RAPPORT D'AVANCEMENT DE THÈSE POUR LE COMITÉ DE SUIVI N° 1

Alexandre Ramé
alexandre.rame@isir.upmc.fr
https://alexrame.github.io/

PhD Title: Diversity in deep networks and data to boost generalization for computer vision
PhD Supervisor: Matthieu Cord
Host laboratory: ISIR (Institut des systèmes intelligents et de robotique)
Doctoral school: EDITE
PhD starting date: January 01, 2021
Funding: VISA-DEEP chair of Matthieu Cord (ANR National French Government program on AI)
Jury: Philippe Esling and Andrei Bursuc

1 INTRODUCTION

Deep learning algorithms are currently used for a wide range of applications, such as natural language processing (Vaswani et al., 2017), acoustics analysis (Hinton et al., 2012) or graph understanding (Scarselli et al., 2008). Another key application is computer vision (Krizhevsky et al., 2012), where networks take pixels as input and try to understand what they represent. The last decade has seen great progress — in network architectures (Ioffe & Szegedy, 2015; Dosovitskiy et al., 2021), learning methods (Goodfellow et al., 2014) or data augmentation (Yun et al., 2019) strategies — mainly seeking to train the more adequate network architecture with the appropriate loss on a predefined dataset. Specifically, they mostly consider one model trained on one dataset. Yet, this goes against everything that is challenging when trying to actually use deep networks to analyze images in real life. In this thesis, we go beyond these simplifying assumptions by considering multiple models trained on multiple datasets.

Regarding the models, they are rarely unique in real applications, where several models are often trained. Indeed, most predictive systems rely on multiple and diverse models; this is because averaging their different predictions reduce the variance of the estimator and significantly improve the generalization ability (Nilsson, 1965). This ensembling strategy is especially efficient in deep learning (Hansen & Salamon, 1990; Lakshminarayanan et al., 2017) and is the de facto solution for most Kaggle competitions (Hin, 2020).

Regarding the data, real life datasets are more complex than those (Deng et al., 2009; Krizhevsky & Hinton, 2009) studied by the community, and this for many different reasons. First, real datasets are generally not identically distributed. Notably, the data in train and in test can be very diverse, as much in the inputs as in the correlations between the input and the label; this goes against a key assumption in most learning theories. Thus, standard strategies lack robustness to these distribution shifts (Ovadia et al., 2019; Arjovsky et al., 2019). Second, datasets are constantly evolving and can grow (Rebuffi et al., 2017) to include more classes or more domains of applications.

This thesis aims at improving the robustness of deep learning networks, not despite the complexity of the models and data, but thanks to their diversity, respectively in predictions and in distribution. In brief, we study how the diversity across these models and datasets can enhance deep learning generalization abilities for classification in computer vision tasks. Moreover, analyzing these notions of diversity gives us the opportunity to better understand why deep learning work so well generally.

We summarize our lines of research as follows:

- First, we analyze deep ensembles, where multiple models are trained independently and then averaged to improve performances. We propose to increase the diversity across the models (Rame & Cord, 2021) and then tackle their computational overhead (Rame et al., 2021b).
- Second, we consider setups in which multiple domains are given. When these domains contain the same classes, we propose a new invariant regularization (Rame et al., 2021a) to promote the learning of a causal mechanism consistent across domains and improve ouf-of-distribution gener-

alization. When these domains are given sequentially, we propose a new transformer architecture for continual learning (Douillard et al., 2022).

- Finally, we discuss our current project where we learn multiple networks on multiple domains.

2 Multiple models: ensembling

2.1 Deep ensembles

Context Ensembling, *i.e.*, averaging the predictions of multiple models, was one of the most popular research topic before the advent of deep learning (Nilsson, 1965): popular examples are random forests (Breiman, 2001) or xgboost (Chen & Guestrin, 2016) algorithms. Yet, in deep computer vision (Krizhevsky et al., 2012), ensembling of networks (Krogh & Vedelsby, 1995) were initially arguably underlooked. Now, their effectiveness is undisputed (Lakshminarayanan et al., 2017) even at fixed number of parameters (Chirkova et al., 2020); they are often analyzed as a simple alternative to fully Bayesian methods (Blundell et al., 2015; Gal & Ghahramani, 2016).

Bias variance covariance decomposition The reason why ensembling improves generalization is easily grasped by analyzing the Bias-Variance-Covariance Decomposition (Ueda & Nakano, 1996), that generalizes the Bias-Variance Decomposition (Kohavi et al., 1996). The expected squared error between the prediction of an ensemble of M members and the true label y can be written as:

$$\mathbb{E}[(\overline{f}-y)^2] = \overline{\text{bias}}^2 + \frac{1}{M}\overline{\text{var}} + (1-\frac{1}{M})\overline{\text{covar}}, \tag{1}$$

where

$$\begin{aligned}
\overline{\text{bias}} &= \frac{1}{M}\sum_i(\mathbb{E}[f_i]-y),\\
\overline{\text{var}} &= \frac{1}{M}\sum_i \mathbb{E}[(\mathbb{E}[f_i]-y)^2],\\
\overline{\text{covar}} &= \frac{1}{M(M-1)}\sum_i\sum_{j\neq i}\mathbb{E}[(f_i-\mathbb{E}[f_i])(f_j-\mathbb{E}[f_j])].
\end{aligned}$$

The reduction factor of the variance component equals to M when errors are uncorrelated, *i.e.*, when members are diverse. Brown et al. (2005a;b) summarized it this way: “in addition to the bias and variance of the individual estimators, the generalisation error of an ensemble also depends on the covariance between the individuals. This raises the interesting issue of why we should ever train ensemble members separately; why shouldn’t we try to find some way to capture the effect of the covariance in the error function?”.

Standard strategies In standard deep ensembles, diversity is mostly due to different initializations (Lakshminarayanan et al., 2017). To obtain more diverse ensembles, we could adapt the training samples through bagging (Breiman, 1996) and bootstrapping (Efron & Tibshirani, 1994), but a reduction of training samples has a negative impact on members with multiple local minima (Lee et al., 2015). Liu & Yao (1999a;b); Brown et al. (2005b) explicitly quantified the diversity and regularized members into having negatively correlated errors. However, these ideas have not significantly improved ensemble accuracy when applied to deep learning (Shui et al., 2018; Pang et al., 2019): even if they did manage to increase diversity, it was at the cost of of drastically decreased individual members’ performances.

2.2 DICE: diversity in features

Diversity in features Our goal in Rame & Cord (2021) is to find a training strategy to reach an improved trade-off between ensemble diversity and individual accuracies (Masegosa, 2020). Our core approach is to encourage all members to predict the same thing, but for different reasons. Therefore we argue the diversity should be enforced in the features space and not on predictions. Intuitively, to maximize the impact of a new member, extracted features should bring information about the target that is unpredictable from another members’ features. As if, to recognize a cow in a picture, we had one network specialized to detect shapes and the other specialized on textures. This is illustrated in Figure 2.

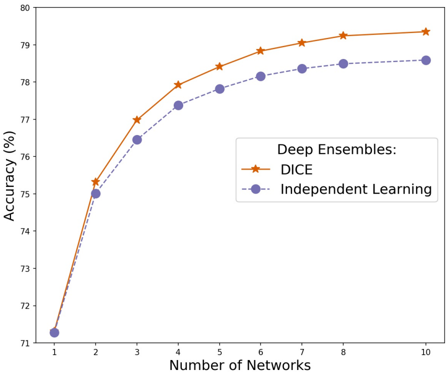


Figure 1: **DICE better leverages additional networks by increasing their diversity**: 5 networks trained with DICE match 7 networks trained independently. Setup: CIFAR-100 (Krizhevsky & Hinton, 2009) with ResNet-32 (He et al., 2016).

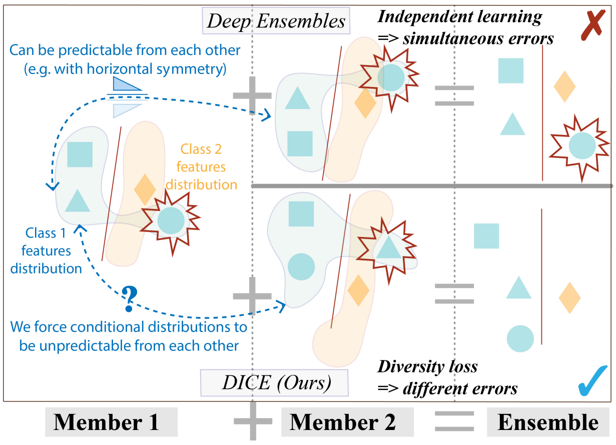


Figure 2: **Outline**. DICE prevents features from being predictable from each other *conditionally* upon the target class. Features extracted by members (1, 2) from one input (●,●) should not share more information than features from two inputs in the same class (●,▲): *i.e.*, (●,-) should not be able to differentiate (-,●) and (-,▲).

Information bottleneck We justify our approach by leveraging information bottleneck (IB) principles (Tishby, 2001; Alemi et al., 2017; Fischer, 2020): features extracted by the networks should have sufficient information to predict the label, but should also be concise, by forgetting the task-irrelevant information — measured in terms of entropy (Shannon, 1948).

DICE: conditional redundancy minimization Applying these principles to deep ensembles and considering the pairwise interactions between members $i, j \in \{1, \ldots, M\}$, we minimize the conditional redundancy $I(Z_i; Z_j|Y)$, where Z_i is the features extracted by member i and Y is the label. This regularization aims at splitting the useful information into the two features. Critically, the conditioning on label Y protects features' informativeness; members are only forced to have independent bias. It encourages diversity while protecting members' individual precision. Overall, this targets spurious correlations, *e.g.*, information redundantly shared among features extracted by different members but useless for label prediction.

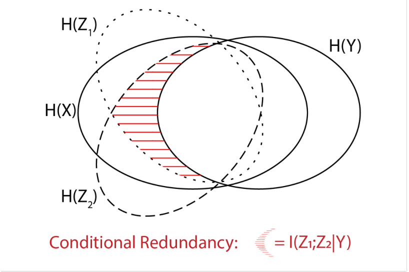


Figure 3: **Venn Information Diagram** (Yeung, 1991). DICE minimizes conditional redundancy.

Neural estimation of mutual information Despite being a pivotal measure, mutual information estimation historically relied on nearest neighbors (Singh et al., 2003; Kraskov et al., 2004; Gao et al., 2018) or density kernels (Kandasamy et al., 2015) that do not scale well in high dimensions. We benefit from recent advances in neural estimation of mutual information (Belghazi et al., 2018), built on optimizing dual representations (Donsker & Varadhan, 1975) of the KL divergence. Mukherjee et al. (2020) extended this formulation for conditional mutual information estimation. This leads to an adversarial strategy, where a discriminator learns to distinguish between features extracted from the same image or from two different images belonging to the same label. This adversarial training removes information shared between members that are not found in other images from the same label.

Experiments We consistently improve accuracy on CIFAR-100 as summarized in Figure 1, with also better uncertainty estimation and calibration. We analyze how of our loss modify the accuracy-diversity trade-off towards a more efficient cooperation. We also improve out-of-distribution detection: while a single network has difficulty saying "I don't know", the disagreements across predictions can detect complex images, for which human expertise would be needed.

2.3 MixMo: Ensembling for free

Ensembling cost Ensembling's drawback is the inherent computational and memory overhead, which increases linearly with the number of members. This is an untenable cost in many real world applications, for example when networks must fit on tiny embedded chips. This is even more prohibitive when one knows the carbon footprint of deep learning. This bottleneck is typically addressed by sacrificing either individual performance or diversity in a complex trade-off, usually through some degree of parameter sharing between models (Lee et al., 2015; Wen et al., 2019).

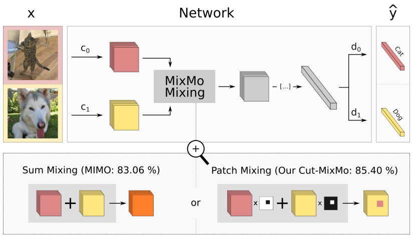

x
Network
ŷ
c_0
c_1
MixMo Mixing
– [...] →
d_0
d_1
Cat
Dog
Sum Mixing (MIMO: 83.06 %)
or
Patch Mixing (Our Cut-MixMo: 85.40 %)
x
x

Figure 4: **MixMo overview**. We embed $M = 2$ inputs into a shared space with convolutional layers (c_1, c_2), mix them, pass the embedding through further layers and output 2 predictions via dense layers (d_1, d_2). The key point of our MixMo (Rame et al., 2021b) is the mixing block. Mixing with patches performs better than basic summing: 85.40% vs. 83.06% (MIMO Havasi et al. (2021)) on CIFAR-100 with WRN-28-10.

MIMO Very recently, the multi-input multi-output MIMO (Havasi et al., 2021) achieves ensemble almost "for free": all of the layers except the first convolutional and last dense layers are shared ($\approx +1\%$ #parameters). The idea is that highly-parameterized CNNs with many inactive filters (Frankle & Carbin, 2019; Molchanov et al., 2017) can fit multiple subnetworks (Veit et al., 2016). The question is how to prevent homogenization among the simultaneously trained subnetworks. To do so, Gao et al. (2019) includes stochastic channel recombination; Durasov et al. (2020) relies on predefined binary masks; in GradAug (Yang et al., 2020), subnetworks only leverage the first channels up to a given percentage. In contrast, MIMO does not need structural differences among subnetworks: rather than treating images one by one, M images are treated simultaneously, as shown on Fig. 4 with $M = 2$. Specifically, they consider M (input, label) pairs at the same time in training. The M inputs are encoded by M separate convolutional layers into a shared latent space before being mixed. The representation is then fed to the core network, which finally branches out into M classifiers. Diverse subnetworks naturally emerge as the i-th classifier learns to classify the label associated with the i-th input ($0 \leq i < M$).

MixMo Our main contribution in MixMo (Rame et al., 2021b) is to shed light on the mixing block - which combines inputs into a shared representation. MIMO (Havasi et al., 2021) just sums those features: we believe this is a missed opportunity. Specifically, we see summing as a balanced and restrictive form of Mixup (Zhang et al., 2018). By analogy, we draw from the considerable mixed sample data augmentation (MSDA) literature to design a more appropriate mixing block. In particular, we leverage binary masking methods to to best tackle the diversity/individual accuracy trade-off in subnetworks. Our framework allows us to create a new Cut-MixMo variant inspired by CutMix (Yun et al., 2019), and illustrated in Fig. 4: a patch of features from the first input is pasted into the features from the second input.

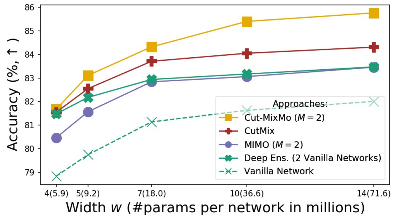


Figure 5: **MixMo results**. CIFAR-100 with WRN-28-w. Our Cut-MixMo variant (patch mixing and $M = 2$) surpasses CutMix and deep ensembles (with half the parameters) by leveraging over-parameterization.

Experiments We (Rame et al., 2021b) consistently perform better than other methods, on CIFAR and TinyImageNet. We notably beat all data augmentation approaches such as CutMix. We also perform better than deep ensembles, but without the inference and memory overhead.

Limits and future work However, MIMO-based strategies perform poorly on more complex datasets such as ImageNet, with smaller architectures or even when dealing with more than $M > 2$ subnetworks. The scaling problem results from the fact that the subnetworks are completely separated in the backbone network, which leads to suboptimal use of parameters. In our current work, we discuss these obstacles preventing feature sharing in MIMO architectures, and we propose solutions to correct this behavior.

3 Multiple datasets: domain generalization and continual learning

3.1 Out-of-distribution generalization

Bias in data The success of deep neural networks in supervised learning (Krizhevsky et al., 2012) relies on the crucial assumption that the train and test data distributions are identical. In particular, the tendency of networks to rely on simple features (Valle-Perez et al., 2019; Geirhos et al., 2020) is generally a desirable behavior reflecting Occam's razor. However, in case of distribution shift, this simplicity bias deteriorates performance when more complex features are needed (Tenenbaum, 2018; Shah et al., 2020). For example, in the recent fight against Covid-19, most of the deep learning methods developed to detect coronavirus from chest scans were shown useless for clinical use (DeGrave et al., 2021; Roberts et al., 2021): indeed, networks exploited simple bias in the training datasets such as patients' age or body position rather than 'truly' analyzing medical pathologies.

Multi domains To better generalize under distribution shifts, we leverage a new notion of diversity, this time in the data rather than in the networks; the training data is divided into different training domains (Blanchard et al., 2011; Muandet et al., 2013). Rather than having multiple networks on a single dataset, we have a single network on multiple datasets. The key assumption is that the causal mechanism is invariant across domains. Under this assumption, the domain is a privileged information (Vapnik & Izmailov, 2015) that may help differentiating between correlation and causation. To remove the domain-dependent explanations, different **invariance criteria across those training domains** have been proposed. Ganin et al. (2016); Sun et al. (2016); Sun & Saenko (2016) enforce similar feature distributions, others (Arjovsky et al., 2019; Krueger et al., 2021) force the classifier to be simultaneously optimal across all domains. Yet, despite the popularity of this research topic, none of these methods perform significantly better than the classical Empirical Risk Minimization (ERM) when applied with controlled model selection and restricted hyperparameter search (Gulrajani & Lopez-Paz, 2021; Ye et al., 2021). These failures motivate the need for new ideas.

3.1.1 Fishr

Invariant gradient variances To foster the emergence of a shared mechanism with consistent generalization properties, our intuition in Fishr (Rame et al., 2021a) is that learning should progress consistently and similarly across domains. Besides, the learning procedure of deep neural networks is dictated by the distribution of the gradients with respect to the network weights (Yin et al., 2018; Sankararaman et al., 2020) — usually backpropagated in the network during gradient descent. Thus, we seek distributional invariance across domains in the gradient space. Our Fishr regularization is summarized in Fig. 6 and **matches the domain-level gradient variances**, *i.e.*, the second moment of the gradient distributions. Our strategy is also motivated by the close relations between the gradient variance, the Fisher Information (Fisher, 1922) and the Hessian. Notably, we show that Fishr forces the model to have similar domain-level Hessians. More generally, Fishr aligns the domain-level loss landscapes and reduces inconsistencies (Parascandolo et al., 2021) across domains.

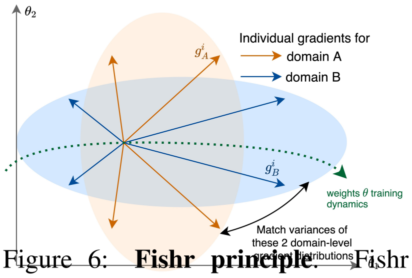

θ_2
Individual gradients for
domain A
domain B
g^i_A
g^i_B
weights θ training dynamics
Match variances of these 2 domain-level gradient distributions
θ_1

Figure 6: **Fishr principle.** Fishr considers the individual (per-sample) gradients of the loss in the network weights θ. Specifically, Fishr matches the domain-level gradient variances of the distributions across the two training domains: A ($\{\boldsymbol{g}^i_A\}_{i=1}^{n_A}$ in orange) and B ($\{\boldsymbol{g}^i_B\}_{i=1}^{n_B}$ in blue). This regularization during the learning of θ improves the out-of-distribution generalization properties by aligning the domain-level loss landscapes at convergence.

Experiments We prove the effectiveness of Fishr on Colored MNIST (Arjovsky et al., 2019). Then, we show that Fishr performs best on the DomainBed benchmark (Gulrajani & Lopez-Paz, 2021) and reaches a new state of the art. Fishr is the only method to systematically and consistently perform better than ERM on all 'real' datasets.

3.1.2 ECML Challenge

By imposing the datasets, the training procedure and controlling the hyperparameter search, DomainBed is arguably the fairer open-source benchmark to rigorously compare the different strategies for OOD generalization. Yet, it only tackles computer vision datasets (with mostly diversity shifts and little correlation shift), imposes the model selection procedure and gives access to the test data by design. To tackle these limits, we are currently organizing an ECML challenge with Criteo named "PRINCE Out-of-distribution Generalization Challenge", available here `https://codalab.lisn.upsaclay.fr/competitions/3353`. We propose a new dataset based on a mix of categorical and continuous features aggregated from user web journeys. To perform well in this challenge, participants will need to both develop models that learn an invariant mechanism across domains, and design an appropriate model selection strategy. We hope this challenge further enriches the understanding of which methods are useful for out-of-distribution generalization.

3.2 Multiple different classes

Datasets with different classes In the previous Section, we had access to different domains sharing the same classes. Now, we tackle the challenge when classes are split across datasets, or more specifically, are added progressively. Due to privacy or memory reasons, the previous data are often no longer inaccessible for training. This is problematic because, if we only finetune the model on the new data, it would catastrophically forgets the old distribution. Continual learning models (Rebuffi et al., 2017) aim at balancing a rigidity/plasticity trade-off where old data are not forgotten (rigidity to changes) while learning new incoming data (plasticity to adapt). Despite recent advances, mostly focused on new distillation losses, it is still an open challenge.

DyTox for continual learning In Douillard et al. (2022), we propose a transformer (Vaswani et al., 2017; Dosovitskiy et al., 2021) architecture based on a dedicated encoder/decoder framework. Critically, the encoder and decoder are shared among all datasets. Through a dynamic expansion of special tokens, we specialize each forward of our decoder network on a task distribution. Our strategy scales to a large number of datasets while having negligible memory and time overheads due to strict control of the parameters expansion. Moreover, this efficient strategy doesn't need any hyperparameter tuning to control the network's expansion. Our model reaches excellent results on CIFAR-100 and state-of-the-art performances on the largescale ImageNet-100 and ImageNet-1000 while having less parameters than concurrent dynamic frameworks (Yan et al., 2021).

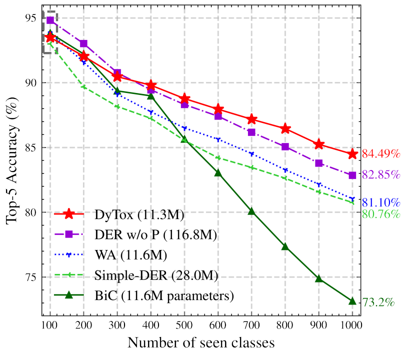


Figure 7: **DyTox's continual learning performance on ImageNet 1000**: for each task, 100 new classes are learned while previously learned classes are not fully accessible but shouldn't be forgotten. Our strategy DyTox (in **red**) is state-of-the-art by a large margin. Note that at the initial step before the continual process begins, our model has performance comparable to other baselines: the performance gain is achieved by reducing catastrophic forgetting. Moreover, we have systematically fewer parameters than previous approaches.

Spurious features in continual learning In the future, we plan to extend this continual scenario to include distribution shifts where some spurious features correlate well with labels within a dataset but not within all datasets (Lesort, 2022).

4 Plan for the second part of the thesis

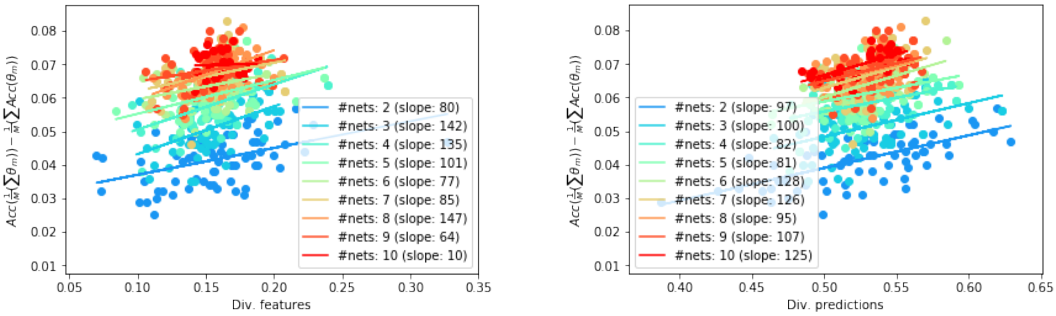
My main goal in the second part of my thesis (*i.e.*, 18 months) is to keep on pushing the state-of-the-art on out-of-distribution (OOD) generalization.

4.1 Current project: ensembling for OOD generalization

Fortunately, I have a serious lead that I am working on and writing currently. The idea follows the work from Cha et al. (2021) — currently state-of-the-art on DomainBed (Gulrajani & Lopez-Paz, 2021) — which leveraged stochastic weight averaging (SWA) (Izmailov et al., 2018) along the training trajectory. They motivated this approach by proving that flatter solution (with low Hessian) should generalize better in OOD. Yet, this does not explain why SWA performs so much better than more complex strategies (Foret et al., 2021) that also flattens the Hessian.

In our research, we try to provide a new explanation of the success of SWA by leveraging its ensembling nature. Indeed, SWA is a first order approximation of the functional ensembling. Then, if we go back to Equation 1 from Section 2.1, we see that growing M can reduce variance. This is important as variances tend to be large in OOD, and cause critical underspecification (D'Amour et al., 2020).

Thus, this shows that ensembling can improve generalization as long as the checkpoints are sufficiently diverse. This theory is validated by our experiments in Figure 8; increased diversity across checkpoints lead to more accurate SWA. While IRM(Arjovsky et al., 2019) or Fishr (Rame et al., 2021a) sought invariance across domains to reduce the bias component, SWA uses the invariance across checkpoints to minimize the variance component. Most importantly, this provides a clear motivation to find and combine diverse checkpoints.



(a) Diversity in features (CKA (Kornblith et al., 2019)).

(b) Diversity in preditions (ratio error (Aksela, 2003)).

Figure 8: Diversity across checkpoints correlates with the performance of SWA, when combining various numbers of checkpoints ($2 \leq M \leq 10$).

To do so, one could think of applying diversity losses (Rame & Cord, 2021) between the current network and its moving average; yet, preliminary experiments were inconclusive. But the best idea seems to combine checkpoints from different runs. This is possible because independently trained SGD solutions (starting from the same initialization) can be connected along one-dimensional paths of near-constant training loss (Benton et al., 2021; Garipov et al., 2018). Our preliminary experiments in Figure 9 suggest that this works even better than ensembling of stochastic weight averages (Arpit et al., 2021). Note that a similar method was applied on ImageNet very recently (Wortsman et al., 2022).

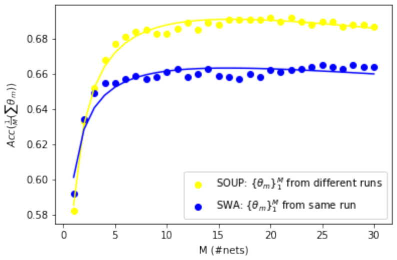


Figure 9: **OOD accuracies on Office-Home**. Averaging weights from multiple runs works better than combining weights from a single run.

4.2 Temporal division

Overall, I plan to split my schedule as follows:

1. Mars - May 2022: ongoing project described in Section 4.1, with a focus on NeurIPS.
2. June - September 2022: organization of the Challenge ECML-KDD https://codalab.lisn.upsaclay.fr/competitions/3353, described in Section 3.1.2, and writing of a follow up paper.
3. November 2022 - January 2023: new project in OOD generalization with hopefully an ICML or CVPR submission. This may be about spurious correlations in continual learning, as introduced in Section 3.2.
4. February - May 2023: new project with hopefully a NeurIPS submission. This may be about (1) features diversity in deep networks, (2) scaling laws for OOD generalization, or (3) the use of transformers to tackle distribution shifts.
5. June - November 2023: writing of the thesis manuscript.
6. December 2023: thesis defense.

References

Matti Aksela. Comparison of classifier selection methods for improving committee performance. In *MCS*, 2003. (page 7).

Alex Alemi, Ian Fischer, Josh Dillon, and Kevin Murphy. Deep variational information bottleneck. In *ICLR*, 2017. (page 3).

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint*, 2019. (pages 1, 5, 7).

Devansh Arpit, Huan Wang, Yingbo Zhou, and Caiming Xiong. Ensemble of averages: Improving model selection and boosting performance in domain generalization. *arXiv preprint*, 2021. (page 7).

Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual information neural estimation. In *ICML*, 2018. (page 3).

Gregory Benton, Wesley Maddox, Sanae Lotfi, and Andrew Gordon Gordon Wilson. Loss surface simplexes for mode connecting volumes and fast ensembling. In *ICML*, 2021. (page 7).

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Generalizing from several related classification tasks to a new unlabeled sample. In *NeurIPS*, 2011. (page 5).

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural networks. In *ICML*, 2015. (page 2).

Leo Breiman. Bagging predictors. *Machine learning*, 1996. (page 2).

Leo Breiman. Random forests. *Machine learning*, 2001. (page 2).

Gavin Brown, Jeremy Wyatt, and Ping Sun. Between two extremes: Examining decompositions of the ensemble objective function. In *MCS*, 2005a. (page 2).

Gavin Brown, Jeremy L Wyatt, and Peter Tiňo. Managing diversity in regression ensembles. *JMLR*, 2005b. (page 2).

Junbum Cha, Sanghyuk Chun, Kyungjae Lee, Han-Cheol Cho, Seunghyun Park, Yunsung Lee, and Sungrae Park. SWAD: Domain generalization by seeking flat minima. In *NeurIPS*, 2021. (page 6).

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *SIGKDD*, 2016. (page 2).

Nadezhda Chirkova, Ekaterina Lobacheva, and Dmitry P. Vetrov. Deep ensembles on a fixed memory budget: One wide network or several thinner ones? *arXiv preprint*, 2020. (page 2).

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *arXiv preprint*, 2020. (page 6).

Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 2021. (page 5).

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. (page 1).

Monroe D Donsker and SR Srinivasa Varadhan. Asymptotic evaluation of certain markov process expectations for large time. *Communications on Pure and Applied Mathematics*, 1975. (page 3).

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. (pages 1, 6).

Arthur Douillard, Alexandre Rame, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *CVPR*, 2022. URL `https://arxiv.org/abs/2111.11326`. (pages 2, 6).

Nikita Durasov, Timur Bagautdinov, Pierre Baque, and Pascal Fua. Masksembles for uncertainty estimation. *arXiv preprint*, 2020. (page 4).

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. 1994. (page 2).

Ian Fischer. The conditional entropy bottleneck. *arXiv preprint*, 2020. (page 3).

Ronald A Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London.*, 1922. (page 5).

Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur. Sharpness-aware minimization for efficiently improving generalization. In *ICLR*, 2021. (page 6).

Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *ICLR*, 2019. (page 4).

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *ICML*, 2016. (page 2).

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 2016. (page 5).

Weihao Gao, Sewoong Oh, and Pramod Viswanath. Demystifying fixed k-nearest neighbor information estimators. *Transactions on Information Theory*, 2018. (page 3).

Yuan Gao, Zixiang Cai, and Lei Yu. Intra-ensemble in neural networks. *arXiv preprint*, 2019. (page 4).

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew G Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *NeurIPS*, 2018. (page 7).

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2020. (page 5).

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. (page 1).

Ishaan Gulrajani and David Lopez-Paz. In search of lost domain generalization. In *ICLR*, 2021. (pages 5, 6).

Lars Kai Hansen and Peter Salamon. Neural network ensembles. *TPAMI*, 1990. (page 1).

Marton Havasi, Rodolphe Jenatton, Stanislav Fort, Jeremiah Liu, Jasper Roland Snoek, Balaji Lakshminarayanan, Andrew Mingbo Dai, and Dustin Tran. Training independent subnetworks for robust prediction. In *ICLR*, 2021. (page 4).

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. (page 3).

David Hin. Stackoverflow vs kaggle: A study of developer discussions about data science. *arXiv preprint*, 2020. (page 1).

Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara Sainath, and Brian Kingsbury. Deep neural networks for acoustic modeling in speech recognition. *Signal Processing Magazine*, 2012. (page 1).

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *ICML*, 2015. (page 1).

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. In *UAI*, 2018. (page 6).

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry Wasserman, et al. Nonparametric von mises estimators for entropies, divergences and mutual informations. In *NeurIPS*, 2015. (page 3).

Ron Kohavi, David H Wolpert, et al. Bias plus variance decomposition for zero-one loss functions. 1996. (page 2).

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In *ICML*, 2019. (page 7).

Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. Estimating mutual information. *Physical review E*, 2004. (page 3).

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, 2009. (pages 1, 3).

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, 2012. (pages 1, 2, 5).

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *NeurIPS*, 1995. (page 2).

David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, 2021. (page 5).

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *NeurIPS*, 2017. (pages 1, 2).

Stefan Lee, Senthil Purushwalkam, Michael Cogswell, David J. Crandall, and Dhruv Batra. Why M heads are better than one: Training a diverse ensemble of deep networks. *arXiv preprint*, 2015. (pages 2, 4).

Timothée Lesort. Continual feature selection: Spurious features in continual learning. *arXiv preprint*, 2022. (page 6).

Yong Liu and Xin Yao. Ensemble learning via negative correlation. *Neural networks*, 1999a. (page 2).

Yong Liu and Xin Yao. Simultaneous training of negatively correlated neural networks in an ensemble. *Cybernetics*, 1999b. (page 2).

Andres R. Masegosa. Learning under model misspecification: Applications to variational and ensemble methods. In *NeurIPS*, 2020. (page 2).

Pavlo Molchanov, Stephen Tyree, Tero Karras, Timo Aila, and Jan Kautz. Pruning convolutional neural networks for resource efficient transfer learning. In *ICLR*, 2017. (page 4).

Krikamol Muandet, David Balduzzi, and Bernhard Schölkopf. Domain generalization via invariant feature representation. In *ICML*, 2013. (page 5).

Sudipto Mukherjee, Himanshu Asnani, and Sreeram Kannan. Ccmi: Classifier based conditional mutual information estimation. In *UAI*, 2020. (page 3).

Nils J. Nilsson. Learning machines: Foundations of trainable pattern-classifying systems. 1965. (pages 1, 2).

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *NeurIPS*, 2019. (page 1).

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *ICML*, 2019. (page 2).

Giambattista Parascandolo, Alexander Neitz, Antonio Orvieto, Luigi Gresele, and Bernhard Schölkopf. Learning explanations that are hard to vary. In *ICLR*, 2021. (page 5).

Alexandre Rame and Matthieu Cord. Dice: Diversity in deep ensembles via conditional redundancy adversarial estimation. In *ICLR*, 2021. URL `https://arxiv.org/abs/2101.05544`. (pages 1, 2, 7).

Alexandre Rame, Corentin Dancette, and Matthieu Cord. Fishr: Invariant gradient variances for out-of-distribution generalization. *arXiv preprint*, 2021a. URL `https://arxiv.org/abs/2109.02934`. (pages 1, 5, 7).

Alexandre Rame, Remy Sun, and Matthieu Cord. Mixmo: Mixing multiple inputs for multiple outputs via deep subnetworks. In *ICCV*, 2021b. URL https://arxiv.org/abs/2103.06132. (pages 1, 4).

Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. icarl: Incremental classifier and representation learning. In *CVPR*, 2017. (pages 1, 6).

Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 2021. (page 5).

Karthik Abinav Sankararaman, Soham De, Zheng Xu, W Ronny Huang, and Tom Goldstein. The impact of neural network overparameterization on gradient confusion and stochastic gradient descent. In *ICML*, 2020. (page 5).

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *ITNN*, 2008. (page 1).

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, 2020. (page 5).

Claude E Shannon. A mathematical theory of communication. *The Bell system technical journal*, 1948. (page 3).

Changjian Shui, Azadeh Sadat Mozafari, Jonathan Marek, Ihsen Hedhli, and Christian Gagné. Diversity regularization in deep ensembles. 2018. (page 2).

Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. Nearest neighbor estimates of entropy. *American journal of mathematical and management sciences*, 2003. (page 3).

Baochen Sun and Kate Saenko. Deep coral: Correlation alignment for deep domain adaptation. In *ECCV*, 2016. (page 5).

Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016. (page 5).

Josh Tenenbaum. Building machines that learn and think like people. In *AAMAS*, 2018. (page 5).

Naftali Tishby. The information bottleneck method. In *Conference on Communication, Control and Computation*, 2001. (page 3).

Naonori Ueda and Ryohei Nakano. Generalization error of ensemble estimators. In *ICNN*, 1996. (page 2).

Guillermo Valle-Perez, Chico Q. Camargo, and Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. In *ICLR*, 2019. (page 5).

Vladimir Vapnik and Rauf Izmailov. Learning using privileged information: Similarity control and knowledge transfer. *JMLR*, 2015. (page 5).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. (pages 1, 6).

Andreas Veit, Michael Wilber, and Serge Belongie. Residual networks behave like ensembles of relatively shallow networks. In *NeurIPS*, 2016. (page 4).

Yeming Wen, Dustin Tran, and Jimmy Ba. Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In *ICLR*, 2019. (page 4).

Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S. Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, and Ludwig Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *arXiv preprint*, 2022. (page 7).

Shipeng Yan, Jiangwei Xie, and Xuming He. Der: Dynamically expandable representation for class incremental learning. In *CVPR*, 2021. (page 6).

Taojiannan Yang, Sijie Zhu, and Chen Chen. Gradaug: A new regularization method for deep neural networks. *NeurIPS*, 2020. (page 4).

Nanyang Ye, Kaican Li, Lanqing Hong, Haoyue Bai, Yiting Chen, Fengwei Zhou, and Zhenguo Li. Oodbench: Benchmarking and understanding out-of-distribution generalization datasets and algorithms. *arXiv preprint*, 2021. (page 5).

R. W. Yeung. A new outlook on shannon's information measures. In *Transactions on Information Theory*, 1991. (page 3).

Dong Yin, Ashwin Pananjady, Max Lam, Dimitris Papailiopoulos, Kannan Ramchandran, and Peter Bartlett. Gradient diversity: a key ingredient for scalable distributed learning. In *AISTATS*, 2018. (page 5).

Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 2019. (pages 1, 4).

Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*, 2018. (page 4).