Data and Parameter Scaling Laws for Neural Machine Translation

Mitchell Gordon
Johns Hopkins University
mitchg@jh.edu

Kevin Duh
Johns Hopkins University
kevinduh@cs.jhu.edu

Jared Kaplan
Johns Hopkins University
jaredk@jhu.edu

Abstract
We observe that the development cross-entropy loss of supervised neural machine translation models scales like a power law with the amount of training data and the number of non-embedding parameters in the model. We discuss some practical implications of these results, such as predicting BLEU achieved by large scale models and predicting the ROI of labeling data in low-resource language pairs.

1 Introduction
As training neural networks becomes an organizational and multi-million dollar venture (Brown et al., 2020), it is imperative to quantifiably predict the benefits of scaling up neural networks. In this paradigm, machine learning is an engineering effort, in which money can buy resources (data, compute) and the main concern is to predict return-on-investment (ROI) while avoiding bottlenecks.

Recent work has observed that the cross entropy loss of neural language models and other autoregressive generative models scales like a power law in the amount of training data, compute, and number of model parameters over several orders of magnitude (Hestness et al., 2019; Kaplan et al., 2020; Henighan et al., 2020). Similar intuitions exist in the realm of supervised MT: doubling the amount of parallel training data leads to roughly a fixed improvement in BLEU in both phrase-based statistical MT (Irvine and Callison-Burch, 2013; Turchi et al., 2008) and neural MT (Koehn and Knowles, 2017; Sennrich and Zhang, 2019).

In Section 2, we show that these MT intuitions can be quantified and explained via cross-entropy power law scaling; using a handful of experiments on small subsets of MT datasets, we precisely predict the performance of large systems trained on orders of magnitude more data. In Section 3, we demonstrate how these trends might be utilized to make ROI predictions when annotating more data for low-resource language pairs.

2 Machine Translation Scaling Laws
To investigate the predictability of MT system performance as parameters/data increase, we train many Transformers of various sizes (Table 2) on randomly selected subsets of data ($\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, ...$) for several standard MT datasets. The smallest data subsets contain $\sim0.1\%$ of the total data available.

We use three language pairs in our experiments:

![Figure 1: The development cross-entropy vs. the amount of data used to train each model from Section 2. Each point is colored by the number of non-embedding parameters in the model. The best fit of Equation 1 via least-squares regression for each language pair is shown as the dotted line.](image)
German-English (de-en), Russian-English (ru-en), and Chinese-English (zh-en). The data for each language pair is a concatenation of WMT 2017 data (which includes news commentary, parliamentary proceedings, and web-crawled data) and OpenSubtitles2018 (Lison and Tiedemann, 2016; Tiedemann, 2016). Datasets are tokenized using the Moses tokenizer, after which a 30k BPE vocabulary is constructed using the full dataset. For evaluation, we use newstest2016 concatenated with the last 2500 lines of OpenSubtitles2018.

Transformers are trained with early stopping and a learning rate of 0.0002 with a plateau-reduce schedule for a maximum of 350k updates. Other training details can be found in the code supplement. The resulting losses are plotted in Figure 1, with model sizes ranging from 393k-56M parameters and data sizes from 40k-50M lines of text.

### 2.1 Cross-entropy vs. Data/Parameters

Kaplan et al. (2020) provide an ansatz that predicts cross-entropy loss given the amount of training data and the size of the neural model:

$$L(N, D) = \left[ \left( \frac{N C}{N} \right)^{\frac{\alpha_N}{\alpha_D}} + \frac{D C}{D} \right]^{\alpha_D}$$

(1)

where $L$ is the per-token development cross-entropy loss (in nats), $N$ is the number of non-embedding parameters, $D$ is the amount of training data (in bytes), and $\alpha_N$, $\alpha_D$, $N_C$, and $D_C$ are constants determined by the particulars of the data distribution and training setup.4

Figure 1 shows that this equation is highly predictive of our results.5 The predictions are also fairly stable; Table 3 shows that the best-fit parameters of this equation stay similar even when restricting ourselves to using only 3-6% of the data. In Appendix A we perform a retrospective analysis of the results from Zhang and Duh (2020) to give some insight into how different hyper-parameter settings may influence scaling coefficients.

As either $N$ or $D$ approaches infinity, $L(N, D)$ simplifies to a “pure power law” in the other variable, which looks like a straight line on a log-log graph. For example, if we assume all models are large enough that data becomes the main performance bottleneck, then:

$$L(D) = \left( \frac{D C}{D} \right)^{\alpha_D}$$

(2)

We will use this assumption later when dealing with very low-resource language pairs.

### 2.2 BLEU vs. Cross-Entropy Loss

Predicting cross-entropy loss by itself does not tell us much about the quality of the translation system; we would really like to predict the achieved

...
Figure 2: (Left) BLEU exponentially decays as cross-entropy increases. Different language pairs may have different exponents and constants. (Right) A retrospective analysis of Zhang and Duh (2020) (Appendix A). Even the same development dataset with different BPE applied may have a different constant multiplier.

BLEU score, which is more interpretable to humans as a measure of adequacy, fidelity, and fluency (Papineni et al., 2002). Figure 2 shows that the relationship between BLEU and cross-entropy can vary between different language pairs and BPE settings. However, when these factors are fixed, BLEU seems to exponentially increase as cross-entropy decreases:

\[
\text{BLEU}(L) \approx Ce^{-kL} \tag{3}
\]

This relationship is fairly predictable for high BLEU values, but becomes noisier as BLEU drops below 15. Notably, changing the BPE encoding does not seem to affect \( k \), but does change the multiplying constant \( C \).

Why should this relationship be exponential? We might gain some insight by re-writing Equation 3 in terms of the per-token perplexity \((P)\):

\[
\text{BLEU}(P) \approx C \left( \frac{1}{P} \right)^k \tag{4}
\]

where \( (1/P) \) can intuitively be interpreted as the expected unigram precision of an autoregressively sampled translation with the same length as the reference sentence (Manning and Schutze, 1999). This is only intuition, however: in practice, we do not sample translations but decode using beam search, and BLEU combines multiple modified n-gram precisions besides unigram precision.\(^6\)

Using a much smaller BPE vocabulary of 2k tokens rectifies this plateau and returns to power law scaling, even with datasets <5 MB. We believe this is because the smaller vocabulary makes it difficult to exploit unigram statistics for rare words. While this is not conclusive evidence, we recommend that cross-entropies near or above unigram LM performance should not be relied upon to extrapolate performance. Dataset subsets which contain less than half of the BPE vocabulary should similarly be avoided.\(^8\)

---

\(^6\)We evaluate BLEU using multi-bleu.perl from the Moses toolkit. De-bpe-ing, de-tokenizing, and using Sacrebleu (Post, 2018) adds a small amount of noise but does not qualitatively change our results. See Appendix Figure 5.

\(^7\)The relationship between precision and perplexity for higher values of \( n \) is not clear. In general, expected bigram precision \( \neq (1/P)^2 \).

---

2.3 Preventing Breakdown At Smaller Dataset Sizes

Some extremely low-resource MT datasets (which we examine in Section 3) can have less than 5 MB of data (~40k sentence pairs). Figure 3 shows that when we extend our previous experiments to datasets smaller than this size, using 0.05% - 0.0125% of the data, the data scaling power law seems to break down, casting doubt on our ability to extrapolate extremely low-resource results to medium and high-resource data regimes.

However, the results are not simply noisy but predictably plateau to an apparent ceiling of 7.8 nats. For reference, a unigram language model trained on only the English part of the training data (with a 30k BPE vocab) achieves a per-token cross-entropy of ~7 nats. This leads us to suspect that models in this data regime are learning to rely on simple unigram statistics that do not change much as we decrease the data size.

In ~5 MB datasets, around half of the 30k BPE vocabulary is never seen during training. In contrast, ~10 MB datasets contain almost every vocab token. See Appendix Figure 6.
Predicting ROI of Annotating Low-Resource Language Pairs

If we assume that data is the main performance bottleneck (as it is in many low-resource language pairs), we can plug Equation 2 into Equation 3 to directly model the relationship between BLEU and data size:

\[
\text{BLEU}(D) = C \exp \frac{K}{D^{\alpha}}
\]  

(5)

where \( K = k(D_C)^{\alpha_D} \). This can be further combined with the hourly cost of fluent human translators to give us an approximate USD-to-BLEU trade-off when annotating more data for low-resource language pairs.

Figure 4 shows some example projections for Tagalog-English (tl-en) and Swahili-English (sw-en), with each dataset containing less than 50k sentence pairs (Zavorin et al., 2020). Under some assumptions about the costs of human translation\(^9\), we predict that spending \( \sim \$60k \) USD to acquire more tl-en/sw-en data (which would roughly double the size of the either dataset) would lead to an improvement of around 10-15 BLEU.

3.1 Limitations

There is a reasonable amount of noise in the cross-entropy/BLEU relationship at this scale (shown in Appendix Figure 8) which limits the precision and reliability of these predictions. In practice, we expect small amounts of data can be acquired in batches and predictions can be re-evaluated before deciding to continue. However, these predictions give a general sense of the cost of progress in low-resource machine translation. When engineering a real-world system, the simple option of acquiring more data and predictably improving performance should always be carefully weighed against more complicated and less predictable options.

That being said, predictably achieving a high BLEU score on a test dataset is not equivalent to “solving translation” for that language pair. Under-specification (D’Amour et al., 2020) still poses a challenge for effectively evaluating machine translation systems in real-world scenarios, especially in low-resource language pairs where evaluation data is usually from a narrow domain. More robust evaluation methods are needed, and it is not clear whether the output of these methods will be as predictable as cross-entropy loss or BLEU.

And finally, while our work demonstrates empirical power law scaling of NMT systems, it does not attempt to provide any causal explanation for these results. We also do not investigate the specific training factors that lead to a particular scaling exponent, but we expect this to be a fruitful research direction for future exploration.\(^11\)

4 Conclusion

We have shown that supervised neural machine translation performance with Transformers scales like a power law in non-embedding parameters and training data, aligning with similar observations in unsupervised auto-regressive modeling. We’ve also seen that as development cross-entropy decreases, BLEU exponentially increases. These two relationships can be combined to predict an effective USD-to-BLEU trade-off when annotating more data, even in low-resource regimes.

\(^9\)We assume translation costs around 0.10 USD per word, each word is composed of 5 characters on average, and each character requires around a byte of space.

\(^10\)We train a 12 layer model using a 2k BPE dataset subsets (100%, 90%, ..., 50%) with five different data shuffling seeds. We also increase the checkpoint frequency for earlier stopping.

\(^11\)For example, Kasai et al. (2020) have found that a deep encoder and a shallower decoder can be more efficient, which may lead to better parameter scaling.
References


Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and tv subtitles.


Utkarsh Sharma and Jared Kaplan. 2020. A neural scaling law from the dimension of the data manifold.


A Parameter Scaling in Japanese-English Translation

In this section, we provide a brief retrospective analysis of the results of Zhang and Duh (2020), in which many MT systems were trained to evaluate the efficacy of hyper-parameter optimization techniques. Specifically, we examine their results on the Japanese-English WMT 2019 Robustness task (Li et al., 2019). Figure 9 shows power-law scaling of the development cross-entropy loss with the number of non-embedding parameters. We see that changing the BPE encoding vocabulary size and the number of layers can affect the constant multiplier $N_C$, but does not seem to affect the exponent $\alpha_N$. Furthermore, multiple attention

In these experiments, only a single dataset size was used so we were unable to verify power-law data scaling.
Figure 9: Non-embedding parameters vs. development cross-entropy for the Japanese-English models described in Section A. Changing the number of layers or the BPE vocab size or the number of Transformer layers seems to impact the multiplying constant $N_C$, but does not seem to change $\alpha_N$ much.

Head settings (8, 16) were trained for each model size but they do not seem to impact scaling trends. We exclude some outliers with unexpectedly large losses at for larger model sizes. This only occurs for specific learning rates, so we believe those models failed to converge due to improper learning rate tuning.