Extrapolation in NLP

Jeff Mitchell, Pasquale Minervini, Pontus Stenetorp and **Sebastian Riedel**
University College London
Department of Computer Science
{j.mitchell, p.minervini, p.stenetorp, s.riedel}@cs.ucl.ac.uk

Abstract

We argue that extrapolation to examples outside the training space will often be easier for models that capture global structures, rather than just maximise their local fit to the training data. We show that this is true for two popular models: the Decomposable Attention Model and word2vec.

1 Introduction

In a controversial essay, Marcus (2018a) draws the distinction between two types of generalisation: *interpolation* and *extrapolation*; with the former being predictions made *between* the training data points, and the latter being generalisation *outside* this space. He goes on to claim that deep learning is only effective at interpolation, but that human like learning and behaviour requires extrapolation.

On Twitter, Thomas Diettrich rebutted this claim with the response that no methods extrapolate; that *what appears to be extrapolation from X to Y is interpolation in a representation that makes X and Y look the same.* [1]

It is certainly true that extrapolation is hard, but there appear to be clear real-world examples. For example, in 1705, using Newton's then new inverse square law of gravity, Halley predicted the return of a comet 75 years in the future. This prediction was not only possible for a new celestial object for which only a limited amount of data was available, but was also effective on an orbital period twice as long as any of those known to Newton. Pre-Newtonian models required a set of parameters (deferents, epicycles, equants, etc.) for each body and so would struggle to generalise from known objects to new ones. Newton's theory of gravity, in contrast, not only described celestial orbits but also predicted the motion of bodies thrown or dropped on Earth.

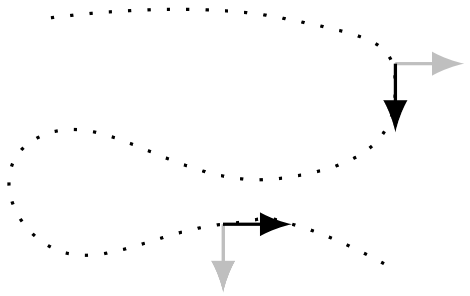

Figure 1: Generalising to unseen data: dotted line = training manifold; black arrows = interpolation; grey arrows = extrapolation. Both directions are represented globally in the training data, but local interpolation is only effective in one of them at each point.

In fact, most scientists would regard this sort of extrapolation to new phenomena as a vital test of any theory's legitimacy. Thus, the question of what is required for extrapolation is reasonably important for the development of NLP and deep learning.

Marcus (2018a) proposes an experiment, consisting of learning the identity function for binary numbers, where the training set contains only the even integers but at test time the model is required to generalise to even numbers. A standard multilayer perceptron (MLP) applied to this data fails to learn anything about the least significant bit in input and output, as it is constant throughout the training set, and therefore fails to generalise to the test set. Many readers of the article ridiculed the task and questioned its relevance. Here, we will argue that it is surprisingly easy to solve Marcus' even-odd task and that the problem it illustrates is actually endemic throughout machine learning.

Marcus (2018a) links his experiment to the systematic ways in which the meaning and use of a word in one context is related to its meaning and

[1] https://twitter.com/tdietterich/status/948811920001282049

Proceedings of the Workshop on Generalization in the Age of Deep Learning, pages 28–33
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

use in another (Fodor and Pylyshyn, 1988; Lake and Baroni, 2017). These regularities allow us to extrapolate from sometimes even a single use of a word to understand all of its other uses.

In fact, we can often use a symbol effectively with no prior data. For example, a language user that has never have encountered the symbol *Socrates* before may nonetheless be able to leverage their syntactic, semantic and inferential skills to conclude that *Socrates is mortal* contradicts *Socrates is not mortal*.

Marcus' experiment essentially requires extrapolating what has been learned about one set of symbols to a new symbol in a systematic way. However, this transfer is not facilitated by the techniques usually associated with improving generalisation, such as L2-regularisation (Tikhonov, 1963), drop-out (Srivastava et al., 2014) or preferring flatter optima (Hochreiter and Schmidhuber, 1995).

In the next section, we present four ways to solve this problem and discuss the role of global symmetry in effective extrapolation to the unseen digit. Following that we present practical examples of global structure in the representation of sentences and words. Global, in these examples, means a model form that introduces dependencies between distant regions of the input space.

2 Four Ways to Learn the Identity Function

The problem is described concretely by Marcus (1998), with inputs and outputs both consisting of five units representing the binary digits of the integers zero to thirty one. The training data consists of the binary digits of the even numbers $(0, 2, 4, 8, \ldots, 30)$ and the test set consists of the odd numbers $(1, 3, 5, 7, \ldots, 31)$. The task is to learn the identity function from the training data in a way that generalises to the test set.

The first model (SLP) we consider is a simple linear single layer perceptron from input to output.

In the second model (FLIP), we employ a change of representation. Although the inputs and outputs are given and fixed in terms of the binary digits **1** and **0**, we will treat these as symbols and exploit the freedom to encode these into numeric values in the most effective way for the task. Specifically, we will represent the digit **1** with the number `0` and the digit **0** with the number `1`. Again, the network will be a linear single layer perceptron without biases.

Model	Train	Test
SLP	8.12e-06	0.99
FLIP	6.79e-05	1.04e-05
ORTHO	1.27e-04	4.09e-05
CONV	1.71e-04	3.20e-05
PROJ	5.15e-06	8.07e-06

Table 1: Mean Squared Error on the Train (even numbers) and Test (odd numbers) Sets.

Returning to the original common-sense representation, $\mathbf{1} \rightarrow \mathtt{1}$ and $\mathbf{0} \rightarrow \mathtt{0}$, the third model (ORTHO) attempts to improve generalisation by imposing a global condition on the matrix of weights in the linear weights. In particular, we require that the matrix is orthogonal, and apply the absolute value function at the output to ensure the outputs are not negative.

For the fourth model (CONV), we use a linear Convolutional Neural Network (ConvNet, Lecun et al., 1998) with a filter of width five. In other words, the network weights define a single linear function that is shifted across the inputs for each output position.

Finally, in our fifth model (PROJ) we employ another change of representation, this time a dimensionality reduction technique. Specifically, we project the 5-dimensional binary digits $\mathbf{d}$ onto an n dimensional vector $\mathbf{r}$ and carry out the learning using an n-to-n layer in this smaller space.

$$\mathbf{r} = \mathbf{A}\mathbf{d} \tag{1}$$

where the entries of the matrix $\mathbf{A}$ are $A_{ij} = e^{\beta(j-i)}$. In each case, our loss and test evaluation is based on squared error between target and predicted outputs.

Training. Each model is implemented in TensorFlow (Abadi et al., 2015) and optimised for 1,000 epochs. In Eq. (1), we find that values of $\beta = \ln(2)$ and $n = 1$ work well in practice.

Results. As can be seen in Table 1, SLP fails to learn a function that generalises to the test set. In contrast, all the other models (FLIP, ORTHO, CONV, PROJ) generalise almost perfectly to the test set. Thus, we are left with four potential approaches to learning the identity function. Is lowest test set error the most appropriate means of choosing between them?

Discussion. This decision probably isn't as momentous as the choice discussed by Galileo in his Dialogue Concerning the Two Chief World Systems, where he presented the arguments for and against the heliocentric and geocentric models of planetary motion. These pre-Newtonian models could, in principle, attain as much predictive accuracy as desired, given enough data, by simply incorporating more epicycloids for each planet. On the other hand, they could not extrapolate beyond the bodies in that training data. Here, we will try to extract something useful from our results by considering how each model might generalise to other data and problems.

Although FLIP has the second lowest test set error, it is at best a cheap hack[2] which works only in the limited circumstance of this particular problem. If there were more than a single fixed digit in the training data, this trick would not work.

ORTHO suffers from the same problem, though it does embody the principle that everything in the input should end up in the output which seems to be part of this task.

CONV on the other hand will generalise to any size of input and output, and will even generalise to multiplication by powers of 2, rather than just learning the identity function.

PROJ, with the values $\beta = \ln(2)$ and $n = 1$, boils down to converting the binary digits into the equivalent single real value and learning the identity function via linear regression. This approach will extrapolate to values of any magnitude[3] and generalise to learning any linear function, rather than just the identity. As such, it is probably the only practically sensible solution, although it cheats by avoiding the central difficulty in the original problem.

At its most general, this central difficulty is the problem of extrapolating in a direction that is perpendicular to the training manifold. The even number inputs lay on a 4 dimensional subspace, while the odd numbers were displaced in a direction at right angles to that subspace. In this general form, the problem of how to respond to variation in the test set that is perpendicular to the training manifold lacks a well-defined unique solution, and this helps to explain why many people dismissed the task entirely.

However, this problem is in fact pervasive in most of machine learning. Training instances will typically lie on a low dimensional manifold and effective generalisation to new data sources will commonly require handling variation that is orthogonal to that manifold in an appropriate manner, e.g. Fig. 1. If prediction is based on local interpolation using a highly non-linear function, then no amount of smoothing of the fit will help.

Convolution is able to extrapolate from even to odd numbers because it exploits the key structure of the ordering of digits that a human would use. A human, given this task, would recognise the correspondence between input and output positions and then apply the same copying operation at each digit, which is essentially what convolution learns to do. It implicitly assumes that there is a global translational symmetry[4] across input positions, and this reduces the number of parameters and allows generalisation from one digit to another.

Returning to the linguistic question that inspired the task, we can think of systematicity in terms of symmetries that preserve the meaning of a word or sentence (Kiddon and Domingos, 2015). Ideally, our NLP models should embody or learn the symmetries that allow the same meaning to be expressed within multiple grammatical structures.

Unfortunately, syntax is complex and prohibits a short and clear investigation here. On the other hand, relations between sentences (e.g. contradiction) sometimes have much simpler symmetries. In the next section, we examine how global symmetries can be exploited in an inference task.

3 Global Symmetries in Natural Language Inference

The Stanford Natural Language Inference (SNLI, Bowman et al., 2015) dataset attempts to provide training and evaluation data for the task of categorising the logical relationship between a pair of sentences. Systems must identify whether each hypothesis stands in a relation of *entailment*, *contradiction* or *neutral* to its corresponding premise. A number of neural net architectures have been

[2] Nonetheless, such tricks are hardly unknown in machine learning research.

[3] Generalisation to values outside the training set would not be so successful had we used an MLP rather than a uniform linear function. Fitting to the training set using sigmoids will not yield a function that continues to approximate the identity very far beyond its range in the training set.

[4] Coincidentally, the rejection of the Earth centred model in favour of planetary motions orbiting the Sun played an important role in the recognition that the laws of physics also have a global translational symmetry, i.e. that no point in space is privileged or special.

proposed that effectively learn to make test set predictions based purely on patterns learned from the training data, without additional knowledge of the real world or of the logical structure of the task.

Here, we evaluate the Decomposable Attention Model (DAM, Parikh et al., 2016) in terms of its ability to extrapolate to novel instances, consisting of contradictions from the original test set which have been reversed. For a human that understands the task, such generalisation is obvious: knowing that A contradicts B is equivalent to knowing that B contradicts A. However, it is not at all clear that a model will learn this symmetry from the SNLI data, without it being imposed on the model in some way. Consequently we also evaluate a modification, S-DAM, where this constraint is enforced by design.

Models. Both models build representations, $\mathbf{v}_p$ and $\mathbf{v}_h$, of premise and hypothesis in attend and compare steps. The original DAM model then combines these representations by concatenating them and then transforming and aggregating the result to produce a final representation $\mathbf{u}_{ph}$, forming the input to a 3-way softmax:

$$\begin{aligned} \mathbf{u}_{ph} &= t(\mathbf{v}_p; \mathbf{v}_h), \\ p(i) &= s(\mathbf{u}_{ph} \cdot \mathbf{W}_i), \quad \text{with } i \in \{c, e, n\}. \end{aligned} \tag{2}$$

In S-DAM, we break the prediction into two decisions: contradiction vs. non-contradiction followed by entailment vs. neutral. The first decision is symmetrised by concatenating the vectors in both orders and then summing the output of the same transformation applied to both concatenations:

$$\begin{aligned} \tilde{\mathbf{u}}_{ph} &= t(\mathbf{v}_p; \mathbf{v}_h) + t(\mathbf{v}_h; \mathbf{v}_p), \\ p(j) &= s(\tilde{\mathbf{u}}_{ph} \cdot \tilde{\mathbf{W}}_j), \quad \text{with } j \in \{c, \neg c\}. \end{aligned} \tag{3}$$

Predictions for entailment and neutral are then made conditioned on $\neg c$:

$$\begin{aligned} \bar{\mathbf{u}}_{ph} &= t(\mathbf{v}_p; \mathbf{v}_h), \\ p(k|\neg c) &= s(\bar{\mathbf{u}}_{ph} \cdot \bar{\mathbf{W}}_k), \quad \text{with } k \in \{e, n\}. \end{aligned} \tag{4}$$

Results. Table 2 gives the accuracies for both models on the whole SNLI test set, the subset of contradictions, and the same set of contradictions reversed. In the last row, the DAM model suffers a significant fall in performance when the contradictions are reversed. In comparison, the S-DAM's performance is almost identical on both sets.

Instances	DAM	S-DAM
Whole Test Set	86.71%	85.95%
Contradictions	85.94%	85.69%
Reversed Contradictions	78.13%	85.20%

Table 2: Accuracy on all instances, contradictions and reversed contradictions from the SNLI test set.

Thus, the S-DAM model extrapolates more effectively because its architecture exploits a global symmetry of the relation between sentences in the task. In the following section, we investigate a global symmetry within the representation of words.

4 Global Structure in Word Embeddings

Word embeddings, such as GloVe (Pennington et al., 2014) and word2vec (Mikolov et al., 2013), have been enormously effective as input representations for downstream tasks such as question answering or natural language inference. One well known application is the $king = queen - woman + man$ example, which represents an impressive extrapolation from word co-occurrence statistics to linguistic analogies (Levy and Goldberg, 2014). To some extent, we can see this prediction as exploiting a global structure in which the differences between analogical pairs, such as $man - woman$, $king - queen$ and $father - mother$, are approximately equal.

Here, we consider how this global structure in the learned embeddings is related to a linearity in the training objective. In particular, linear functions have the property that $f(a + b) = f(a) + f(b)$, imposing a systematic relation between the predictions we make for a, b and $a + b$. In fact, we could think of this as a form of translational symmetry where adding a to the input has the same effect on the output throughout the space.

We hypothesise that breaking this linearity, and allowing a more local fit to the training data will undermine the global structure that the analogy predictions exploit.

Models. These embedding models typically rely on a simple dot product comparison of target and context vectors as the basis for predicting some measure of co-occurrence s:

$$s = f\left(\sum_i \text{target}_i \cdot \text{context}_i\right). \tag{5}$$

D	Linear	Non-Linear
100	50.38%	42.96%
200	53.18%	40.66%
400	50.77%	32.43%

Table 3: Accuracy on the analogy task.

We replace this simple linear function of the context vectors, with a set of non-linear broken-stick functions $g_i(\,\cdot\,)$.

$$s = f\left(\sum_i g_i\left(\text{context}_i\right)\right),$$

$$g_i\left(x\right) = \begin{cases} m_i x & \text{if } n_i x + c_i < 0, \\ \left(m_i + n_i\right) x + c_i & \text{otherwise.} \end{cases}$$

We modify the CBOW algorithm in the publicly available word2vec code to incorporate this non-linearity and train on the commonly used *text8* corpus of 17M words from Wikipedia. As this modification doubles the number of parameters used for each word, we test models of dimensions 100, 200 and 400.

Results. Table 3 reports the performance on the standard analogy task distributed with the word2vec code. The non-linear modification of CBOW is substantially less successful than the original linear version on this task. This is true on all the sizes of models we evaluated, indicating that this decrease is not simply a result of over-parameterisation.

Thus, destroying the global linearity in the embedding model undermines extrapolation to the analogy task.

5 Conclusions

Language is a very complex phenomenon, and many of its quirks and idioms need to be treated as local phenomena. However, we have also shown here examples in the representation of words and sentences where global structure supports extrapolation outside the training data.

One tool for thinking about this dichotomy is the *equivalent kernel* (Silverman, 1984), which measures the extent to which a given prediction is influenced by nearby training examples. Typically, models with highly local equivalent kernels - e.g. splines, sigmoids and random forests - are preferred over non-local models - e.g. polynomials - in the context of general curve fitting (Hastie et al., 2001).

However, these latter functions are also typically those used to express fundamental scientific laws - e.g. $E = mc^2$, $F = G\frac{m_1 m_2}{r^2}$ - which frequently support extrapolation outside the original data from which they were derived. Local models, by their very nature, are less suited to making predictions outside the training manifold, as the influence of those training instances attenuates quickly.

We suggest that NLP will benefit from incorporating more global structure into its models. Existing background knowledge is one possible source for such additional structure (Marcus, 2018b; Minervini et al., 2017). But it will also be necessary to uncover novel global relations, following the example of the other natural sciences.

We have used the development of the scientific understanding of planetary motion as a repeated example of the possibility of uncovering global structures that support extrapolation, throughout our discussion. Kepler and Newton found laws that went beyond simply maximising the fit to the known set of planetary bodies to describe regularities that held for every body, terrestrial and heavenly.

In our SNLI example, we showed that simply maximising the fit on the development and test sets does not yield a model that extrapolates to reversed contradictions. In the case of word2vec, we showed that performance on the analogy task was related to the linearity in the objective function.

More generally, we want to draw attention to the need for models in NLP that make meaningful predictions outside the space of the training data, and to argue that such extrapolation requires distinct modelling techniques from interpolation within the training space. Specifically, whereas the latter can often effectively rely on local smoothing between training instances, the former may require models that exploit global structures of the language phenomena.

Acknowledgments

The authors are immensely grateful to Ivan Sanchez Carmona for many fruitful disagreements. This work has been supported by the European Union H2020 project SUMMA (grant No. 688139), and by an Allen Distinguished Investigator Award.

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Jerry A. Fodor and Zenon W. Pylyshyn. 1988. Connectionism and cognitive architecture. *Cognition*, 28(1-2):3–71.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. 2001. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

Sepp Hochreiter and Jrgen Schmidhuber. 1995. Simplifying neural nets by discovering flat minima. In *Advances in Neural Information Processing Systems 7*, pages 529–536. MIT Press.

Chloé Kiddon and Pedro Rauel Cândido Domingos. 2015. Symmetry-based semantic parsing. Https://homes.cs.washington.edu/ pedrod/papers/sp14.pdf.

B. M. Lake and M. Baroni. 2017. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. *ArXiv e-prints*.

Yann Lecun, Lon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. In *Proceedings of the IEEE*, pages 2278–2324.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014*, pages 171–180.

G. Marcus. 2018a. Deep Learning: A Critical Appraisal. *ArXiv e-prints*.

G. Marcus. 2018b. Innateness, AlphaZero, and Artificial Intelligence. *ArXiv e-prints*.

Gary F. Marcus. 1998. Rethinking eliminative connectionism. *Cognitive Psychology*, 37:243–282.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA. Curran Associates Inc.

Pasquale Minervini, Thomas Demeester, Tim Rocktäschel, and Sebastian Riedel. 2017. Adversarial sets for regularising neural link predictors. In *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, UAI 2017, Sydney, Australia, August 11-15, 2017*. AUAI Press.

Ankur P. Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2249–2255.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *In EMNLP*.

B. W. Silverman. 1984. Spline smoothing: The equivalent variable kernel method. *Ann. Statist.*, 12(3):898–916.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

A. N. Tikhonov. 1963. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038.