi^g(A', B')}{\sqrt{d_{A'}}}\right) \tag{4}$$

where

$d_{B'}$ represents the dimension of second input features
$d_{A'}$ represents the dimension of first input features

The global pairwise function Ψ^g is calculated by,

$$\Psi^g_{ij}(A', B') = W_g[W_{a'}a'_i; W_{b'}b'_j] \tag{5}$$

where

a'_i is the i-th first input region, b'_j is the j-th second input region
W_g, $W_{a'}$ and $W_{b'}$ are weight matrices to be learned
$[W_{a'}a'_i; W_{b'}b'_j]$ denotes the concatenation of $W_{a'}a'_i$ and $W_{b'}b'_j$

2.2 Guided Regional Relation Module

First and second input features are divided into K regions,

$$A'^R = \left\{A_1^{'R}, A_2^{'R}, A_3^{'R}, ..., A_K^{'R}\right\}$$
$$B'^R = \left\{B_1^{'R}, B_2^{'R}, B_3^{'R}, ..., B_K^{'R}\right\}$$

$B_k^{'R}$ and $A_k^{'R}$ denote second and first input features in k-th corresponding region,

where

$k \in \{1, 2, 3, ..., K\}$ is the region index

Regional relation maps for second input features in k-th region $\alpha^{r(k)}$ is defined as,

$$\alpha^{r(k)} = \frac{1}{C_r} Softmax\left(\Psi^{r(k)}(A', B')\right) \tag{6}$$

Where $\Psi^{r(k)}(A', B')$ projects the topographic correspondence $B' - A'$ feature vector to a scalar expressed as,

$$\Psi_{ij}^{r(k)} = R_{ij}^{r(k)} W_r [W_{a'} a'_i; W_{b'} b'_j] \tag{7}$$

with

$$R_{ij}^{r(k)} = \begin{cases} 1, a'_i \in A_k^{'R} \, and \, b'_j \in B_k^{'R} \\ 0, otherwise \end{cases} \tag{8}$$

Where region pair factor $R_{ij}^{r(k)}$ is 1 when a'_i and b'_j are in the k-th topographic correspondence region pair or 0 when they are not in the k-th topographic correspondence region.
W_r, $W_{a'}$ and $W_{b'}$ are learnable weight vectors which are weight sharing for computing K regional relations.

The normalization factor is given by,

$$C_r = \frac{\sum_{i,j,k} \left(R_{ij}^{r(k)} \right)}{\sqrt{d_{B'}}} \tag{9}$$

Regional relation maps for first input features in k-th region $\beta^{r(k)}$ is calculated in a similar way.

3 Interaction Transformer Module (RNN enhanced)

Relation modules stage local recurrence feed-forward equations for $\alpha^{'g}$ and $\beta^{'g}$, and $\alpha^{'r(k)}$ and $\beta^{'r(k)}$ are calculated in a similar way as outputs at the feature extraction stage.

The information transformation process can be formulated as,

$$T_{b' \to a'_i}^{\sigma} = \sum_{\forall j} \alpha_{ij}^{'\sigma} W_{\tilde{b'}} b'_j \tag{10}$$

$$T_{a' \to b'_j}^{\sigma} = \sum_{\forall i} \beta_{ij}^{'\sigma} W_{\tilde{a'}} a'_i \tag{11}$$

where

$\sigma \in g, r(k)$
$W_{\tilde{b'}}$ and $W_{\tilde{a'}}$ are weight matrices to be learned

The outputs of the information transformation process is combined as,

$$A'_{b'\to a'} = Fusion\left(T^{g}_{b'\to a'}, \sum_{k} T^{r(k)}_{b'\to a'}\right) W_t \tag{12}$$

$$B'_{a'\to b'} = Fusion\left(T^{g}_{a'\to b'}, \sum_{k} T^{r(k)}_{a'\to b'}\right) W_t \tag{13}$$

where

$T^{g}_{b'\to a'}$ is computed based on the global relation map
$T^{r(k)}_{b'\to a'}$ is computed based on the k-th regional relation map
W_t is the parameter matrices to be learned
W_t in Eq. 12 and Eq. 13 are weight sharing
$A'_{b'\to a'}$ is the transformed first input features
$B'_{a'\to b'}$ is the transformed second input features

Three different fusion methods (sum, maxout or concatenation) can be used to combine global and regional representations.

The original and transformed features are combined as the output of the interaction transformer module as,

$$Z_{ab} = A' + A'_{b'\to a'} \tag{14}$$

$$Z_{ba} = B' + B'_{a'\to b'} \tag{15}$$

Second input (Z_{ba}) and first input (Z_{ab}) representations are combined by average fusion then passed to a three-layer fully connected classifier to generate a classification.