Emergent Equivariance in Deep Ensembles

Jan E. Gerken







in collaboration with



Pan Kessel



Philipp Misof













Equivariance



T.S.COHEN@UVA.NL

M.WELLING@UVA.NL

Group Equivariant Convolutional Networks

Taco S. Cohen University of Amsterdam Max Welling University of Amsterdam

University of Amsterdam University of California Irvine Canadian Institute for Advanced Research

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symmeCorrolution layers can be used effectively in a deep network because all the layers in such a network net translative equivariaser: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature mars call leads un to edue-effects. In

T.S.COHEN@UVA.NL

M.WELLING@UVA.NL

Group Equivariant Convolutional Networks

Equivariant Transformer Networks

Kai Sheng Tai ¹	Peter Bailis ¹	Gregory Valiant

Abstract

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivariant Transformers (ETs), a famiily of differentiable image-to-image mappings that improve the robustness of models towards pre-defined continuous transformation groups. Through the use of specially-derived camerical coontinue wastern ETs incorporate functions that scaling to each training image). While data augmentation typically helps reduce the test error of CNN-based models, there is no guarantee that transformation invariance will be enforced for data not seen during training.

In contrast to training time approaches like data augmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016; Dieleman et al., 2016; Marcos et al., 2017; Worrall et al., 2017; Horniques & Vedaldi, 2017; Cohen et al., 2018) has explored new CNN architectures that are survented to the moreoid modified to noteriodite transforma-

Taco S. Cohen University of Amsterdam Max Welling

University of Amsterdam University of California Irvine Canadian Institute for Advanced Research

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symmeConvolution layers can be used effectively in a deep network because all the layers in such a network are translarice equivariant: shifting the image and then feeding it through a number of layers in the same as feeding the original image through the same layers and then shifting the useful fatture mark (a least not not desceffert).

3

Group Equivariant Convolutional Networks

T.S.COHEN@UVA.NL University of Amsterdam M.WELLING@UVA.NL University of Amsterdam

University of California Irvine Canadian Institute for Advanced Research

Taco S. Cohen

Max Welling

Abstract

We introduce Group emissiont Completional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces somely complexity by exploiting symmeConvolution layers can be used effectively in a deep network because all the layers in such a network are translation equivariant: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature many (at least up to obse-effects). In

Equivariant Transformer Networks

Kai Sheng Tai 1 Peter Bailis 1 Gregory Valiant

Abstract

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivariant Transformers (ETs), a family of differentiable image-to-image mappings that improve the robustness of models towards. pre-defined continuous transformation groups, Through the use of specially derived canonical cointrough the use of spectally-derived caronical co-onlinate systems. FTs incomorate functions that scaling to each training image). While data augmentation tynically helps reduce the test error of CNN-based models. there is no guarantee that transformation invariance will be enforced for data not seen during training.

In contrast to training time approaches like data sugmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016: Dieleman et al., 2016: Marcos et al., 2017: Worrall et al. 2017: Henriques & Vadable 2017: Cohenet al. 2018) has explored new CNN architectures that are

Theory for Equivariant Quantum Neural Networks

Ouvnh T. Nemen 3.2 Louis Schatzki 3.4 Paolo Braeria 3.5 Michael Barone 3.6 Patrick J. Coles,¹ Frédéric Sauvage,¹ Martin Larocca,^{1,7} and M. Cerezo³ ¹Theoretical Division, Los Alamos National Laboratory, Los Alamos, New Mexico 87515, USA ²School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138, USA. Information Sciences, Les Alemas National Laboratory, Les Alemas, New Merico 87515, USA University of Blinois at Urbana-Chempaign, Urbana, Elinois 61801, USA

^bDipartimento di Fisica e Astronomia, Università di Firenze, Sesto Fiorentino (FI), 50019, Italy ⁶Department of Mathematics, University of California Davis, Davis, California 95616, USA

Quantum neural network architectures that have little-to-no inductive biases are known to face trainability and generalization issues. Inspired by a similar problem, recent breakthroughs in machine learning address this challenge by creating models encoding the symmetries of the learning task. This is materialized through the usage of equivariant neural networks whose action com-

Group Equivariant Convolutional Networks

Tara S Calan T.S. COREN @ UNA.NL University of Annolam M. WELLING @ UNA.NL University of Collifornia Freier M. WELLING @ UNA.NL University of Collifornia Freier M. WELLING @ UNA.NL

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symmeCorrectution layers can be used effectively in a deep network because all the layers in such a network are translation equivariant: shifting the image and then feeding in through a transfer of layers is the same as feeding the original image through the same layers and then shifting the resulting feature mans (at least un to advece effect). In

Equivariant Transformer Networks

Kai Sheng Tai⁺ Peter Bailis⁺ Gregory Valiant⁺

Abstract

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivarian Transformers (IFUs), a famity of differentiable image-to-image mappings that improve the robustness of models towards pre-defined continuous transformation groups. Through the use of specially-derived carotical coonlinate washime. FC incorporate formition the other scaling to each training image). While data augmentation typically helps reduce the test error of CNN-based models, there is no guarantee that transformation imariance will be enforced for data not seen during training.

In contrast to training time approaches like data sugmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016; Dieleman et al., 2017; Hornov et al., 2017; Wernal et al., 2017; Henriques & Vedalds, 2017; Cohen et al., 2018) has explored new CNN architectures that are established to nurveal unell-tothic to satisfast transformasarranted to nurveal unell-tothic to satisfast transforma-

Theory for Equivariant Quantum Neural Networks

Quynh T. Negrus,^{1,2} Louis Schattal,^{1,4} Pinalo Buscela,^{1,3} Michael Ragnen,^{1,4} Patrick, J. Chel,^{1,4} Piofiti Schurger,¹ Muth. Lacces,¹ and M. Crewpl³ ¹Thematical Division. Les Alarone Neural Liberatory. Les Alarone, New Maxime 8744, USA ²Theorem and Applet Science, Revered University, Contrologi, Estavathanelli 1923, USA ³Departments Neural Control (1998), 2008. Control (1998), 2008. ⁴Departments Neural (1998), 2009. ⁴Department (1998), 2018. ⁴Departments (1998), 2019. ⁴Department (1998), 2019. ⁴Department (1998), 2019. ⁴Department of Microsof Microsoft (1998), 2019. ⁴Department (1998), ⁴Dep

^bDipartimoda i Puise Astronomia, Università di Primar, State Forentino (PJ), 60019, Raly Dipartimoda di Puise Astronomia, Università di Primar, State Forentino (PJ), 60019, Raly Department of Mathematica, Università qui Catifornia Davis, Davis, Caldornia 9606, USA Center for Noulicees Zhalica, Les Altamas Mathematica, Las Alaxans, New Mexico ST45, USA

Quantum neural network architectures that have little-to-no inductive biases are known to face trainability and generalization issues. Inspired by a similar problem, necessi trackthroughs in machino learning address this challenge by creating models encoding the symmetries of the learning task. This is materialized largely the usage of equivariant neural networks whose action comneurs with their of the symmetry. In this ways of the innerest them ishness to the maximum ends has

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging

Shiqi Gong^{a,c,1} Qi Meng^b Jue Zhang^b Huilin Qu^c Congqiao Li^d Sitian Qian^d Weitao Du^c Zhi-Ming Ma^a Tie-Yan Liu^b

*Androng of Mathematics and Systems Science, Chance Andreny of Sciences, Damogeneous Deat Road, Beijan (1019), China *Mercang Research Asia, Daming Street, Breijan (1008), China *CERN, RP Doputment, CH1111 Graves B., SosterAnda *School of Physics, Poling University, Chaneta Beaul, Brains (1008), China

Group Equivariant Convolutional Networks

Taro S. Calon T.S. CONTON (FUNCAS). University of Annotation MARTING University of Annotation MARTING (FUNCAS). University of Annotation (Funcas) Contact (Funcas).

Abstract

We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that reduces sample complexity by exploiting symmeConvolution layers can be used effectively in a deep network because all the layers in such a network are translation equivariant: shifting the image and then feeding in through a number of layers in the same as feeding the original image through the same layers and then shifting the provider fortune same (the low in the order-effects).

Equivariant Transformer Networks

Kai Sheng Tai⁺ Peter Bailis⁺ Gregory Valiant⁺

Abstract

How can prior knowledge on the transformation imrainaces of a domain be incorporated into the architecture of a neural network? We propose Equivatian Transformers (ETs), a family of differentiable image-to-image mappings that improve the robustness of models towards pre-defined continuous transformation groups. Through the use of specially-derived carotical coonlines weatmen. FTs incorporates functions that scaling to each training image). While data augmentation typically helps reduce the test error of CNN-based models, there is no guarantee that transformation imariance will be enforced for data not seen during training.

In contrast to training time approaches like data augmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016; Dietensnet 41a, 2017; Weerall et al., 2017; Henriques & Vedaldi, 2017; Cohen et al., 2013) has explored new CNN architectures that are surrestated to recorded meet CNN architectures that are

Theory for Equivariant Quantum Neural Networks

Qeynh T. Nguya, ^{1,6} Lothi Schatzli, ^{2,4} Panho Broscia, ^{1,5} Michael Bagour, ^{1,6} Patrick J. Coles,¹ Friddric Samagn,¹ Martin Larocea,^{1,7} and M. Coron³ ¹ Thermical Dirismi, Lot Masson Missinal Laboratory, Londons, Von Bernes 1971, USA ²School of Degenerating and Applied Sciences, Bierswei University, Cambridge, Massekoutte 0714, USA Papermains University, ¹ Control and Constraints, Cambridge, Massekoutte, 0713, USA ¹ Directed and Berneta and Computer Engineering.

¹ Opperature of Internation and Computer Engineering University of Illicois al Urbane-Champign, Urbane, Minosi 61801, USA ⁵ Dipartimento di Fusion e Astronomis, Università di Firmas, Scate Fiorentino (FI), 50019, Italy ⁶ Department of Mathematics, University of California Divisi, Davis, California 55164, USA ⁷ Center for Noulisseer Studies, Los Alamas Nalional Laboratory, Los Alamon, New Mexico 87255, USA

Quantum neural network architectures that have little-to-no inductive biases are known to face trainability and generalization issues. Inspired by a similar problem, recent trackthroughs in machino learning address this challenge by creating models encoding the symmetries of the learning task. This is materialized through the mags of equivariant neural networks whose action comneurs with their of the symmetry. In b this work more more three inducts to the surrounder neuro-

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging

Shiqi Gong^{a,c,1} Qi Meng^b Jue Zhang^b Huilin Qu^c Congqiao Li^d Sitian Qian^d Weitao Du^c Zhi-Ming Ma^a Tie-Yan Liu^b

*Androng of Mathematics and Systems Science, Chance Andreny of Sciences, Damogeneous Deat Road, Beijan (1019), China *Mercang Research Asia, Daming Street, Breijan (1008), China *CERN, RP Doputment, CH1111 Graves B., SosterAnda *School of Physics, Poling University, Chaneta Beaul, Brains (1008), China

E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials

Simon Batzner⁴,¹ Albert Musaelian,¹ Lixin Sun,¹ Mario Geiger,² Jonathan P. Mailon,³ Mordechai Kornbluth,³ Nicola Molinari,¹ Tess E. Smidt,^{4,5} and Boris Kozinsky^{4,1,3}

¹John A., Fundom Rohn of Digmorray and Applied Sciences, Branned University, Candrolph, M. 20155, USA Basel Polycologue Foldenk et Lansauxe, 2015 Lansauxe, Distributi Rohn Polycologue Foldenk et Lansauxe, 2015 Lansauxe, 2016 Control Computational Research Distance and Contexp For Manned Mathematics for Europy Research Applications, Lawrence Revisite National Laboratory, Berkels, CA 3720, USA Engineering and Computer Science, Carding Mathematics, Experiment and Computer Science, Carding Mathematics, 2018.

This work presents Neural Equivalent Interatoric Potentials (RequiP), an E(3)-equivalent energy large the large difference in the energy simulation for molecular dynamics inimitations. While most contemporary symmetry-source models use intraviate convolutions and only act on source and Neurope energy E(3)-equivalent action of generative transverse models are interaction of generative transver, resulting in a more information-rich and faithful representation of atomic environments. The method advises state-of-the-anti-context and a duality and the environments of the method and barries state-of-the-anti-context and a duality and difference state of the-anti-context and a duality of the method advised in the state of the environments.

Group Equivariant Convolutional Networks

Taco S. Cohen University of Amsterdam	T.S.COHEN@UVA.NL
Max Welling	M.WELLING@UVA.NL
University of Amsterdam	
University of California Irvine	
Canadian Institute for Advanced Research	

Abstract

We introduce Group emissiont Completional Neural Networks (G-CNNs), a natural generalization of convolutional neural networks that redaces sounds complexity by exploiting symmeConvolution layers can be used effectively in a deep network because all the layers in such a network are translation equivariant: shifting the image and then feeding it through a number of layers is the same as feeding the original image through the same layers and then shifting the resulting feature many (at least up to obse-effects). In

Equivariant Transformer Networks

Kai Sheng Tai 1 Peter Bailis 1 Gregory Valiant

Abstract

How can prior knowledge on the transformation invariances of a domain be incorporated into the architecture of a neural network? We propose Equivariant Transformers (ETs), a family of differentiable image-to-image mappings that improve the robustness of models towards. pre-defined continuous transformation groups, Through the use of specially-derived caponical coonlinate systems. FTs incomprate functions that scaling to each training image). While data augmentation tynically helps reduce the test error of CNN-based models. there is no guarantee that transformation invariance will be enforced for data not seen during training.

In contrast to training time approaches like data sugmentation, recent work on group equivariant CNNs (Cohen & Welling, 2016: Dieleman et al., 2016: Marcos et al., 2017: Worrall et al. 2017: Henriques & Vadaldi. 2017: Cohenet al. 2018) has explored new CNN architectures that are

Theory for Equivariant Quantum Neural Networks

Ouvnh T. Nemen 3.2 Louis Schatzki 3.4 Paolo Braeria 3.5 Michael Barone 3.6 Patrick J. Coles,¹ Frédéric Sauvage,¹ Martín Larocea,^{1,7} and M. Cerezo³ ¹Theoretical Division, Los Alamos National Laboratoru, Los Alamos, New Mexico 87515, USA ²School of Engineering and Applied Sciences. Harvard University, Cambridge, Mussachusetts 02138, USA. Information Sciences, Les Alemas National Laboratory, Les Alemas, New Merico 87515, USA ⁴Department of Electrical and Computer Engineering University of Blinois at Urbana-Chempaign, Urbana, Elinois 61801, USA ^bDipartimento di Fisica e Astronomia, Università di Firenze, Sesto Fiorentino (FI), 50019, Italy

⁶Department of Mathematics, University of California Davis, Davis, California 95616, USA

Quantum neural network architectures that have little to no inductive biases are known to face trainability and generalization issues. Inspired by a similar problem, recent breakthroughs in machine learning address this challenge by creating models encoding the symmetries of the learning task. This is materialized through the usage of equivariant neural networks whose action com-

An Efficient Lorentz Equivariant Graph Neural Network for Jet Tagging

Shini Gong^{4,c,1} Qi Meng⁶ Jue Zhang⁶ Huilin Qu^c Congniao Li^d Sitian Qian^d Weitao Du^o Zhi-Ming Ma^o Tio-Yan Liu^b

* Academy of Mathematics and Sostems Science. Chinese Academy of Sciences. Thomassensore Fost Road Reiling 100190 China ^bMicrosoft Research Asia. Danling Street, Briting 100089, China CERN, EP Department, CH-1211 Geneva 23. Switzerland distant of Dission Dahima University Chengla Road. Beijing 100871. China

E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials

Simon Batzner^{* 1} Albert Musaelian ¹ Livin Sun ¹ Mario Geiger ² Jonathan P. Mallon ³ Mordechai Kornbluth.³ Nicola Molinari.¹ Tess E. Smidt.^{4,5} and Boris Kozinsky^{*1,3}

John A. Postson School of Engineering and Ambiol Sciences. Harvard University, Cambridae, MA 02138, USA ³Robert Basch Research and Technology Center, Cambridge, MA 02139, USA 4 Computational Research Division and Center for Advanced Mathematics for Energy Research Applications. Engineering and Commuter Science, Cambridge, MA 02119, USA

This work networks Neural Equivariant Interatomic Potentials (NegaIP), an E(3)-equivariant dynamics simulations. While most contemporary symmetry-aware models use invariant convolutions and only act on scalars. NanuIP condens E(3) conjunctant consolutions for interactions of reconctrice tensors, resulting in a more information-rich and faithful representation of atomic environments. The method achieves state-of-the-art accuracy on a challenging and diverse set of molecules and

HIERARCHICAL, ROTATION-EQUIVARIANT NEURAL NETWORKS TO SELECT STRUCTURAL MODELS OF PROTEIN COMPLEXES

Stephan Eismann*	Raphael J.L. Townshend*	Nathaniel Thomas"
Department of Applied Physics	Department of Computer Science	Department of Physics
Stanfeed University	Stanford University	Stanford University
seismanz@stanford.edu	raphael@cs.stanford.edu	nthomas1030gmail.com
Millind Jagota	Bowen Jing	Ron O. Dror
epertment of Electrical Engineering	Department of Computer Science	Department of Computer Science
Stanford University	Stanford University	Stanford University
mjagota@atanford.edu	bjing@stanford.edu	rondror@cs.stanford.edu

ABSTRACT

Predicting the structure of multi-protein complexes is a grand challenge in biochemistry, with make implications for basic science and drug discovery. Computational structure prediction methods presently investigation of the structural features to distinguish accurate structural markels from less accurate ones. This raises the operation of whether it is rowaible to learn characteristics of accurate models directly from atomic coordinates of protein complexes, with no prior sequentions. Here we introduce a machine learning method that learns directly from the 3D positions of all atoms to

Group Equivariant (Convolutional Networks		Theory for Equivariant Quantum Neural Networks
		-	Quynh T. Nguyen, ^{1,2} Louis Schatzki, ^{3,4} Paolo Braccia, ^{1,3} Michael Ragone, ^{1,6}
Taco S. Cohen University of Amsterdam	T.S.COHEN@UVA.NL	Vol Gioro Tol I. Bates Balls I. Conserv. Vallant I	Patrick J. Coles, ¹ Frédéric Sauvage, ¹ Martin Larocca, ^{1,7} and M. Cerezo ³ ¹ Thorretical Division, Los Alaxos National Laboratory, Los Alaxos, New Maxio 87545, USA School of Evoluciering and Asplica Sciences. Hierard Visioeristic, Canaschastle, 60158, USA
Max Welling University of Amsterdam University of California Irvine Canadian Institute for Advanced Research	M.WELLING @UVA.	Geometric Deep Learning and Equivariant Neural Networks	¹ Information Sciences, Les Alamos National Laboratory, Les Alamos, New Mesico 8744, USA Department of Behrierica and Compared Fusioness, 1998, USA University of Biosis at Orbano-Champion, Urbano, Biomise (1894), USA ¹ Dapartmento of Jantamanico, Usanerati de primes, Sarde Faveriani (UT, 2007), Ibaly ² Dispertinento of Jantamanico, Usanerati de primes, Janaka, New Academia 9644, USA ² Ornie for Woltano Subato, Les Alamos Maina Laboratory, Las Manos, New Beisen 8744, USA
Abstract We introduce Group equivariant Convolutional Neural Networks (G-CNNs), a natural general- ization of correlational neural networks that re- duces sample complexity by exploiting symme-	Correducion layers can be used effectively in a deep work because all the layers is such a network <i>new</i> <i>natives</i> equivariants: athilting the image and then fore it through a <i>stamber</i> of layers is the same as feeding original image through the same layers and then all the resultine <i>feature</i> mass (at least un to obte-effects)	JAN E. GERKEN ¹ , JIMMY ARONSSON ¹⁸ , OSCAR CARLSSON ¹⁴ , HAMPUS LIMANDER ² , FREDRIK OHLSSON ³ , CHRISTOFFER PETERSSON ¹⁴ AND DANIEL PERSSON ¹ ¹ Chalmers University of Technology, Department of Mathematical Sciences 85:21266 Generalized, Sweden	Quantum moral serveds sufficient was than lume little isons inductive biases are known to five translativity and productions insets. Inclusively by a smaller profiles more translationships in more than the state of the task. This is materialised through the usage of explorational results where is the state of the state of the state with the ord the constants. It is the more than there is the state of the state of the state of the state of the state of the state of the state of the state of the
		² Gothenburg University, Department of Physics SE-41296, Gothenburg, Sweden ³ Umeà University, Department of Mathematics and Mathematical Statistics	ERARCHICAL, ROTATION-FOULVARIANT NEURAL NETWORKS
An Efficient Lorentz Equ	ivariant Graph Neural	SE-901 87, Umeå, Sweden	O SELECT STRUCTURAL MODELS OF PROTEIN COMPLEXES
Network for Jet Tagging	illin Qu ^c Congqiao Li ^d Sitian Qian ^d Weitao	⁴ Zenseacet SE-41736, Gothenburg, Sweden * equal contribution	Signate Economic Spannese of Applied Physics Dependent of Computer Science Standed University Standed University Standed University Standed University and Standed University Standed Un
DU" Zhi-Ming Ma ⁵ Tie-Yan Liu ⁴ * Anateng yi Mathematica and Kapitran Scien Zhanganasean East Book, Briging 100180; C * Meroscaft Remark Asia, CERN, EP Department, CHI-121 Chennes 23, Statisticand * GRN, EP Department, CHI-121 Chennes 23, Statisticand * School of Physics, Poking University, Chennefs Road, Berlins (2007); China	ce, Chinese Anderny of Sciences, Tana	International Accession of Astronom Spin Legislation in Astronomics Disputering and Computer Simon, Cambridge, J.A. (2012), USA This werk process Neural Department International Technological Computing and a stread approach for international protocols and a stread computing and and only of a stread of the strength international comparison of the stread and only of a stread of the strength Computing Simon Terms and the strength Simon Simon Simon Simon Simon Simon Simon Terms and Simon Simon Simon Simon Simon Simon Simon Simon Simon Terms and Simon Terms and Simon Terms Simon Simon Simon Simon Simon Simon Simon Simon Simon Simon Simon Simon Simon Simon Simon Simon Simon Simon S	Name Jages Bener Jage Bener Jage Summer of Chromosovies Digital Chromosovies Digital Chromosovies Digital Chromosovies Summer of Chromosovies Digital Chromosovies Digital Chromosovies Digital Chromosovies Digital Chromosovies Summer of Chromosovies Digital Chromosovies Digital Chromosovies Digital Chromosovies Digital Chromosovies Network Digital Chromosovies Digital





of complexes including proteins, nucleic acids, small molecules, ions and modified molecules. The new AlphaExid model demonstrates substantially immoved accuracy.

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

htps://doi.org/103038/s41586-024-07483 w	Josh Abramson ¹⁷ , Jonas Adlar ¹⁷ , Jack Dunger ¹² , Richard Evans ¹⁷ , Tim Green ¹⁷ ,	
Received 19 December 2023	Alexander Pritzel", Olaf Runneberger", Lindsoy Willreore", Andrew J. Ballard,	
Accepted: 29 April 2024	Michael O'Neill', David Reiman', Kathryn Turypeuwanekoo?, Zachary Wu', Akvilé Zengelyté	
Published online: 8 May 2024	Eirir Arvantif, Churles Bastini, Ottavia Bertolf, Alex Bridghard, Alexey Cherepanov, Miles Congreye, Alexander L. Oswar, Rivers', Andrew Cowiel, Michael Figurrov ¹ , Fabian B. Fuche', Hannah Gladman', Sishab Jain', Yasaf A. Khan', Casoline M. R. Lew',	
Open access		
Check for updates	Kala Perlin', Anas Poloperake', Pascel Servi, Suddolog Bingh', Advin Biecola', Advin Thelian Selected', Catholen Grog', Sergi Valoner', Blan Zhang'', Michal Zelinski', Azgustin Zidak', Victor Bapat'', Pashenest Kabil'', Hax Jakoberg'' ⁴⁴⁰ , Domin Hasabis' ¹⁴⁴⁰ & John M. Aregor ¹⁴⁴⁰	
	The introduction of AlphaFold 2 th has spurred a revolution in modeling the structure of proteins and their interactions, enabling a huge range of a polications in protein modeling and design ¹⁺ . Here we describe our AlphaFold 3 model with a substantial by undated diffusion haved architecture that is canable of producting the bias tracture structures.	

5

residues. The new Alpha Fold model demonstrates substantially informed accuracy

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

https://doi.org/103038/s41588-024-07483 w	w Josh Abramson ¹⁷ , Jonas Adler ¹⁷ , Jack Dunger ¹³ , Richard Evans ¹³ , Tim Green ¹⁷		
Received 19 December 2023	Alaxander Pritzel", Olaf Ronneberger", Lindsay Willmore", Andrew J. Ballard, Joshua Barabicki, Sabartian W. Bodesatelo, David J. Europi, Chin. Char. March.		
Accepted 29 April 2024 Published extine: 8 May 2024 Open access M. Ohiok for updates	Michael O'Hell, Casel Heiner, Astroya Turyasovanies, T. Zahary W., Janki Surgaryi Seria Asurit, Tonis Battis, O'Cutol Anton, T. Anton Solar, M. Kang Gargaryi Milas Gregoryi, Nasanife L, Gones Houri, Antone Gonei, Histohel Tgarano, Yadaha T, Hendi Sharano, Tabaka Miri, Yaand J, Narris, Garaina M, B. Levi, Katal Print, Jana Propertiel, Janes Bary, Sakhaba Bary, Anton Beeck, Milan Theori, Sana Propertiel, Janes Bary, Sakhaba Bary, Anton Beeck, Milan Tahan, Katan Katala, Sakhaba Bary, Sakhaba Bary, Anton Beeck, Milan Zhaoha, Yanan Katayan Taka, Yana Bary, Nariha Bary, Anton Beeck, Milan Zhaoha, Yanan Katayan Taka, Yana Bary, Nariha Bary, Mana Baron, Sakhaba Barta Barahaba Katayan Taka, Yana Bary, Barta Barta Barta, Barta Barta, Barta Hasabata ¹⁹⁹⁷ & Jaha K. Araper ⁴⁰¹		
			The introduction of Alphafold 2 ¹ has spurred a revolution in modelling the structure of proteins and their interactions, enabling a huge range of applications in protein modelling and design ¹⁻⁸ . Here we describe our Alphafold 3 model with a substantially updated diffusion haved architecure that is capable of predicting the joint structure updated.

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric Qu UC Berkeley ericqu@berkeley.edu Aditi S. Krishnapriyan UC Berkeley, LBNL aditik10berkeley.edu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields or machine learning. It incrives how a model's performance charges with increases in model size or input data, as well as how efficiently comparisonal resources are utilized to support this growth. Despite successes in scaling other types of machine learning models, the study of scaling in Neural Network Interasonic Peternials (NTNP) terminis limited. NINPs act as

residues. The new Alpha Fold model demonstrates substantially informed accuracy

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

	Htps://doi.org/103038/s41586-024-07483 w	Josh Abramson ¹⁷ , Jonas Adlar ¹⁹ , Jack Danger ¹³ , Richard Exam ¹⁷ , Tim Green ¹⁹ ,
	Received: 19 December 2023	Alexander Pritzel ¹⁴ , Olaf Ronneberger ¹⁴ , Lindsoy Willmore ¹⁴ , Andrew J. Ballard ² , Joshua Barabach ² Sabastian W Berfenstels ¹⁵ , David A. Europi ¹⁴ ("bio/"bio/"bio/"bio/"
	Accepted 29 April 2024	Michael O'Neill', David Reimar/, Kathryn Turyesuwunakool', Zachary Wu', Akulté Zereg
	Published online: 8 May 2024	Einini Arvanitr', Charles Beattie', Ottavia Bertolli', Alex Bridgland', Alexey Cherepanov', Miles Congreve', Alexander I, Cover, Elvera', Andrew Covia', Michael Fizarnov',
	Open access	Fabian B. Fuchs", Hannah Gladman", Fishub Jain", Yausuf A. Khan ¹² , Caroline M. R. Lew',
	Check for updates	Kala Perlini, Jona Pologorkai, Pascel Bory, Buhdeep Bingh, Adrian Borokai, Adhish Thilasokanen, Casheen Gorgi, Sargai Yakanen, Halan D. Dinagi, Michai Zelinaki, Jozgashi Zakiri, Victor Bopat ¹⁴ , Padreset Kahli ¹⁴ , Haz Jaketserg ¹⁴⁰ , Denis Hazakir ¹⁴⁰⁰ & John M. Arrgen ¹⁴⁰¹
		The introduction of AlphaFold 2 th has spurred a revolution in modeling the structure of proteins and their interactions, enabling a huge range of a protections in protein modeling and design ¹¹ . If there we derive our AlphaFold 31 model with a substantial by optato diffusion has derivative corrections (coupled of producting the joint structure of compression hult large models an unclearing the production model for a derivative structure and a substantial by the compression huber and architecture that is coupled of producting the joint structure of compression huber and architecture huber and the coupled of producting the joint structure.

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric Qu UC Berkeley ericqu@berkeley.edu Aditi S. Krishnapriyan UC Berkeley, LBNL aditik1@berkeley.edu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. In incrives how a model's performance charges with increases in model size or input data, as well as how efficiently comparational resources are utilized to support this growth. Despite successes in scaling other types of machine learning models, the study of scaling in Neural Network Interastorie Potentials (NTMP) series and in Neural Network Interastorie Potentials (NTMP) remains limited, NTMP set as Swallowing the Bitter Pill: Simplified Scalable Conformer Generation

Yuyang Wang⁺ Ahmed A. Elhag⁺² Navdeep Jaitly⁺ Joshua M. Susskind⁺ Miguel Ångel Bautista⁺

Abstract

We present a novel way to predict molecular conferences through a simple formulation that subscreps many of the heuristics of prior works and achieves state of the art results by using the abstrateges of scale. By mining additional generative model directly on DD atomic positions without miking assumptions about the exploit attracture of moleculas (e.g. medding protonal agels) or ex are table to subscribtly analyfily interis the used complexity of the 3D arraytne space, necespossing factors was do hand regime and noricinal angles. Despite the nucleotide graph distating potential, 2D conformers through specific constraints, such as boad types and uptial arrangements dotermined by chiral contex, the conformation of pace experisonce sequencial growth while the expansion of the graph size and the number of restable bands (Astello et Gornez-Ionsharzhi 2022). This complex cates brate losse and exbansive approaches, making them visually articasalite for even moleratory small molecules.

Systematic methods, like OMEGA (Hawkins et al., 2010)

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

Htps://doi.org/103038/s41586-024-07483 w	Josh Abramson ¹⁷ , Jonas Adlar ¹⁷ , Jack Dunger ¹⁰ , Richard Exans ¹⁷ , Tim Green ¹⁷ ,
Received: 19 December 2023	Alexander Pritzel", Olaf Ronneberger", Lindsoy Willenore", Andrew J. Ballard, Joshun Berehrich, Schusting W. Bodenstalo, David A. Europe, Chin. Chun Harry,
Accepted 29 April 2024	Michael O'Neill', David Reiman', Kathryn Turyaeuwanakoo?, Zachary Wu', Akolik Zeregelyti
Published online: 8 May 2024	Erkin Arvaniti', Charles Bastini', Ottavis Bertolf', Alos Bridgland', Alesey Cheropanov', Miles Congreve', Alexander L. Conver, Floves', Andrew Cowie', Michael Figurnov', Fabian B. Fucho', Hannah Gladman', Sishub Jain', Yousaf A. Khan'', Gaoline M. B. Lee',
Open access Fabian B. Check for updates Aske Peri Michail 24 Dentil Has	
	Kata Perlin', Ania Nobyekan', Taskat Bary', Buhtabep Bingh', Adrian Barada', Anhai Thibashalanan', Catherine Torg', Sengi Yahanan', Elan Di Zhang'', Michal Zialinah', Augustin Zidak', Victor Bapat'', Pashwaset Kabil'', Hax Jakoberg' ^{ani} , Denis Hasabia' ¹⁶⁰ & John H. Aurger ¹⁰¹
	The interduction of Alphafold 2 th has spurred a revolution in modelling the structure of protosts and their interactions, enabling a huge range of applications in proceen modelling and degrees. ¹ Inter our degree from a Alphafold 10 model with automating or complementation than approximate and an alphafold 10 model and automating applications and applications and applications and an alphafold 10 model and a structure applications and applications and applications and applications and an alphafold 10 model 10 mod

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric On UC Berkeley ericgu@berkeley.edu

Aditi S. Krishnaneiyan UC Berkeley, LBNL aditiki@harkalay adu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. It involves how a model's performance changes with increases in model size or input data, as well as how efficiently computational resources are utilized to support this growth. Despite successes in scaline other types of machine learning models, the study of scaline in Neural Network Interatomic Potentials (NNIPs) remains limited. NNIPs act as

Swallowing the Ritter Pill: Simplified Scalable Conformer Generation

Yuyang Wang⁺ Ahmed A, Elhag⁺² Navdeen Jaitly⁺ Joshua M, Susskind⁺ Micuel Ángel Bautista⁺

Abstract

We research a need way to medici melaratur conformers through a simple formulation that sidestens many of the hearistics of prior mode and achieves state of the art results by using the advantages of scale. By training a diffusion generative model directly on 3D atomic positions without making assumptions about the explicit structure of molecules (e.e. modeline torsional angles) we are able to radically simplify struc-

is the vast complexity of the 3D structure space, encomressing factors such as bond lengths and torsional angles. Despite the molecular graph dictating potential 3D conformers through specific constraints, such as bond types and spatial arrangements determined by chiral centers, the conformational space experiences exponential prowth with the expansion of the graph size and the number of rotatable bonds (Aselrod & Gomez-Bombarelli, 2022). This complicates brate force and exhaustive approaches, making them virtually unfeasible for even moderately small molecules.

Systematic methods, like OMEGA (Headcins et al., 2010)

Probing the effects of broken symmetries in machine learning

Marcel F Langer¹, Sergey N Pozdnyakov¹ and Michele Ceriotti¹

Laboratory of Conversational Science and Medeline and National Centre for Conversational Design and Discovery of Need Materials * Author to where any correspondence should be addressed.

E-mails michale contestilitenthick

Kerwords: machine learning, symmetry-constrained models, atomistic modeling, molecular simulations Samplementary material for this article is available online

Abstract

Symmetry is one of the most central concepts in physics, and it is no surprise that it has also been widely adopted as an inductive bias for machine-learning models applied to the physical sciences. This is especially true for models targeting the properties of matter at the atomic scale. Both established and state-of-the-art approaches, with almost no exceptions, are built to be exactly equivariant to translations, nermutations, and rotations of the atoms, incorporating symmetries-rotations in particular-constrains the model design space and implies more complicated architectures that are often also computationally demanding. There are indications

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

43ps://doi.org/103038/s41588-024-01483 w	Josh Abramson ¹⁷ , Jonas Adlar ¹⁷ , Jack Danger ¹⁰ , Richard Exans ¹⁷ , Tim Green ¹⁷ ,	
loceived. 19 December 2023	Alexander Pritzel ¹⁴ , Olaf Ronne berger ¹⁴ , Lindsoy Willmore ¹⁴ , Andress J. Ballard ² , Joshua Barabaluk ¹ , Sabastian W. Budeastak ¹⁴ , David A. Europi ¹⁴ ("bia: "basa")	
Accepted: 29 April 2024	Michael O'Neill', David Reiman', Kathryn Turypeuwunakoo?, Zachary Wu', Akulte Zeregel	
Addished online: 8 May 2024	Erich Arwardt, Churlas Bastill, Ottoria Barrill, Alles Bridgert, Alles Didgert, Marce Diarquert, Marce Corpora, Sansahar L. Oward, Brain, K. Marce M. Sansahar L. Oward, Brain, K. Marce M. Sansahar L. Oward, Brain, S. Marce M. Sansahar L. Oward, Brain, S. Marce M. Sansahar L. Sansahar S. Marce Sciences, Cartonis et al. (2014). Electric Marce	
Open access		
Check for updates		
	The introduction of AlphaFold 2 ⁴ has spurred a revolution in modelling the structu of proteins and their interactions, enabling a huge range of a polications in protein	

updated diffusion based architecture that is capable of predicting the joint structure of complexes including parteins, nucleic acids, small molecules, ions and modified residues. The new AlphaErdd model demonstrates substantially improved accuracy

Abuild Zerranduni

Probing the effects of broken symmetries in machine learning

Marcel F Langer . Servey N Pondovakov . and Michele Ceriotti .

Laboratory of Commutational Science and Medeline and National Centre for Commutational Dision and Discourse of Neural Materials * Author to where any correspondence should be addressed.

E-mails michale contestilitenthick

Kerwords: machine learning, symmetry-constrained models, atomistic modeling, molecular simulations Samplementary material for this article is available online

Abstract

Symmetry is one of the most central concepts in physics, and it is no surprise that it has also been widely adopted as an inductive bias for machine-learning models applied to the physical sciences. This is especially true for models targeting the properties of matter at the atomic scale. Both established and state-of-the-art approaches, with almost no exceptions, are built to be exactly equivariant to translations, permutations, and rotations of the atoms. Incorporating symmetries-rotations in particular-constrains the model design space and implies more complicated architectures that are often also computationally demanding. There are indications

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric On UC Berkeley ariconfbarkalay adu

Aditi S. Krishnaneiyan UC Berkeley, LBNL aditiki@harkalay adu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. It involves how a model's performance changes with increases in model size or input data, as well as how efficiently computational resources are utilized to support this growth. Despite successes in scaling other types of machine learning models, the study of scaling in Neural Network Interatomic Potentials (NNIPs) remains limited. NNIPs act as

Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics

Marloes Arts, *11,0 Victor Garcia Satorras, *10 Chin-Wei Huang, Daniel Zügner 1 Marco Federici 11 Cecilia Clementi 11 Frank Noé 1 Robert Pinsler # and Rianne van den Berg

1 Work done during an internship at Microsoft Research (Amsterdam). ¹University of Copenhaarn, Department of Computer Science, Universitetaparken 1. Conenhagen, 2100, Denmark

ALIScience, Microsoft Research, Evert van de Beekstraat 354, Amsterdam, 1118 CZ, The Netherlands

SALIScience Microsoft Research Karl-Lichbrecht-Straffe 39 Berlin 10178 Cermann University of Amsterdam, Informatics Institute, Science Park 901, Amsterdam, 1098 XII. The Metherlands

1 Preis Universität Rerlin, Department of Physics, Arnimalle 19, Berlin, 1/195, Germann, #AUScience, Microsoft Research, 21 Station Road, Cambridae, CB1 2FB, United Kinadom.

* E-mail: maßlei ku die vietoerar@microsoft.com

Abstract

Coarse-grained (CG) molecular dynamics enables the study of biological processes at temporal and anatial scales that would be intractable at an atomistic resolution. However, accurately learning a CC force field remains a challence. In this work, we leverage connections between score-based generative models, force fields and molecular

Swallowing the Ritter Pill: Simplified Scalable Conformer Generation

Yuyang Wang⁺ Ahmed A, Elhag⁺² Navdeen Jaitly⁺ Joshua M, Susskind⁺ Miguel Ángel Bautista

Abstract

We remember a movel were to readict molecular conformers through a simple formulation that and achieves state of the art results by using the advantages of scale. By training a diffusion generative model directly on 3D atomic positions without making assumptions about the explicit structure of molecules (e.e. modeline torsional angles) we are able to radically simplify strucis the vast complexity of the 3D structure space, encomressing factors such as bond lengths and torsional angles. Despite the molecular graph dictating potential 3D conform ers through specific constraints, such as bond types and spatial arrangements determined by chiral centers, the conformational space experiences exponential growth with the expansion of the graph size and the number of rotatable bonds (Aselrod & Gomez-Bombarelli, 2022). This complicates brate force and exbansive antroaches, meking them virtually unfeasible for even moderately small molecules.

Systematic methods, like OMEGA (Headcins et al., 2010)

Article

Accurate structure prediction of biomolecular interactions with AlphaFold 3

htps://doi.org/103038/s41588-024-01483 w	Josh Abramson ¹⁷ , Jonas Adlar ¹⁷ , Jack Danger ¹⁷ , Richard Exam ¹⁷ , Tim Green ¹⁷ ,
Received 19 December 2023	Alexander Pritzel ¹⁴ , Olaf Ronne berger ¹⁴ , Lindsoy Willmore ¹⁴ , Andress J. Ballard ² , Joshua Barabaluk ¹ , Sabastian W. Budeastak ¹⁴ , David A. Europi ¹⁴ ("bia: "basa")
Accepted: 29 April 2024	Michael O'Neill', David Reiman', Kathryn Turypeuwunakoo?, Zachary Wu', Akulte Zeregel
Published online: 8 May 2024	Einini Arvaniti", Charles Beattle", Ottavia Bertolli", Alex Bridgland", Alexey Cherepanov", Miles Comment, Alexender I, Coner, Elezen ¹ , Antone Conels ¹ , Michael Elezence ¹
Open access	Polisti R. Pochi, 'Issendo Gladinan', Rahok Jain', Yasudi A. Nava', Gavilio R. B. Lawi, Kalar Arcini, Yaso Nozperiek', 'Nava' Bary, 'Bahlwan Shari, Anishi Resala, Anishi Polisianisharan', Catherine Forq', Ingal Yahanon', Elan D. Zhangi', Mahuk Zahakar, Agustan Shari, Yator Tagari Yahanon', Elan D. Zhangi', Demis Hasasha' ¹⁴⁴⁷ B. Jahn M. Arngar ¹⁴¹¹
Check for systeme	
	The introduction of AlphaFold 2 th as spurred a revolution in modelling the structu of proteins and their interactions, enabling a hage range of applications in protein

updated diffusion based architecture that is capable of predicting the joint structure of complexes including parteins, nucleic acids, small molecules, ions and modified residues. The new AlphaErdd model demonstrates substantially improved accuracy

Abuild Zerranduni

Probing the effects of broken symmetries in machine learning

Marcel F Langer . Servey N Pondovakov . and Michele Ceriotti .

Laboratory of Commutational Science and Medeline and National Centre for Commutational Dision and Discourse of Neural Materials * Author to where any correspondence should be addressed.

E-mails michale contestilitenthick

Kerwords: machine learning, symmetry-constrained models, atomistic modeling, molecular simulations Samplementary material for this article is available online

Abstract

Symmetry is one of the most central concepts in physics, and it is no surprise that it has also been widely adopted as an inductive bias for machine-learning models applied to the physical sciences. This is especially true for models targeting the properties of matter at the atomic scale. Both established and state-of-the-art approaches, with almost no exceptions, are built to be exactly equivariant to translations, permutations, and rotations of the atoms. Incorporating symmetries-rotations in particular-constrains the model design space and implies more complicated architectures that are often also computationally demanding. There are indications

The Importance of Being Scalable: Improving the Speed and Accuracy of Neural Network Interatomic Potentials Across Chemical Domains

Eric On UC Berkeley ariconfbarkalay adu

Aditi S. Krishnaneiyan UC Berkeley, LBNL aditiki@harkalay adu

Abstract

Scaling has been a critical factor in improving model performance and generalization across various fields of machine learning. It involves how a model's performance changes with increases in model size or input data, as well as how efficiently computational resources are utilized to support this growth. Despite successes in scaline other types of machine learning models, the study of scaline in Neural Network Interatomic Potentials (NNIPs) remains limited. NNIPs act as

Two for One: Diffusion Models and Force Fields for Coarse-Grained Molecular Dynamics

Marloes Arts, *1.1.¹⁰ Victor Garcia Satorras, *1.¹⁰ Chin-Wei Huang,¹ Daniel Zügner 1 Marco Federici 11 Cecilia Clementi 11 Frank Noé 1 Robert Pinsler # and Rianne van den Berg

1 Work done during an internship at Microsoft Research (Amsterdam). ¹University of Copenhaarn, Department of Computer Science, Universitetaparken 1. Conenhagen, 2100, Denmark ALIScience, Microsoft Research, Evert van de Beekstraat 354, Amsterdam, 1118 CZ, The Netherlands

SALIScience Microsoft Research Karl-Lichbrecht-Straffe 39 Berlin 10178 Cermann University of Amsterdam, Informatics Institute, Science Park 901, Amsterdam, 1098 XII. The Metherlands

1 Preis Universität Rerlin, Department of Physics, Arnimalle 19, Berlin, 1/195, Germann, #AUScience, Microsoft Research, 21 Station Road, Cambridae, CB1 2FB, United Kinadom.

* E-mail: maßlei ku die vietoerar@microsoft.com

Abstract

Coarse-grained (CG) molecular dynamics enables the study of biological processes at temporal and anatial scales that would be intractable at an atomistic resolution. However, accurately learning a CC force field remains a challence. In this work, we leverage connections between score-based generative models, force fields and molecular

Swallowing the Ritter Pill: Simplified Scalable Conformer Generation

Yuyang Wang⁺ Ahmed A, Elhag⁺² Navdeen Jaitly⁺ Joshua M, Susskind⁺ Miguel Ángel Bautista

Abstract

We remember a movel were to readict molecular conformers through a simple formulation that and achieves state of the art results by using the advantures of coals. By training a diffusion and erative model directly on 3D atomic positions without making assumptions about the explicit structure of molecules (e.e. modeline torsional angles) we are able to radically simplify strucis the vast complexity of the 3D structure space, encomressing factors such as bond lengths and torsional angles. Despite the molecular graph dictating potential 3D conform ers through specific constraints, such as bond types and spatial arrangements determined by chiral centers, the conformational space experiences exponential prowth with the expansion of the graph size and the number of rotatable bonds (Aselrod & Gomez-Bombarelli, 2022). This complicates brate force and exbansive antroaches, meking them virtually unfeasible for even moderately small molecules.

Systematic methods, like OMEGA (Headcins et al., 2010)

DOES EQUIVARIANCE MATTER AT SCALE?

Johann Brohmer' Sieke Behrends Pim de Haan' Toco Cohen' Oudcorrers AI Research

ABSTRACT

Given large data sets and sufficient compate, is it beneficial to design neural architectures for the structure and symmetries of each problem? Or is it more efficient to leave there from date? We study empirically have empirized and non-conjuncted networks acide with commute and training samples. Econome on a benchmark medhere of rigid, had vinteractions and on seneral memore transformer architectures, we perform a series of experiments, varying the model size, training steps, and dataset size. We find esidence for these conclusions. First conjugations interrores data efficiency, but training non-consistent models with data anotheristical can close this cap river sufficient erochs. Second, scaling with compute follows a power law, with equivariant models outeerforming non-equivariant ones at each tested commute budget. Finally, the optimal allocation of a commute budget onto model

🖒 Easy to implement

- 🖒 Easy to implement
- 凸 No specialized architecture necessary
- ர No exact equivariance

🖒 Easy to implement

凸 No specialized architecture necessary

ர No exact equivariance

Can we understand data augmentation theoretically?

• Consider continuous gradient descent

$$\frac{\mathrm{d}\theta_{\mu}}{\mathrm{d}t} = -\eta \frac{\partial \mathcal{L}(\mathcal{N}_{\theta}, \mathcal{D})}{\partial \theta_{\mu}} = -\frac{\eta}{N} \sum_{i=1}^{N} \frac{\partial L(\mathcal{N}_{\theta}(x_i), y_i)}{\partial \theta_{\mu}}$$

• Consider continuous gradient descent

$$\frac{\mathrm{d}\theta_{\mu}}{\mathrm{d}t} = -\eta \frac{\partial \mathcal{L}(\mathcal{N}_{\theta}, \mathcal{D})}{\partial \theta_{\mu}} = -\frac{\eta}{N} \sum_{i=1}^{N} \frac{\partial L(\mathcal{N}_{\theta}(x_i), y_i)}{\partial \theta_{\mu}}$$

• Then, the network evolves according to

$$\frac{\mathrm{d}\mathcal{N}_{\theta}(x)}{\mathrm{d}t} = \sum_{\mu} \frac{\partial\mathcal{N}_{\theta}(x)}{\partial\theta_{\mu}} \frac{\mathrm{d}\theta_{\mu}}{\mathrm{d}t}$$

• Consider continuous gradient descent

$$\frac{\mathrm{d}\theta_{\mu}}{\mathrm{d}t} = -\eta \frac{\partial \mathcal{L}(\mathcal{N}_{\theta}, \mathcal{D})}{\partial \theta_{\mu}} = -\frac{\eta}{N} \sum_{i=1}^{N} \frac{\partial L(\mathcal{N}_{\theta}(x_i), y_i)}{\partial \theta_{\mu}}$$

• Then, the network evolves according to

$$\frac{\mathrm{d}\mathcal{N}_{\theta}(x)}{\mathrm{d}t} = \sum_{\mu} \frac{\partial\mathcal{N}_{\theta}(x)}{\partial\theta_{\mu}} \frac{\mathrm{d}\theta_{\mu}}{\mathrm{d}t} = -\frac{\eta}{N} \sum_{i=1}^{N} \sum_{\mu} \frac{\partial\mathcal{N}(x)}{\partial\theta_{\mu}} \frac{\partial\mathcal{N}(x_{i})}{\partial\theta_{\mu}} \frac{\partial\mathcal{L}(\mathcal{N}_{\theta}(x_{i}), y_{i})}{\partial\mathcal{N}(x_{i})}$$

• Consider continuous gradient descent

$$\frac{\mathrm{d}\theta_{\mu}}{\mathrm{d}t} = -\eta \frac{\partial \mathcal{L}(\mathcal{N}_{\theta}, \mathcal{D})}{\partial \theta_{\mu}} = -\frac{\eta}{N} \sum_{i=1}^{N} \frac{\partial L(\mathcal{N}_{\theta}(x_i), y_i)}{\partial \theta_{\mu}}$$

• Then, the network evolves according to

$$\frac{\mathrm{d}\mathcal{N}_{\theta}(x)}{\mathrm{d}t} = \sum_{\mu} \frac{\partial\mathcal{N}_{\theta}(x)}{\partial\theta_{\mu}} \frac{\mathrm{d}\theta_{\mu}}{\mathrm{d}t} = -\frac{\eta}{N} \sum_{i=1}^{N} \sum_{\mu} \frac{\partial\mathcal{N}(x)}{\partial\theta_{\mu}} \frac{\partial\mathcal{N}(x_{i})}{\partial\theta_{\mu}} \frac{\partial\mathcal{L}(\mathcal{N}_{\theta}(x_{i}), y_{i})}{\partial\mathcal{N}(x_{i})}$$

With the empirical neural tangent kernel (NTK)

$$\Theta_{ij}^{\theta}(x,x') = \sum_{\mu} \frac{\partial \mathcal{N}_i(x)}{\partial \theta_{\mu}} \frac{\partial \mathcal{N}_j(x')}{\partial \theta_{\mu}}$$

Training dynamics under continuous gradient descent:



Infinite width limit

[Jacot et al. 2018]


[Jacot et al. 2018]



Consider an MLP in NTK parametrization

$$z^{(\ell)} = \frac{1}{\sqrt{n_{\ell-1}}} W^{(\ell)} \sigma(z^{(\ell-1)}(x)), \quad W^{(\ell)} \in \mathbb{R}^{n_{\ell} \times n_{\ell-1}}, \quad W^{(\ell)}_{ij} \sim \mathcal{N}(0, 1)$$

Consider an MLP in NTK parametrization

$$z^{(\ell)} = \frac{1}{\sqrt{n_{\ell-1}}} W^{(\ell)} \sigma(z^{(\ell-1)}(x)), \quad W^{(\ell)} \in \mathbb{R}^{n_{\ell} \times n_{\ell-1}}, \quad W^{(\ell)}_{ij} \sim \mathcal{N}(0,1)$$

At infinite width

$$z_{j}^{(\ell)}(x) = \sqrt{n_{\ell-1}} \frac{1}{n_{\ell-1}} \sum_{j=1}^{n_{\ell}} W_{ij}^{(\ell)} \sigma(z_{j}^{(\ell-1)}(x))$$

Consider an MLP in NTK parametrization

$$z^{(\ell)} = \frac{1}{\sqrt{n_{\ell-1}}} W^{(\ell)} \sigma(z^{(\ell-1)}(x)), \quad W^{(\ell)} \in \mathbb{R}^{n_{\ell} \times n_{\ell-1}}, \quad W^{(\ell)}_{ij} \sim \mathcal{N}(0,1)$$

At infinite width

$$z_{j}^{(\ell)}(x) = \sqrt{n_{\ell-1}} \underbrace{\frac{1}{n_{\ell-1}} \sum_{j=1}^{n_{\ell}} \underbrace{W_{jj}^{(\ell)} \sigma(z_{j}^{(\ell-1)}(x))}_{\text{i.i.d.}}}_{\text{i.i.d.}}$$

Consider an MLP in NTK parametrization

$$z^{(\ell)} = \frac{1}{\sqrt{n_{\ell-1}}} W^{(\ell)} \sigma(z^{(\ell-1)}(x)), \quad W^{(\ell)} \in \mathbb{R}^{n_{\ell} \times n_{\ell-1}}, \quad W^{(\ell)}_{ij} \sim \mathcal{N}(0,1)$$

At infinite width

$$z_{j}^{(\ell)}(x) = \sqrt{n_{\ell-1}} \underbrace{\frac{1}{n_{\ell-1}} \sum_{j=1}^{n_{\ell}} \underbrace{W_{ij}^{(\ell)} \sigma(z_{j}^{(\ell-1)}(x))}_{\text{i.i.d.}}}_{\text{i.i.d.}} \sim \mathcal{N}(0, \underbrace{\operatorname{Cov}(z_{j}^{(\ell)}, z_{j}^{(\ell)})}_{\text{NNGP}})$$

[Jacot et al. 2018]



[Jacot et al. 2018]



凸 NTK becomes independent of initialization

[Jacot et al. 2018]



凸 NTK becomes independent of initialization

Ճ NTK becomes constant in training



凸 NTK becomes independent of initialization

凸 NTK becomes constant in training

凸 NTK can be computed for most networks



- 凸 NTK becomes independent of initialization
- 凸 NTK becomes constant in training
- 凸 NTK can be computed for most networks
- ✓ Training dynamics can be solved

NTK at initialization

When taking the layer widths to infinity sequentially, the empirical NTK $\Theta_{ij}^{\theta}(x, x')$ at initialization converges in probability to a deterministic kernel $\Theta(x, x')\delta_{ij}$

When taking the layer widths to infinity sequentially, the empirical NTK $\Theta_{ij}^{\theta}(x,x')$ at initialization converges in probability to a deterministic kernel $\Theta(x,x')\delta_{ij}$

• The deterministic kernel is given in terms of a recursion over layers

When taking the layer widths to infinity sequentially, the empirical NTK $\Theta_{ij}^{\theta}(x,x')$ at initialization converges in probability to a deterministic kernel $\Theta(x,x')\delta_{ij}$

- The deterministic kernel is given in terms of a recursion over layers
- For most common architectures, this recursion can be performed explicitly, e.g. using neural-tangents Python package [Noval

[Novak et al. 2020]

Frozen NTK

[Jacot et al. 2020]

For a nonlinearity which is Lipschitz, twice differentiable and has bounded second derivative,

$$\Theta_{ij}^{\theta_t}(x,x') \to \Theta(x,x')\delta_{ij}$$

uniformly in t as the layer widths go to infinity sequentially.

Frozen NTK

[Jacot et al. 2020]

For a nonlinearity which is Lipschitz, twice differentiable and has bounded second derivative,

$$\Theta_{ij}^{\theta_t}(x,x') \to \Theta(x,x')\delta_{ij}$$

uniformly in *t* as the layer widths go to infinity sequentially.

• Intuitively, this happens because the weight updates vanish in the limit $n \rightarrow \infty$

Frozen NTK

[Jacot et al. 2020]

For a nonlinearity which is Lipschitz, twice differentiable and has bounded second derivative,

$$\Theta_{ij}^{\theta_t}(x,x') \to \Theta(x,x')\delta_{ij}$$

uniformly in *t* as the layer widths go to infinity sequentially.

- Intuitively, this happens because the weight updates vanish in the limit $n \rightarrow \infty$
- However, the network still learns

[Jacot et al. 2018]

$$\mu_t(x) = \Theta(x, X) \Theta(X, X)^{-1} (\mathbb{I} - e^{-\eta \Theta(X, X)t}) Y$$

[Jacot et al. 2018]

$$\mu_t(x) = \Theta(x, X) \Theta(X, X)^{-1} (\mathbb{I} - e^{-\eta \Theta(X, X)t}) Y$$

[Jacot et al. 2018]

$$\mu_t(x) = \Theta(x, X) \Theta(X, X)^{-1} (\mathbb{I} - e^{-\eta \Theta(X, X)t}) Y$$
train data

[Jacot et al. 2018]



[Jacot et al. 2018]



Data augmentation

$$\mu_t(x) = \Theta(x, X) \Theta(X, X)^{-1} (\mathbb{I} - e^{-\eta \Theta(X, X)t}) Y$$





Kernel transformation

The neural tangent kernel Θ as well as the NNGP kernel K transform according to

$$\Theta(\rho(g)x,\rho(g)x') = \rho_{\mathcal{K}}(g)\Theta(x,x')\rho_{\mathcal{K}}^{\mathsf{T}}(g),$$

$$\mathcal{K}(\rho(g)x,\rho(g)x') = \rho_{\mathcal{K}}(g)\mathcal{K}(x,x')\rho_{\mathcal{K}}^{\mathsf{T}}(g),$$

for all $g \in G$ and $x, x' \in X$.

Kernel transformation

The neural tangent kernel Θ as well as the NNGP kernel K transform according to

$$\Theta(\rho(g)x,\rho(g)x') = \rho_{\mathcal{K}}(g)\Theta(x,x')\rho_{\mathcal{K}}^{\mathsf{T}}(g),$$

$$\mathcal{K}(\rho(g)x,\rho(g)x') = \rho_{\mathcal{K}}(g)\mathcal{K}(x,x')\rho_{\mathcal{K}}^{\mathsf{T}}(g),$$

for all $g \in G$ and $x, x' \in X$.

Hence, for MLPs,

$$\Theta(\rho(g)x,\rho(g)x') = \Theta(x,x') \quad \Rightarrow \quad \Theta(\rho(g)x,x') = \Theta(x,\rho^{-1}(g)x')$$

Permutation shift

• On the training data, group transformations permute the samples

$$\rho(g)x_i = x_{\pi_g(i)}, \qquad \pi_g \in S_N$$

Permutation shift

• On the training data, group transformations permute the samples

$$\rho(g)x_i = x_{\pi_g(i)}, \qquad \pi_g \in S_N$$

• Therefore, for a permutation of training samples associate to g

$$(g)\Theta(X,X) = \Theta(\rho(g)X,X)$$
$$= \Theta(X,\rho^{-1}(g)X)$$
$$= \Theta(X,X)(\Pi^{-1}(g))^{\top}$$
$$= \Theta(X,X)\Pi(g)$$







$$\mu_t(\rho(g)x) = \Theta(x,X)\Theta(X,X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\underbrace{\rho(g)Y}_{=Y}$$
for invariance

$$\mu_t(\rho(g)x) = \Theta(x,X)\Theta(X,X)^{-1}(\mathbb{I} - e^{-\eta\Theta(X,X)t})\underbrace{\rho(g)Y}_{=Y}$$

$$= \mu_t(x)$$
for invariance

Mean prediction

 $\mu_t(x)$

Mean prediction

$$\mu_t(x) = \mathbb{E}_{\theta_0 \sim \text{initializations}}[\mathcal{N}_{\theta_t}(x)]$$
Mean prediction

$$\mu_t(x) = \mathbb{E}_{\theta_0 \sim \text{initializations}} [\mathcal{N}_{\theta_t}(x)] = \lim_{n \to \infty} \frac{1}{n} \sum_{\theta_0 = \text{init}_1}^{\text{init}_n} \mathcal{N}_{\theta_t}(x)$$

Mean prediction

$$\mu_t(x) = \mathbb{E}_{\theta_0 \sim \text{initializations}} [\mathcal{N}_{\theta_t}(x)] = \lim_{n \to \infty} \underbrace{\frac{1}{n} \sum_{\theta_0 = \text{init}_1}^{\text{init}_n} \mathcal{N}_{\theta_t}(x)}_{\text{mean prediction of deep ensemble}}$$

- ✓ Proof of exact equivariance for
 - full data augmentation
 - infinite ensembles

- ✓ Proof of exact equivariance for
 - full data augmentation
 - infinite ensembles
- ✓ Equivariance holds for all training times

- ✓ Proof of exact equivariance for
 - full data augmentation
 - infinite ensembles
- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

Deep ensembles trained with data augmentation are equivariant.

- ✓ Proof of exact equivariance for
 - full data augmentation
 - infinite ensembles
- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data
- ✓ Holds also for finite-width networks

[Nordenfors, Flinth 2024]

Intuitive explanation

- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

Intuitive explanation

- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

• At infinite width, the mean output at initialization is zero everywhere.

Intuitive explanation

- ✓ Equivariance holds for all training times
- ✓ Equivariance holds away from the training data

- At infinite width, the mean output at initialization is zero everywhere.
- Training with full data augmentation leads to an equivariant function.

Toy example









After 1 Training Step



After 2 Training Steps



After 3 Training Steps



After 2000 Training Steps



After 2000 Training Steps





After 1 Training Step



After 2 Training Steps



After 3 Training Steps



After 2000 Training Steps



After 2000 Training Steps



What Does An Augmented Ensemble Converge To?











Consider two ensembles:

trained without data augmentation

trained with data augmentation

Consider two ensembles:



Consider two ensembles:



Then

$$\mu_t^{\text{non-aug}}(x) = \mu_t^{\text{aug}}(x)$$

at infinite width.

Consider two ensembles:



Then

$$\mu_t^{non-aug}(x) = \mu_t^{aug}(x) \quad \forall t$$

at infinite width.
Data augmentation and NTKs

Consider two ensembles:



Then

$$\mu_t^{\text{non-aug}}(x) = \mu_t^{\text{aug}}(x) \quad \forall t \quad \forall x$$

at infinite width.

Data augmentation and NTKs

$$\Theta^{\mathsf{non-aug}}(f,f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\mathsf{aug}}(f,\rho_{\mathsf{reg}}(g)f')$$

Data augmentation and NTKs

$$\Theta^{\mathsf{non-aug}}(f,f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\mathsf{aug}}(f, \rho_{\mathsf{reg}}(g)f')$$

 Given an architecture with NTK ⊖^{aug}, find an architecture with NTK ⊖^{non-aug}

[Cohen, Welling 2016]

[Cohen, Welling 2016]

Group conv's are the (unique) linear layers equivariant wrt ρ_{reg}

[Cohen, Welling 2016]

Group conv's are the (unique) linear layers equivariant wrt ρ_{reg}

• Ordinary convolutions

$$f'(y) = \int_X \mathrm{d}x \,\kappa(x-y) \,f(x)$$

[Cohen, Welling 2016]

Group conv's are the (unique) linear layers equivariant wrt ρ_{reg}

• Ordinary convolutions

$$f'(y) = \int_X \mathrm{d}x \,\kappa(x-y) \,f(x)$$

• Group convolutions

$$f'(g) = \int_X dx \kappa(\rho(g^{-1})x) f(x)$$
 lifting

[Cohen, Welling 2016]

Group conv's are the (unique) linear layers equivariant wrt ρ_{reg}

• Ordinary convolutions

$$f'(y) = \int_X \mathrm{d}x \,\kappa(x-y) \,f(x)$$

• Group convolutions

$$f'(g) = \int_{X} dx \,\kappa(\rho(g^{-1})x) f(x) \qquad \text{lifting}$$
$$f'(g) = \int_{G} dg \,\kappa(g^{-1}h) f(h) \qquad \text{group convolution}$$

[Cohen, Welling 2016]

Group conv's are the (unique) linear layers equivariant wrt ρ_{reg}

• Ordinary convolutions

$$f'(y) = \int_X \mathrm{d}x \,\kappa(x-y) \,f(x)$$

• Group convolutions

$$f'(g) = \int_{X} dx \,\kappa(\rho(g^{-1})x) f(x) \qquad \text{lifting}$$

$$f'(g) = \int_G dg \kappa(g^{-1}h) f(h)$$
 group convolution

$$f' = \frac{1}{\operatorname{vol}(G)} \int_G \mathrm{d}g f(g)$$
 group pooling















For GCNN-layers, define the NNGP and NTK via

$$\mathcal{K}_{\boldsymbol{g},\boldsymbol{g}'}^{(\ell)}(\boldsymbol{f},\boldsymbol{f}') = \mathbb{E}\left[\left[\mathcal{N}^{(\ell)}(\boldsymbol{f})\right](\boldsymbol{g})\left(\left[\mathcal{N}^{(\ell)}(\boldsymbol{f}')\right](\boldsymbol{g}')\right)^{\mathsf{T}}\right]$$

For GCNN-layers, define the NNGP and NTK via

$$\begin{split} & \mathcal{K}_{g,g'}^{(\ell)}(f,f') = \mathbb{E}\left[[\mathcal{N}^{(\ell)}(f)](g) \left([\mathcal{N}^{(\ell)}(f')](g') \right)^{\mathsf{T}} \right] \\ & \Theta_{g,g'}^{(\ell)}(f,f') = \mathbb{E}\left[\sum_{\ell'=1}^{\ell} \frac{\partial [\mathcal{N}^{(\ell)}(f)](g)}{\partial \theta^{(\ell')}} \left(\frac{\partial [\mathcal{N}^{(\ell)}(f')](g')}{\partial \theta^{(\ell')}} \right)^{\mathsf{T}} \right] \end{split}$$

$$[\mathcal{N}^{(\ell)}(f)](g) = \int_{\mathcal{G}} \mathrm{d}g \,\kappa(g^{-1}h) \left[\mathcal{N}^{(\ell-1)}(f)\right](h)$$

The layer-recursion for a GCNN-layer is given by

$$\mathcal{K}_{g,g'}^{(\ell+1)}(f,f') = \frac{1}{|S_{\kappa}|} \int_{S_{\kappa}} \mathrm{d}h \,\mathcal{K}_{gh,g'h}^{(\ell)}(f,f')$$

$$[\mathcal{N}^{(\ell)}(f)](g) = \int_{\mathcal{G}} \mathrm{d}g \,\kappa(g^{-1}h) \left[\mathcal{N}^{(\ell-1)}(f)\right](h)$$

The layer-recursion for a GCNN-layer is given by

$$\begin{aligned} & \mathcal{K}_{g,g'}^{(\ell+1)}(f,f') = \frac{1}{|S_{\kappa}|} \int_{S_{\kappa}} dh \, \mathcal{K}_{gh,g'h}^{(\ell)}(f,f') \\ & \Theta_{g,g'}^{(\ell+1)}(f,f') = \mathcal{K}_{g,g'}^{(\ell+1)}(f,f') + \frac{1}{|S_{\kappa}|} \int_{S_{\kappa}} dh \, \Theta_{gh,g'h}^{(\ell)}(f,f') \end{aligned}$$



Stack GConv-layers to obtain an invariant network



Stack GConv-layers to obtain an invariant network



Compute NTK with layer-wise recursion

0

Stack GConv-layers to obtain an invariant network



$$0 \longrightarrow \Theta_{g,g'}^{(1)}(f,f')$$

Stack GConv-layers to obtain an invariant network



$$0 \longrightarrow \Theta_{g,g'}^{(1)}(f,f') \longrightarrow \Theta_{g,g'}^{(2)}(f,f')$$

Stack GConv-layers to obtain an invariant network



$$0 \longrightarrow \Theta_{g,g'}^{(1)}(f,f') \longrightarrow \Theta_{g,g'}^{(2)}(f,f') \longrightarrow \Theta_{g,g'}^{(3)}(f,f')$$

Stack GConv-layers to obtain an invariant network



$$0 \longrightarrow \Theta_{g,g'}^{(1)}(f,f') \longrightarrow \Theta_{g,g'}^{(2)}(f,f') \longrightarrow \Theta_{g,g'}^{(3)}(f,f') \longrightarrow \Theta_{g,g'}^{(4)}(f,f')$$

Stack GConv-layers to obtain an invariant network



$$0 \longrightarrow \Theta_{g,g'}^{(1)}(f,f') \longrightarrow \Theta_{g,g'}^{(2)}(f,f') \longrightarrow \Theta_{g,g'}^{(3)}(f,f') \longrightarrow \Theta_{g,g'}^{(4)}(f,f') \longrightarrow \Theta_{g,g'}^{(5)}(f,f')$$

Stack GConv-layers to obtain an invariant network



$$\Theta \longrightarrow \Theta_{g,g'}^{(1)}(f,f') \longrightarrow \Theta_{g,g'}^{(2)}(f,f') \longrightarrow \Theta_{g,g'}^{(3)}(f,f') \longrightarrow \Theta_{g,g'}^{(4)}(f,f') \longrightarrow \Theta_{g,g'}^{(5)}(f,f') \longrightarrow \Theta(f,f')$$

• Consider two neural networks

• Consider two neural networks

An MLP



• Consider two neural networks

An MLP







• Consider two neural networks

An MLP



• Consider two neural networks

An MLP


An MLP



An MLP



An MLP



An MLP



- Consider two neural networks
 - An MLP



• Then

$$\Theta^{\mathsf{GCNN}}(f,f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\mathsf{MLP}}(f,\rho_{\mathsf{reg}}(g)f')$$

$$\Theta^{\mathsf{GCNN}}(f,f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\mathsf{MLP}}(f, \rho_{\mathsf{reg}}(g)f')$$

before: non-aug

$$\Theta^{\text{GCNN}}(f, f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\text{MLP}}(f, \rho_{\text{reg}}(g)f')$$

before: non-aug

$$\Theta^{\text{GCNN}}(f, f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\text{MLP}}(f, \rho_{\text{reg}}(g)f')$$

before: non-aug

$$\Theta^{\text{GCNN}}(f, f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\text{MLP}}(f, \rho_{\text{reg}}(g)f')$$

training the MLP on *G*-augmented data

before: non-aug

$$\Theta^{\text{GCNN}}(f, f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\text{MLP}}(f, \rho_{\text{reg}}(g)f')$$

ᡌ

G-augmented data = training the GCNN on unaugmented data

before: non-aug

$$\Theta^{\text{GCNN}}(f, f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\text{MLP}}(f, \rho_{\text{reg}}(g)f')$$



training the MLP on *G*-augmented data

training the GCNN on unaugmented data

in the ensemble mean

before: non-aug

$$\Theta^{\text{GCNN}}(f, f') = \frac{1}{|G|} \sum_{g \in G} \Theta^{\text{MLP}}(f, \rho_{\text{reg}}(g)f')$$

⇔

training the MLP on *G*-augmented data

training the GCNN on unaugmented data

- in the ensemble mean, $\forall t, \forall x$

• Consider a CNN



• Consider a CNN and a GCNN invariant wrt. roto-translations



• Consider a CNN and a GCNN invariant wrt. roto-translations



• Then

$$\Theta^{\text{GCNN}}(f, f') = \frac{1}{n} \sum_{r \in C_n} \Theta^{\text{CNN}}(f, \rho_{\text{reg}}(r)f')$$

• Consider a CNN and a GCNN invariant wrt. roto-translations



• Then

$$\Theta^{\text{GCNN}}(f,f') = \frac{1}{n} \sum_{r \in C_n} \Theta^{\text{CNN}}(f,\rho_{\text{reg}}(r)f')$$

⇒ By training the CNN on rotated images, one obtains a roto-translation invariant GCNN

Experiments









Relative Standard Deviation









Histological slices

[Kather et al. 2018]



Histological slices

[Kather et al. 2018]



Histological slices







Histological slices



Ensemble size 5








Out of distribution results



✓ Emergent invariance for rotated FashionMNIST

- ✓ Emergent invariance for rotated FashionMNIST
- ✓ Partial augmentation for continuous symmetries

- ✓ Emergent invariance for rotated FashionMNIST
- ✓ Partial augmentation for continuous symmetries
- ✓ Emergent equivariance (as opposed to invariance)

Comparison to other methods

Comparison to other methods

➡ Models trained on rotated FashionMNIST

Comparison to other methods

➡ Models trained on rotated FashionMNIST

Orbit same predictions out of distribution:

	C4	<i>C</i> ₈	C ₁₆
DeepEns+DA	$3.85{\pm}0.12$	7.72±0.34	15.24±0.69
only DA	$3.41{\pm}0.18$	$6.73 {\pm} 0.24$	12.77 ± 0.71
E2CNN ¹	4 ± 0.0	7.71±0.21	$\textbf{15.08}{\pm}\textbf{0.34}$
Canon ²	4 ± 0.0	7.45±0.14	12.41 ± 0.85

¹[Weiler et al. 2019], ²[Kaba et al. 2022]







Key takeaways

Key takeaways

If you need ensembles

பி use data augmentation to obtain an equivariant model.

Key takeaways

If you need ensembles

பி use data augmentation to obtain an equivariant model.

Papers

- Emergent Equivariance in Deep Ensembles Jan E. Gerken*, Pan Kessel* ICML 2024 (Oral)
 - * Equal contribution
- Equivariant Neural Tangent Kernels Philipp Misof, Pan Kessel, Jan E. Gerken ICML 2025



Thank you

Backup

Emergent equivariance of cross products



Histological Data – OOD samples



Emergent continuous symmetry on FashionMNIST



Kernel convergence



NNGP



NTK

Equivariant NTKs for medical image classification



Equivariant NTKs for molecular property regression



OOD samples for CNN to GCNN convergence

MNIST



