Muppet Models

Natural Language Processing

University of Maryland

Efficiency Optimization

Quantization

- Suppose you want to represent 0.0 to 4.0 with an 8-bit integer: what is S and Z?
- What is the mapping for π?
- What are the largest and smallest numbers that will have a different mapping than π?

Quantization

- Suppose you want to represent 0.0 to 4.0 with an 8-bit integer: what is S and Z?

$$S = \frac{4.0}{255} = 0.0156862745 \tag{1}$$

- What is the mapping for π?
- What are the largest and smallest numbers that will have a different mapping than π?

Quantization

- Suppose you want to represent 0.0 to 4.0 with an 8-bit integer: what is S and Z? **Because 0 is at the start of the range, -128 is our zero point.**
- What is the mapping for π?
- What are the largest and smallest numbers that will have a different mapping than π?

Quantization

- Suppose you want to represent 0.0 to 4.0 with an 8-bit integer: what is S and Z?
- What is the mapping for π?

$$x_q = \text{round}\left(\frac{\pi}{0.0156862745} - 128\right) \tag{1}$$

$$x_q = \text{round}(200.3 - 128) \tag{2}$$

$$x_q = 72 \tag{3}$$

$$\tag{4}$$

- What are the largest and smallest numbers that will have a different mapping than π?

Quantization

- Suppose you want to represent 0.0 to 4.0 with an 8-bit integer: what is S and Z?
- What is the mapping for π?
- What are the largest and smallest numbers that will have a different mapping than π? 3.1294 (71), 3.1455 (73)

Speculative Decoding

- Let's say that you have $q(x) = \frac{1}{V}$. Is this a good proxy distribution for speculative decoding with a modern Muppet Model?

Speculative Decoding

- Let's say that you have $q(x) = \frac{1}{V}$. Is this a good proxy distribution for speculative decoding with a modern Muppet Model?
  No, it's horrible. You'll just randomly pick words and completely ignore context.

Speculative Decoding

- Let's say that you have $q(x) = \frac{1}{V}$. Is this a good proxy distribution for speculative decoding with a modern Muppet Model?
- Let's say that $p(x) = \frac{1}{V_n}$ where V_n is the number of nouns (i.e., it's a uniform distribution over all nouns). Describe when you will accept from speculative decoding.

Speculative Decoding

- Let's say that you have $q(x) = \frac{1}{V}$. Is this a good proxy distribution for speculative decoding with a modern Muppet Model?
- Let's say that $p(x) = \frac{1}{V_n}$ where V_n is the number of nouns (i.e., it's a uniform distribution over all nouns). Describe when you will accept from speculative decoding.
  Accept when you get nouns

Speculative Decoding

- Let's say that you have $q(x) = \frac{1}{V}$. Is this a good proxy distribution for speculative decoding with a modern Muppet Model?
- Let's say that $p(x) = \frac{1}{V_n}$ where V_n is the number of nouns (i.e., it's a uniform distribution over all nouns). Describe when you will accept from speculative decoding.
- When you reject, what distribution do you sample from?

Speculative Decoding

- Let's say that you have $q(x) = \frac{1}{V}$. Is this a good proxy distribution for speculative decoding with a modern Muppet Model?
- Let's say that $p(x) = \frac{1}{V_n}$ where V_n is the number of nouns (i.e., it's a uniform distribution over all nouns). Describe when you will accept from speculative decoding.
- When you reject, what distribution do you sample from?

$$p'(x) \propto \max(0, p(x) - q(x)) \tag{1}$$

Speculative Decoding

- Let's say that you have $q(x) = \frac{1}{V}$. Is this a good proxy distribution for speculative decoding with a modern Muppet Model?
- Let's say that $p(x) = \frac{1}{V_n}$ where V_n is the number of nouns (i.e., it's a uniform distribution over all nouns). Describe when you will accept from speculative decoding.
- When you reject, what distribution do you sample from?

$$p'(x) \propto \max(0, p(x) - q(x)) \tag{1}$$

$$p'(x) = \begin{cases} 0 & \text{if } x \notin X_{\text{noun}} \\ \frac{V - V_N}{V_N(V - V_N)} & \text{if } x \in X_{\text{noun}} \end{cases} \tag{2}$$