A Probabilistic Interpretation of Transformers
International Conference on Machine Learning (ICML 2021)

Anonymous Authors[1]

Abstract

This document provides a basic paper template and submission guidelines. Abstracts must be a single paragraph, ideally between 4–6 sentences long. Gross violations will trigger corrections at the camera-ready phase.

1. Introduction

Transformers have reached state of the art results in language models, significantly outperforming LSTMs. One conceptual explanation for their increased performance is the ability of attention to utilize long range dependencies, whereas Recurrent Neural Networks were limited to encoding past information within a fixed-size hidden state. What this explanation does not explain is how certain architectural choices of transformers, specifically exponential dot product attention, also somewhat ambiguously referred to as softmax attention, outperforms alternatives.

Exponential dot product attention has been popularized in contrastive learning and metric learning. In language models, exponential dot product probability were used to model conditional probabilities in Word2Vec as well as in memory network. Contrastive loss originated as Noisy Contrastive Estimation and continues to be used in seminal papers such as SimCLR, which achieved state of the art results, as have many variants based off SimCLR.

The successes of transformers has been verified empirically, but little work has focused upon a theoretical framework for transformers. We offer a probabilistic explanation, based off of distributions of the exponential family, for attention and contrastive probabilities. Expressing attention as an exponential family allows us to utilize related theory in statistics, machine learning, and statistical mechanics, offering insightful interpretations of the transformer architecture.

We also explicitly state the limitations of our theory, noting that the modern Hopfield network interpretation shares many of these limitations. Moreover, for some of these limitations, we speculate connections between other areas of research which may reconcile the theoretical inconsistencies, motivating directions for future research.

2. Background

2.1. Exponential Dot Product Attention

Word2vec used a skip-gram model to predict neighboring words using a conditional distribution defined by a normalized exponential dot product multinoulli function (Mikolov et al., 2013)

Attention was proposed through normalized exponential alignment functions, often referred to as softmax attention in literature, for Neural Machine Translation (Graves, 2013; Bahdanau et al., 2014), and later work parallelizing computation on sequential data introduced normalized exponential dot product similarity (Parikh et al., 2016) (A Decomposable Attention Model for Natural Language Inference). Neural Turing Machines gated memory updates using normalized exponential cosine similarity, in what is referred to as soft attention (Graves et al., 2014).

Other transformer precursors parallelized attention updates over the entire sequence into a layer and switched to convolution-based attention weights (Kaiser & Sutskever, 2016; Kaiser & Bengio, 2016). The transformer architecture incorporated exponential dot product attention scaled by dimensionality (Vaswani et al., 2017).

2.2. Contrastive Learning and Metric learning

Noise-Contrastive Estimation (NCE) creates a mixture distribution between real data and noisy data to convert an unsupervised learning problem into a semi-supervised learning problem, modeling the distribution of the class conditioned on the data sample as a bernoulli distribution, and using the ratio of the model data distribution and noise distribution to define the logits (Gutmann & Hyvärinen, 2010).

In metric learning, Multidimensional scaling calculates pairwise distances between projected points (Cox & Cox, 2000),

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

and for Euclidean distances calculating pairwise distances is equivalent to calculating PCA's covariance matrix (Bishop, 2007), the individual terms of which are calculated by using dot products. Neighborhood Components Analysis learns a low dimensional linear embedding matrix and models a probability of a neighbor by comparing the exponential negative distance to a neighbor to the sum of the exponential negative distances to non-neighbors (Goldberger et al., 2004)

Due to slow convergence of Bernoulli contrastive loss and triplet loss, Sohn proposed an exponential dot product probability over multiple examples (Sohn, 2016), which is mathematically consistent with NCE for multiple distributions. The papers roots in metric learning motivated the dot product form, with a direct influence from Neighborhood Component Analysis.

More recent contrastive learning research adopted a contrastive loss based off of exponential dot product probabilities, including papers that achieve state of the art semi-supervised learning (Wu et al., 2018; Chen et al., 2020).

2.3. Shortcut Connections and Dynamical Systems

Long Short-Term Memory (LSTM) combined a shortcut connection to deal with the vanishing and exploding gradient problem along with gating functions to incorporate and forget information (Hochreiter & Schmidhuber, 1997). Residual connections similarly formulated the hidden layer as an update to an identity mapping, though without a gating mechanism (He et al., 2015). Recurrent Neural Networks have been interpreted as a discrete time approximation to a continuous dynamical system (Jaeger, 2001), where gating acts as a warping of time (Tallec & Ollivier, 2018). Residual connections have been interpreted as a discretized update to a differential equation (Weinan, 2017; Lu et al., 2020).

Interpreting residual networks as discretized differential equations, researchers have posed alternative methods for performing forward updates to converge to equilibrium points and backwards updates to the parameters from the equilibrium points (Chen et al., 2019; Bai et al., 2019). Further work has used monotone operator theory in convex analysis for solving for equilibrium points, interpreting layers as an operator (Winston & Kolter, 2021).

In a work most similar to ours, transformers have been interpreted as an update of modern hopfield networks and fixed points have been calculated with respect to a fixed set of patterns (Ramsauer et al., 2020). Our work similarly views the attention sublayer as an operator update over a class of discretized probability distribution, though with a changing set of patterns.

2.4. Log Normalizer and Free Energy

Partition functions, or the normalizer function, in statistical physics defines a normalization factor of the Hamiltonian with respect to a parameter defining the temperature. The Boltzmann distribution can be derived through Lagrange multipliers as the distribution which maximizes entropy with a conservation of energy constraint. Jaynes adapted the Boltzmann distributions to maximum entropy distributions with multiple expected statistics constraints by converting the maximum entropy problem into the dual problem of optimizing the log normalizer (Jaynes, 1982), which is known in statistical mechanics as free energy.

Variational methods have been used to approximate log probabilities of observations in machine learning, borrowing from ideas in statistical mechanics. By viewing the joint as an unnormalized probability distribution, the log normalizer is known as the evidence lower-bound, and it has connections to Helmholtz Free Energy (Hinton et al., 1995; Koller & Friedman, 2009).

The sum of exponents loss of AdaBoost (Collins et al., 2000) has been interpreted as the dual form of generalized KL divergence. The log sum of exponents is well known in convex optimization to be the dual form to the maximum entropy objective for a discrete probability distribution (Boyd & Vandenberghe, 2004). Notably, in the modern Hopfield network interpretation of transformers as part of the energy function (Ramsauer et al., 2020).

3. Exponential Dot Product Attention

3.1. Exponential Families

The natural parameter form, also known as the canonical form or a linear exponential family, of a distribution of the exponential family can be written as

$$p(x|\eta) = \frac{1}{Z(\eta)} h(x) e^{u(x)^\intercal \eta} \tag{1}$$

where x is the random variable, $u(x)$ the sufficient statistic, η the natural parameter, $h(x)$ the intrinsic measure or carrier measure, and $Z(\eta) := \int h(x) e^{u(x)^\intercal \eta}$ the normalizer or partition function. An exponential family distribution can be generated from an intrinsic measure $h(x)$ by defining the sufficient statistic $u(x) := x$, and defining the normalizer as the Laplace transform applied to the intrinsic measure:

$$Z(\eta) = \mathcal{L} \circ h(x) = \int h(x) e^{x^\intercal \eta} dx \tag{2}$$

The unnormalized probability mass or probability density may be written as $\tilde{p}(x|\eta)$.

When $h(x)$ is chosen to be a uniform discrete measure over a finite set of points $\{x_n\}_{n=1}^N$ in a continuous space, the probability density converts to a probability mass function, and the normalizer is the summation $Z(\eta) = \sum_{n=1}^{N} e^{u(x_n)^\intercal \eta}$ instead of an integral.

When the query is chosen as the natural parameter and the keys as the finite set of points, exponential dot product attention weights are of the form of an exponential family

$$p(k|q) = \frac{e^{k \cdot q}}{Z(q)} \tag{3}$$

The expected sufficient statistic of an exponential family can be written as a one-to-one function of the natural parameter

$$E_{x \sim P(x|\eta)}\left[u(x)\right] = \nabla_\eta \log Z(\eta) \tag{4}$$

For the exponential family defined by attention, we observe that attention averaging is exactly the gradient of the log normalizer

$$\nabla_\eta \log Z(\eta)|_{\eta=q} = \sum_{n=1}^{N} \frac{e^{k_n \cdot q}}{\sum_{n'=1}^{N} e^{k_{n'} \cdot q}} k_n \tag{5}$$

When the attention sublayer is added to the residual connection, we observe that we are performing a gradient update of the log normalizer with respect to the natural parameters, which are the hidden states.

The log normalizer for this discrete distribution is the log sum of exponents, which is a component of the modern Hopfield energy function, where the attention sublayer also acts as an update for the hidden states

$$\log Z(q) = \log \sum_{n=1}^{N} e^{k_n \cdot q} \tag{6}$$

3.2. Expansion and contraction

The attention sublayer outputs a convex combination of the keys. Without a residual connection, repeated applications of attention would contract the hidden states into the interior, intuitively towards a single fixed point.

With a residual connection, assuming our hidden states are recentered around the origin through layer normalization or some other normalization, we could intuitively imagine that roughly radially symmetric hidden states would push each hidden state further away from the origin. Since the log normalizer is the dual form of the maximum entropy distribution, our log normalizer ascent should result in increased entropy, resulting in an expansion of the hidden states away from their starting points.

If we are to hope that our transformer layers would converge to a certain distribution, we would require a compensating contractive operator. In transformers, layer normalization performs this contractive operation, and in contrastive learning, cosine similarity projects points back on to the unit sphere.

4. Electronic Submission

Submission to ICML 2021 will be entirely electronic, via a web site (not email). Information about the submission process and LaTeX templates are available on the conference web site at:

`http://icml.cc/`

The guidelines below will be enforced for initial submissions and camera-ready copies. Here is a brief summary:

- Submissions must be in PDF.
- Submitted papers can be up to eight pages long, not including references, plus unlimited space for references. Accepted papers can be up to nine pages long, not including references, to allow authors to address reviewer comments. Any paper exceeding this length will automatically be rejected.
- **Do not include author information or acknowledgements** in your initial submission.
- Your paper should be in **10 point Times font**.
- Make sure your PDF file only uses Type-1 fonts.
- Place figure captions *under* the figure (and omit titles from inside the graphic file itself). Place table captions *over* the table.
- References must include page numbers whenever possible and be as complete as possible. Place multiple citations in chronological order.
- Do not alter the style template; in particular, do not compress the paper format by reducing the vertical spaces.
- Keep your abstract brief and self-contained, one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase. The title should have content words capitalized.

4.1. Submitting Papers

Paper Deadline: The deadline for paper submission that is advertised on the conference website is strict. If your full, anonymized, submission does not reach us on time, it will not be considered for publication.

Anonymous Submission: ICML uses double-blind review: no identifying author information may appear on the title page or in the paper itself. Section 5.3 gives further details.

Simultaneous Submission: ICML will not accept any paper which, at the time of submission, is under review for another conference or has already been published. This policy also applies to papers that overlap substantially in technical content with conference papers under review or previously published. ICML submissions must not be submitted to other conferences and journals during ICML's review period. Informal publications, such as technical reports or papers in workshop proceedings which do not appear in print, do not fall under these restrictions.

Authors must provide their manuscripts in **PDF** format. Furthermore, please make sure that files contain only embedded Type-1 fonts (e.g., using the program `pdffonts` in linux or using File/DocumentProperties/Fonts in Acrobat). Other fonts (like Type-3) might come from graphics files imported into the document.

Authors using **Word** must convert their document to PDF. Most of the latest versions of Word have the facility to do this automatically. Submissions will not be accepted in Word format or any format other than PDF. Really. We're not joking. Don't send Word.

Those who use **LaTeX** should avoid including Type-3 fonts. Those using `latex` and `dvips` may need the following two commands:

```
dvips -Ppdf -tletter -G0 -o paper.ps paper.dvi
ps2pdf paper.ps
```

It is a zero following the "-G", which tells dvips to use the config.pdf file. Newer TeX distributions don't always need this option.

Using `pdflatex` rather than `latex`, often gives better results. This program avoids the Type-3 font problem, and supports more advanced features in the `microtype` package.

Graphics files should be a reasonable size, and included from an appropriate format. Use vector formats (.eps/.pdf) for plots, lossless bitmap formats (.png) for raster graphics with sharp lines, and jpeg for photo-like images.

The style file uses the `hyperref` package to make clickable links in documents. If this causes problems for you, add `nohyperref` as one of the options to the `icml2021` usepackage statement.

4.2. Submitting Final Camera-Ready Copy

The final versions of papers accepted for publication should follow the same format and naming convention as initial submissions, except that author information (names and affiliations) should be given. See Section 5.3.2 for formatting instructions.

The footnote, "Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute." must be modified to "*Proceedings of the 38^{th} International Conference on Machine Learning*, Online, PMLR 139, 2021. Copyright 2021 by the author(s)."

For those using the **LaTeX** style file, this change (and others) is handled automatically by simply changing `\usepackage{icml2021}` to

```
\usepackage[accepted]{icml2021}
```

Authors using **Word** must edit the footnote on the first page of the document themselves.

Camera-ready copies should have the title of the paper as running head on each page except the first one. The running title consists of a single line centered above a horizontal rule which is 1 point thick. The running head should be centered, bold and in 9 point type. The rule should be 10 points above the main text. For those using the **LaTeX** style file, the original title is automatically set as running head using the `fancyhdr` package which is included in the ICML 2021 style file package. In case that the original title exceeds the size restrictions, a shorter form can be supplied by using

```
\icmltitlerunning{...}
```

just before `\begin{document}`. Authors using **Word** must edit the header of the document themselves.

5. Format of the Paper

All submissions must follow the specified format.

5.1. Dimensions

The text of the paper should be formatted in two columns, with an overall width of 6.75 inches, height of 9.0 inches, and 0.25 inches between the columns. The left margin should be 0.75 inches and the top margin 1.0 inch (2.54 cm). The right and bottom margins will depend on whether you print on US letter or A4 paper, but all final versions must be produced for US letter size.

The paper body should be set in 10 point type with a vertical spacing of 11 points. Please use Times typeface throughout the text.

5.2. Title

The paper title should be set in 14 point bold type and centered between two horizontal rules that are 1 point thick, with 1.0 inch between the top rule and the top edge of the page. Capitalize the first letter of content words and put the rest of the title in lower case.

5.3. Author Information for Submission

ICML uses double-blind review, so author information must not appear. If you are using LaTeX and the `icml2021.sty` file, use `\icmlauthor{...}` to specify authors and `\icmlaffiliation{...}` to specify affiliations. (Read the TeX code used to produce this document for an example usage.) The author information will not be printed unless `accepted` is passed as an argument to the style file. Submissions that include the author information will not be reviewed.

5.3.1. Self-Citations

If you are citing published papers for which you are an author, refer to yourself in the third person. In particular, do not use phrases that reveal your identity (e.g., "in previous work (**?**), we have shown ...").

Do not anonymize citations in the reference section. The only exception are manuscripts that are not yet published (e.g., under submission). If you choose to refer to such unpublished manuscripts (**?**), anonymized copies have to be submitted as Supplementary Material via CMT. However, keep in mind that an ICML paper should be self contained and should contain sufficient detail for the reviewers to evaluate the work. In particular, reviewers are not required to look at the Supplementary Material when writing their review.

5.3.2. Camera-Ready Author Information

If a paper is accepted, a final camera-ready copy must be prepared. For camera-ready papers, author information should start 0.3 inches below the bottom rule surrounding the title. The authors' names should appear in 10 point bold type, in a row, separated by white space, and centered. Author names should not be broken across lines. Unbolded superscripted numbers, starting 1, should be used to refer to affiliations.

Affiliations should be numbered in the order of appearance. A single footnote block of text should be used to list all the affiliations. (Academic affiliations should list Department, University, City, State/Region, Country. Similarly for industrial affiliations.)

Each distinct affiliations should be listed once. If an author has multiple affiliations, multiple superscripts should be placed after the name, separated by thin spaces. If the authors would like to highlight equal contribution by multiple first authors, those authors should have an asterisk placed after their name in superscript, and the term "*Equal contribution" should be placed in the footnote block ahead of the list of affiliations. A list of corresponding authors and their emails (in the format Full Name <email@domain.com>) can follow the list of affiliations. Ideally only one or two names should be listed.

A sample file with author names is included in the ICML2021 style file package. Turn on the `[accepted]` option to the stylefile to see the names rendered. All of the guidelines above are implemented by the LaTeX style file.

5.4. Abstract

The paper abstract should begin in the left column, 0.4 inches below the final address. The heading 'Abstract' should be centered, bold, and in 11 point type. The abstract body should use 10 point type, with a vertical spacing of 11 points, and should be indented 0.25 inches more than normal on left-hand and right-hand margins. Insert 0.4 inches of blank space after the body. Keep your abstract brief and self-contained, limiting it to one paragraph and roughly 4–6 sentences. Gross violations will require correction at the camera-ready phase.

5.5. Partitioning the Text

You should organize your paper into sections and paragraphs to help readers place a structure on the material and understand its contributions.

5.5.1. Sections and Subsections

Section headings should be numbered, flush left, and set in 11 pt bold type with the content words capitalized. Leave 0.25 inches of space before the heading and 0.15 inches after the heading.

Similarly, subsection headings should be numbered, flush left, and set in 10 pt bold type with the content words capitalized. Leave 0.2 inches of space before the heading and 0.13 inches afterward.

Finally, subsubsection headings should be numbered, flush left, and set in 10 pt small caps with the content words capitalized. Leave 0.18 inches of space before the heading and 0.1 inches after the heading.

Please use no more than three levels of headings.

5.5.2. Paragraphs and Footnotes

Within each section or subsection, you should further partition the paper into paragraphs. Do not indent the first line of a given paragraph, but insert a blank line between succeeding ones.

You can use footnotes[1] to provide readers with additional information about a topic without interrupting the flow of the paper. Indicate footnotes with a number in the text where the point is most relevant. Place the footnote in 9 point type at the bottom of the column in which it appears. Precede the first footnote in a column with a horizontal rule of 0.8 inches.[2]

[1] Footnotes should be complete sentences.

[2] Multiple footnotes can appear in each column, in the same

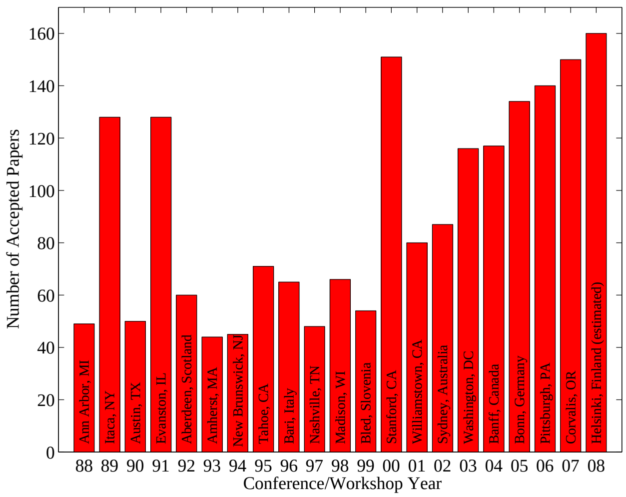

Figure 1. Historical locations and number of accepted papers for International Machine Learning Conferences (ICML 1993 – ICML 2008) and International Workshops on Machine Learning (ML 1988 – ML 1992). At the time this figure was produced, the number of accepted papers for ICML 2008 was unknown and instead estimated.

5.6. Figures

You may want to include figures in the paper to illustrate your approach and results. Such artwork should be centered, legible, and separated from the text. Lines should be dark and at least 0.5 points thick for purposes of reproduction, and text should not appear on a gray background.

Label all distinct components of each figure. If the figure takes the form of a graph, then give a name for each axis and include a legend that briefly describes each curve. Do not include a title inside the figure; instead, the caption should serve this function.

Number figures sequentially, placing the figure number and caption *after* the graphics, with at least 0.1 inches of space before the caption and 0.1 inches after it, as in Figure 1. The figure caption should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left. You may float figures to the top or bottom of a column, and you may set wide figures across both columns (use the environment `figure*` in LaTeX). Always place two-column figures at the top or bottom of the page.

5.7. Algorithms

If you are using LaTeX, please use the "algorithm" and "algorithmic" environments to format pseudocode. These require the corresponding stylefiles, algorithm.sty and algorithmic.sty, which are supplied with this package. Algorithm 1 shows an example.

---

order as they appear in the text, but spread them across columns and pages if possible.

Algorithm 1 Bubble Sort

Input: data x_i, size m
repeat
  Initialize $noChange = true$.
  for $i = 1$ **to** $m - 1$ **do**
    if $x_i > x_{i+1}$ **then**
      Swap x_i and x_{i+1}
      $noChange = false$
    end if
  end for
until $noChange$ is $true$

Table 1. Classification accuracies for naive Bayes and flexible Bayes on various data sets.

DATA SET	NAIVE	FLEXIBLE	BETTER?
BREAST	95.9$\pm$ 0.2	96.7$\pm$ 0.2	$\surd$
CLEVELAND	83.3$\pm$ 0.6	80.0$\pm$ 0.6	$\times$
GLASS2	61.9$\pm$ 1.4	83.8$\pm$ 0.7	$\surd$
CREDIT	74.8$\pm$ 0.5	78.3$\pm$ 0.6	
HORSE	73.3$\pm$ 0.9	69.7$\pm$ 1.0	$\times$
META	67.1$\pm$ 0.6	76.5$\pm$ 0.5	$\surd$
PIMA	75.1$\pm$ 0.6	73.9$\pm$ 0.5	
VEHICLE	44.9$\pm$ 0.6	61.5$\pm$ 0.4	$\surd$

5.8. Tables

You may also want to include tables that summarize material. Like figures, these should be centered, legible, and numbered consecutively. However, place the title *above* the table with at least 0.1 inches of space before the title and the same after it, as in Table 1. The table title should be set in 9 point type and centered unless it runs two or more lines, in which case it should be flush left.

Tables contain textual material, whereas figures contain graphical material. Specify the contents of each row and column in the table's topmost row. Again, you may float tables to a column's top or bottom, and set wide tables across both columns. Place two-column tables at the top or bottom of the page.

5.9. Citations and References

Please use APA reference format regardless of your formatter or word processor. If you rely on the LaTeX bibliographic facility, use `natbib.sty` and `icml2021.bst` included in the style-file package to obtain this format.

Citations within the text should include the authors' last names and year. If the authors' names are included in the sentence, place only the year in parentheses, for example when referencing Arthur Samuel's pioneering work (**?**). Otherwise place the entire reference in parentheses with the authors and year separated by a comma (**?**). List multiple references separated by semicolons (**???**). Use the 'et al.' construct only for citations with three or more authors or after listing all authors to a publication in an earlier reference (**?**).

Authors should cite their own work in the third person in the initial version of their paper submitted for blind review. Please refer to Section 5.3 for detailed instructions on how to cite your own papers.

Use an unnumbered first-level section heading for the references, and use a hanging indent style, with the first line of the reference flush against the left margin and subsequent lines indented by 10 points. The references at the end of this document give examples for journal articles (**?**), conference publications (**?**), book chapters (**?**), books (**?**), edited volumes (**?**), technical reports (**?**), and dissertations (**?**).

Alphabetize references by the surnames of the first authors, with single author entries preceding multiple author entries. Order references for the same authors by year of publication, with the earliest first. Make sure that each reference includes all relevant information (e.g., page numbers).

Please put some effort into making references complete, presentable, and consistent. If using bibtex, please protect capital letters of names and abbreviations in titles, for example, use {B}ayesian or {L}ipschitz in your .bib file.

Software and Data

If a paper is accepted, we strongly encourage the publication of software and data with the camera-ready version of the paper whenever appropriate. This can be done by including a URL in the camera-ready copy. However, **do not** include URLs that reveal your institution or identity in your submission for review. Instead, provide an anonymous URL or upload the material as "Supplementary Material" into the CMT reviewing system. Note that reviewers are not required to look at this material when writing their review.

Acknowledgements

Do not include acknowledgements in the initial version of the paper submitted for blind review.

If a paper is accepted, the final camera-ready version can (and probably should) include acknowledgements. In this case, please place such acknowledgements in an unnumbered section at the end of the paper. Typically, this will include thanks to reviewers who gave useful comments, to colleagues who contributed to the ideas, and to funding agencies and corporate sponsors that provided financial support.

References

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate, 2014.

Bai, S., Kolter, J. Z., and Koltun, V. Deep equilibrium models, 2019.

Bishop, C. M. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 1 edition, 2007. ISBN 0387310738. URL http://www.amazon.com/Pattern-Recognition-Learning-Information-Statisti dp/0387310738%3FSubscriptionId%3D13CT5CVB80YFWJEPWS02%26tag%3Dws%26linkCode%3Dxm2%26camp%3D2025%26creative%3D165953%26creativeASIN%3D0387310738.

Boyd, S. and Vandenberghe, L. *Convex Optimization*. Cambridge University Press, March 2004. ISBN 0521833787. URL http://www.amazon.com/exec/obidos/redirect?tag=citeulike-20&path=ASIN/0521833787.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. Neural ordinary differential equations, 2019.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. E. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020. URL https://arxiv.org/abs/2002.05709.

Collins, M., Schapire, R. E., and Singer, Y. Logistic regression, adaboost and bregman distances. In *Computational Learing Theory*, pp. 158–169, 2000. URL citeseer.nj.nec.com/article/collins00logistic.html.

Cox, T. F. and Cox, M. *Multidimensional Scaling, Second Edition*. Chapman and Hall/CRC, 2 edition, 2000. ISBN 1584880945. URL http://www.amazon.com/Multidimensional-Scaling-Second-Trevor-Cox/dp/1584880945.

Goldberger, J., Roweis, S. T., Hinton, G. E., and Salakhutdinov, R. Neighbourhood components analysis. In *NIPS*, pp. 513–520, 2004. URL http://dblp.uni-trier.de/db/conf/nips/nips2004.html#GoldbergerRHS04.

Graves, A. Generating sequences with recurrent neural networks. *CoRR*, abs/1308.0850, 2013. URL http://arxiv.org/abs/1308.0850.

Graves, A., Wayne, G., and Danihelka, I. Neural turing machines. *CoRR*, abs/1410.5401, 2014. URL http://arxiv.org/abs/1410.5401.

Gutmann, M. and Hyvärinen, A. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Teh, Y. and Titterington, M. (eds.), *Proc. Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, volume 9 of *JMLR W&CP*, pp. 297–304, 2010.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition, 2015.

Hinton, G. E., Dayan, P., Frey, B. J., and Neal, R. M. The wake-sleep algorithm for unsupervised neural networks. *Science*, 268:1158–1161, 1995.

Hochreiter, S. and Schmidhuber, J. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Jaeger, H. The echo state approach to analysing and training recurrent neural networks. *GMD-Report 148, German National Research Institute for Computer Science*, 01 2001.

Jaynes, E. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70(9):939–952, 1982. doi: 10.1109/PROC.1982.12425.

Kaiser, L. and Bengio, S. Can active memory replace attention? *CoRR*, abs/1610.08613, 2016. URL http://arxiv.org/abs/1610.08613.

Kaiser, L. and Sutskever, I. Neural gpus learn algorithms, 2016.

Koller, D. and Friedman, N. *Probabilistic Graphical Models: Principles and Techniques*. Adaptive computation and machine learning. MIT Press, 2009. ISBN 9780262013192. URL https://books.google.co.in/books?id=7dzpHCHzNQ4C.

Lu, Y., Zhong, A., Li, Q., and Dong, B. Beyond finite layer neural networks: Bridging deep architectures and numerical differential equations, 2020.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. Distributed representations of words and phrases and their compositionality. *CoRR*, abs/1310.4546, 2013. URL http://arxiv.org/abs/1310.4546.

Parikh, A. P., Täckström, O., Das, D., and Uszkoreit, J. A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933, 2016. URL http://arxiv.org/abs/1606.01933.

Ramsauer, H., Schäfl, B., Lehner, J., Seidl, P., Widrich, M., Gruber, L., Holzleitner, M., Pavlovic, M., Sandve, G. K., Greiff, V., Kreil, D. P., Kopp, M., Klambauer, G., Brandstetter, J., and Hochreiter, S. Hopfield networks is all you need. *CoRR*, abs/2008.02217, 2020. URL https://arxiv.org/abs/2008.02217.

Sohn, K. Improved deep metric learning with multi-class n-pair loss objective. In *NIPS*, 2016.

Tallec, C. and Ollivier, Y. Can recurrent neural networks warp time?, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*, pp. 5998–6008, 2017.

Weinan, E. A proposal on machine learning via dynamical systems. 2017.

Winston, E. and Kolter, J. Z. Monotone operator equilibrium networks, 2021.

Wu, Z., Xiong, Y., Yu, S., and Lin, D. Unsupervised feature learning via non-parametric instance-level discrimination, 2018.

A. Do *not* have an appendix here

Do not put content after the references. Put anything that you might normally include after the references in a separate supplementary file.

We recommend that you build supplementary material in a separate document. If you must create one PDF and cut it up, please be careful to use a tool that doesn't alter the margins, and that doesn't aggressively rewrite the PDF file. pdftk usually works fine.

Please do not use Apple's preview to cut off supplementary material. In previous years it has altered margins, and created headaches at the camera-ready stage.