Non-recurrent Models for Sequence Processing
Recurrent Models

- Variable input and output length
- Structured output
- Memory

- Hard to train
- Cannot learn long-term dependencies
  - LSTMs work up to ~100 steps
Temporal Convolution

I really love this product

Positive?
Autoregressive Models

Use previous outputs as inputs to predict a sequence

\[ x \]
\[ y_1 \]
Network

\[ x, y_1 \]
\[ y_2 \]
Network

\[ x, y_1, y_2 \]
\[ y_3 \]
Network

\[ h_0 \]
\[ \hat{y}_1 \]
\[ \hat{y}_2 \]
\[ \hat{y}_3 \]

Linear
Sigmoid

Linear
Sigmoid

Linear
Sigmoid
Autoregressive Models

\[ x \rightarrow y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow y_4 \rightarrow y_5 \]
Autoregressive Models

```
\[ \begin{align*}
  y_1 &= \text{dilation=1} \\
  y_2 &= \text{dilation=2} \\
  y_3 &= \text{dilation=4} \\
  y_4 &= \text{dilation=1} \\
  y_5 &= \text{dilation=2} \\
  y_6 &= \text{dilation=4} \\
  y_7 &= \text{dilation=1} \\
  y_8 &= \text{dilation=2}
\end{align*} \]

“Causal Convolution”
```
How do we generate output?

\[ \hat{y} = \arg\max_y P(y_1, y_2, y_3, \ldots | x) \]

\[ P(y_1, y_2, y_3, \ldots | x) = P(y_1 | x) \cdot P(y_2 | x, y_1) \cdot P(y_3 | x, y_1, y_2) \cdot \ldots \]
Greedy Sampling

\[ \hat{y}_i = \text{argmax}_{y_i} P(y_i|x, y_1, \ldots, y_{i-1}) \]

\[ P(y_1^a, y_2^a|x) = 0.36 \]
\[ P(y_1^a, y_2^b|x) = 0.24 \]
\[ P(y_1^b, y_2^a|x) = 0.4 \]
\[ P(y_1^b, y_2^b|x) = 0.0 \]
Sequential Sampling

\[ \hat{y}_i \sim P(y_i|x, y_1, \ldots, y_{i-1}) \]

Unbiased

\[ \hat{y} \sim P(y_1, y_2, \ldots | x) \]

Sample inefficient

\[ \hat{y} = \arg\max_y P(y_1, y_2, y_3, \ldots | x) \]
Beam Search

\[(\text{Assume each } y_i \in Y)\]

\[S = \{ (x) \} \]

Repeat N times

For \((x, y_1, y_2, \ldots, y_i) \in S\)

For \(y_{i+1} \in Y\)

Compute \(P(y_{i+1} | x, y_1, y_2, \ldots, y_i)\)

Find the top k sequences and store them in \(S\)
WaveNet

- Generate sounds as raw waveforms
- Text-to-speech
- Needs to look far back in time
  - ~40k samples/sec to match human hearing
  - 8k for speech

WaveNet Block

- Dilated causal convolution
- Gated activation
WaveNet

- Efficient to train
  - Shifted labels
- State-of-the-art music and English speech
- Slow to generate

Block, dilation=1
Block, dilation=2
Block, dilation=4
Block, dilation=8
Block, dilation=16
Block, dilation=32
Block, dilation=64
Block, dilation=128
Block, dilation=256
Block, dilation=512

Receptive field: 1024 per stack
Parallel WaveNet

- Inverse Autoregressive Flow (IAF)
  - Input: Text + random noise $x, z_1, z_2, \ldots$
  - Output: All audio samples in parallel

$$P(y_i | x, z_1, z_2, \ldots, z_{i-1})$$

- Trained to mimic the original WaveNet
- Can produce 500k samples / sec, 10x faster than necessary for real-time

Modern Approach to NLP

Transformers