Overview

Some recent sequence-to-sequence models like the Transformer (Vaswani et al., 2017) can score all output positions in parallel. We propose a simple algorithmic technique that exploits this property to generate multiple tokens in parallel at decoding time with little to no loss in quality. Our fastest models exhibit wall-clock speedups of up to 4x over standard greedy decoding on the tasks of machine translation and image super-resolution.

Implementation and Training

- Augment the decoder architecture to predict the next k tokens in parallel with sub-models $p_1, ..., p_k$
- Apply the original vocabulary projection
- Add k output layers
- Add a hidden layer
- Original decoder output

Basic Approach

Predict the next k tokens using the base scoring model and k - 1 auxiliary models; verify the predictions in parallel using the base model; accept the prefix that agrees with the greedy predictions.

Combined Approach

Combining the scoring and proposal models allows us to merge the current verify substep with the next predict substep, reducing the number of parallel model invocations during inference by a factor of 2.

Examples

**English-German machine translation** using a model trained with k = 10:

Input: The James Webb Space Telescope (JWST) will be launched into space on board an Ariane5 rocket by 2018 at the earliest.


- Step 1 10 tokens [Das_, James_, Webb_, Space_, Tele, sko, p_, (...)]
- Step 2 5 tokens [w, ST_, ...) wird, bin]