Entropy and Information Theory

Definition (Extrinsically): LMs' embedded performance on other tasks.

Definition (Intrinsically): How accurately LMs predict language.

Information:

$$S(x) = \log_2 \frac{1}{p(x)} = -\log_2 p(x).$$

Entropy:

$$H(X) = \sum_x p(x) \log_2 \frac{1}{p(x)} = -\sum_x p(x) \log_2 p(x).$$

Entropy is a lower bound on the average number of bits necessary to encode X.

Per-Word Entropy Rate:

$$H_{\text{rate}}(X) = \lim_{N \to \infty} \frac{1}{N} H(X_1, \cdots, X_N) \leq \log_2 V.$$

Joint Entropy:

$$H(X, Y) = -\sum_x \sum_y p(x, y) \log_2 p(x, y) = H(X) + H(Y) - I(X; Y).$$

Conditional Entropy:

$$H(Y|X) = \sum_{x \in X} p(x) H(Y|X = x).$$

Mutual Information:

$$I(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = \sum_{x,y} p(x, y) \log_2 \frac{p(x, y)}{p(x)p(y)}.$$