Evolutionary Path from Bahdanau Attention to Transformers

Pedagogical Notes

These notes accompany the intermediate seq2seq notebooks and highlight the main teaching goals, equations, and diagram hints for each step in the progression.

Step 1 – Multi-Head Attention in a Bahdanau Decoder (`bahdanau-multihead-at`

- **Objective**: keep the GRU encoder/decoder loop intact while allowing the decoder to align through multiple subspaces.
- **Key equations** (head i):

$$
\begin{aligned}
Q_i &= H_{\text{dec}} W_q^{(i)}, \qquad K_i = H_{\text{enc}} W_k^{(i)}, \qquad V_i = H_{\text{enc}} W_v^{(i)}, \\
\text{head}_i &= \text{softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_h}}\right) V_i, \\
\text{MHA}(H_{\text{dec}}, H_{\text{enc}}) &= [\text{head}_1; \ldots; \text{head}_h] W_o.
\end{aligned}
$$

- **Diagram cue**: draw a Bahdanau decoder time step, split the query into heads, attend to encoder states in parallel, concatenate, feed into GRU.
- **Teaching tips**: emphasize reshaping ((batch, seq, hidden) $\rightarrow$ (batch, heads, seq, hidden/head)), mention dropout on weights, and contrast with scalar Bahdanau scores.

Step 2 – GRU Encoder with Added Self-Attention (`encoder-self-attention-hy`

- **Objective**: enhance encoder representations using Transformer blocks stacked on GRU outputs while keeping the decoder unchanged.
- **Block math**:

$$
\begin{aligned}
Z &= \text{LayerNorm}\big(H + \text{MHA}(H, H, H, \text{mask})\big), \\
H' &= \text{LayerNorm}\big(Z + \text{FFN}(Z)\big), \qquad \text{FFN}(x) = \sigma(xW_1 + b_1)W_2 + b_2.
\end{aligned}
$$

- **Diagram cue**: show GRU outputs feeding AddNorm+MHA, then AddNorm+FFN; resulting H' goes to the decoder.
- **Teaching tips**: explain residual/LayerNorm stabilization, valid length masking, and encourage comparing validation curves with/without attention.

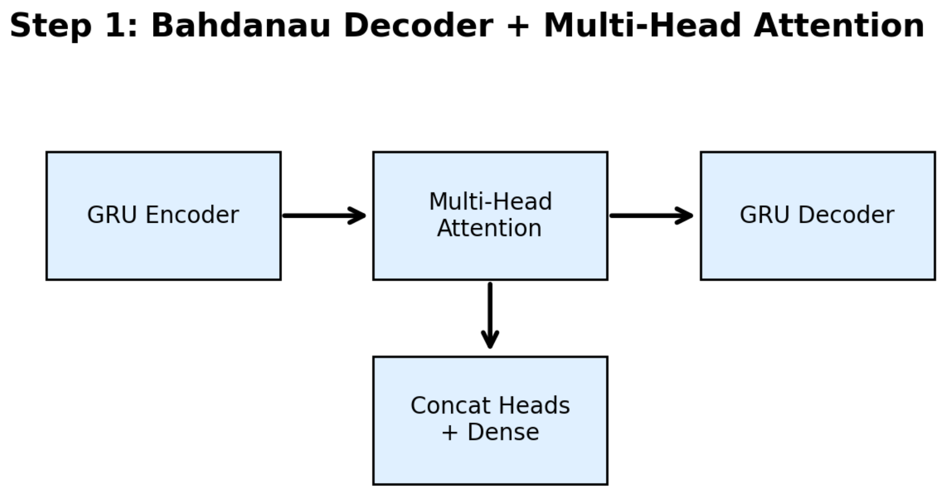

Figure 1: Step 1: GRU encoder-decoder with multi-head attention inserted between context and decoder input.

Step 3 – Decoder with Masked Self-Attention + GRU (`decoder-self-attention`

- **Objective**: let the decoder attend to its own history before cross-attending to the encoder while still leveraging a GRU state.
- **Math**:

$$\begin{aligned} Y &= \mathrm{LayerNorm}\big(X + \mathrm{MHA}(X, X, X, M_{\mathrm{causal}})\big), \\ Z &= \mathrm{LayerNorm}\big(Y + \mathrm{MHA}(Y, H_{\mathrm{enc}}, H_{\mathrm{enc}})\big), \\ \mathrm{logits} &= \mathrm{GRU}(Z) \to \mathrm{Dense}. \end{aligned}$$

- **Diagram cue**: masked self-attention block, then encoder-decoder attention block, then GRU unrolled over timesteps.
- **Teaching tips**: write the causal mask matrix, discuss why masking is required even with teacher forcing, and contrast GRU state vs. self-attention context.

Step 4 – Transformer Decoder on GRU Encoder (`transformer-decoder-on-gru-`

- **Objective**: replace recurrent decoding with stacked Transformer decoder blocks while reusing the GRU encoder.
- **Block math**:

$$\begin{aligned} Y_1 &= \mathrm{AddNorm}(X, \mathrm{MHA}_{\mathrm{mask}}(X, X, X)), \\ Y_2 &= \mathrm{AddNorm}(Y_1, \mathrm{MHA}(Y_1, H_{\mathrm{enc}}, H_{\mathrm{enc}})), \\ Z &= \mathrm{AddNorm}(Y_2, \mathrm{FFN}(Y_2)). \end{aligned}$$

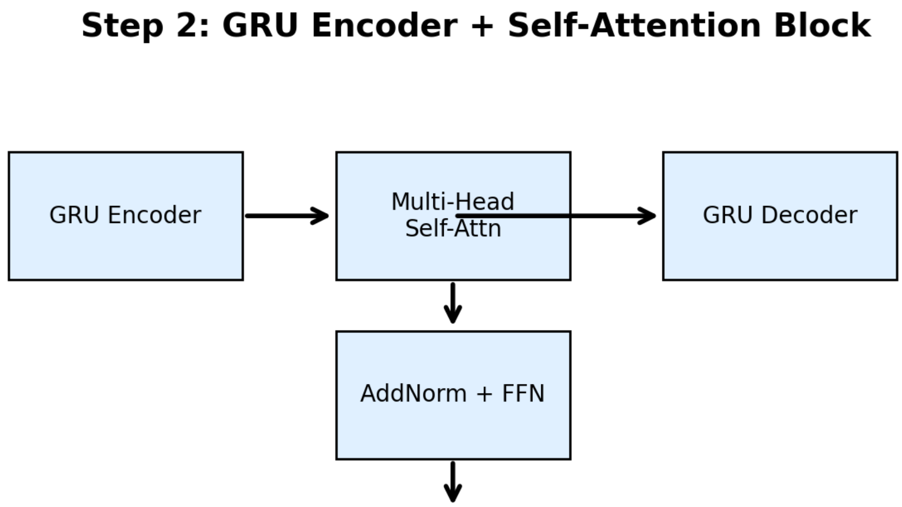

Figure 2: Step 2: GRU encoder outputs flow through a self-attention + FFN stack before reaching the baseline decoder.

- **Positional encoding**: $\mathrm{PE}_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d}}\right)$, $\mathrm{PE}_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d}}\right)$.
- **Diagram cue**: GRU encoder on the left, Transformer decoder stack on the right, positional encodings added before attention blocks.
- **Teaching tips**: highlight parallel decoding during training, caching of key/value pairs for inference, and ‘clone-state‘ for beam search.

Step 5 – Full Transformer (`transformer.ipynb`)

- **Objective**: fully parallel self-attention on both encoder and decoder with positional encodings everywhere.
- **Architecture**: repeat $[\mathrm{MHA}+\mathrm{AddNorm}, \mathrm{FFN}+\mathrm{AddNorm}]$ N times on the encoder; decoder uses masked self-attention followed by cross-attention and FFN blocks.
- **Regularization suggestions**: label smoothing, Adam with weight decay, higher dropout in attention and FFN sublayers.
- **Diagram cue**: canonical Transformer figure showing stacked encoder/decoder blocks with cross-attention connections.
- **Teaching tips**: compare to convolution/RNN receptive fields, note masking only in decoder self-attention, discuss training vs. inference behavior.

Suggested Pedagogical Flow

1. Revisit classic Bahdanau attention; motivate multi-head alignments.

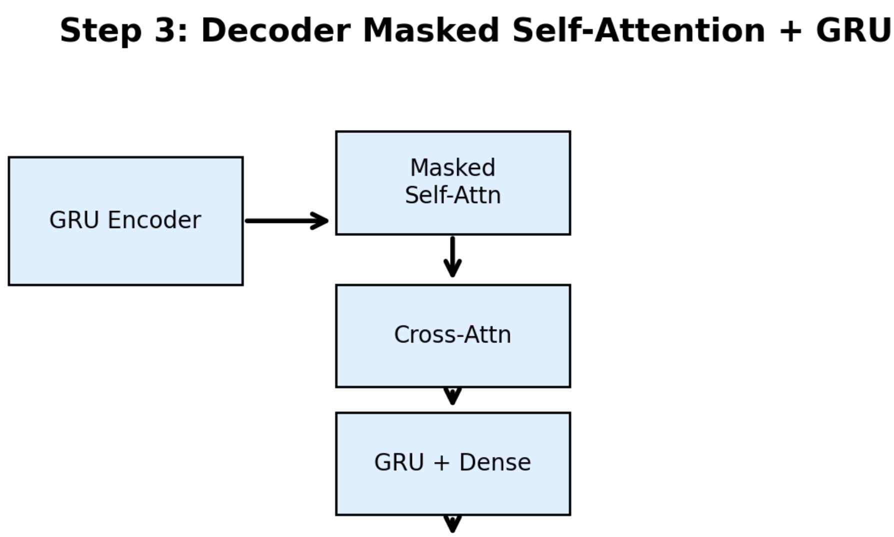

Figure 3: Step 3: Decoder pipeline showing masked self-attention, cross-attention, then GRU processing.

2. Introduce residual/LayerNorm intuition before stacked blocks.
3. Demonstrate causal masks with a toy 4×4 matrix.
4. Explain state caching and cloning for beam search in Transformer decoders.
5. Discuss why label smoothing and weight decay help once the model is fully attention-based.

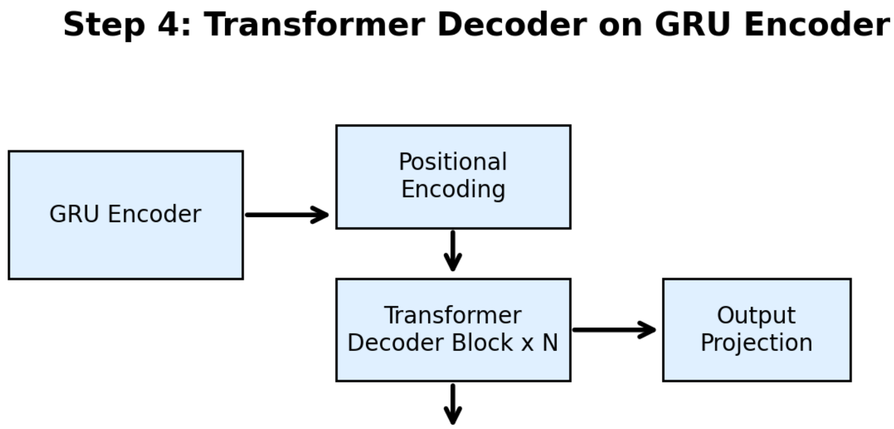

Figure 4: Step 4: GRU encoder provides memory for stacked Transformer decoder blocks with positional encodings.

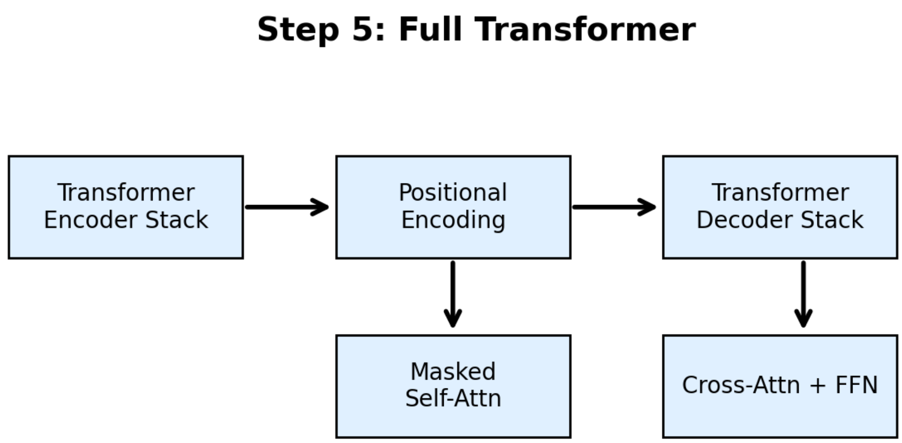

Figure 5: Step 5: Full Transformer with stacked encoder and decoder blocks connected via cross-attention.