Conditional GRU with attention, plus additional encoder for previous sentence and also an additional attention mechanism. Similar to Jean et al. 2017.

1 Arbitrary depth in deep transition decoder with additional attention

Arbitrary transition depth L_t, additional annotation vectors C_2 produced by additional encoder, and additional attention mechanism ATT_2:

$$s_{j,1} = GRU_1(y_{j-1}, s_{j-1,L_t})$$

$$s_{j,2} = GRU_2(ATT_1(C_1, s_{j,1}), s_{j,1})$$

$$s_{j,3} = GRU_3(ATT_2(C_2, s_{j,2}), s_{j,2})$$

$$s_{j,k} = GRU_k(0, s_{j,k-1}) \; for \; 3 < k \leq L_t$$

Extension of "Deep Architectures for Neural Machine Translation" (Miceli Barone et al. 2017), notation also follows this publication. Could be generalized to an arbitrary number of blocks with an attention mechanism.

2 For exactly a reccurrence transition depth of 3

Following the Nematus system description from EACL 2017. Overall decoder state s_j:

$$s_j = cGRU_{new} \; (s_{j-1}, y_{j-1}, C_m, C_c)$$

where C_m are annotation vectors obtained from bidirectional encoding of the sentence that is currently being translated (the "main" sentence) and C_c are annotation vectors from a different encoder that encoded previous context (for instance, the previous sentence).

s_j would be computed as follows:

$$s_j = GRU_3 \; (s'_j, c_j) = (1 - z_j) \odot \underline{s}_j + z_j \odot s''_j$$

GRU_3 is an additional transition block to accomodate the additional attention mechanism. The context vector c_j is computed by an attention mechanism that takes as input an intermediate proposed state s''_j and a set of annotation vectors C_c:

$$c_j = ATT(C_c, s''_j)$$

s''_j in turn is generated by the second transition block GRU_2 that takes as input the proposed intermediate state s'_j and a context vector c'_j:

$$s''_j = GRU_2(s'_j, c'_j) = (1 - z''_j) \odot \underline{s}''_j + z''_j \odot s'_j$$

The main context vector c'_j is the output of the main attention mechanism:

$$c'_j = ATT(C_m, s'_j)$$

and finally, the first GRU transition block computes s'_j:

$$s'_j = GRU_1(y_{j-1}, s_{j-1}) = (1 - z'_j) \odot \underline{s}'_j + z'_j \odot s_j$$