SWAG

A Simple Baseline for Bayesian Uncertainty in Deep Learning

Wesley Maddox, Timur Garipov, Pavel Izmailov, Dmitry Vetrov, Andrew Gordon Wilson

Matias Cosarinsky

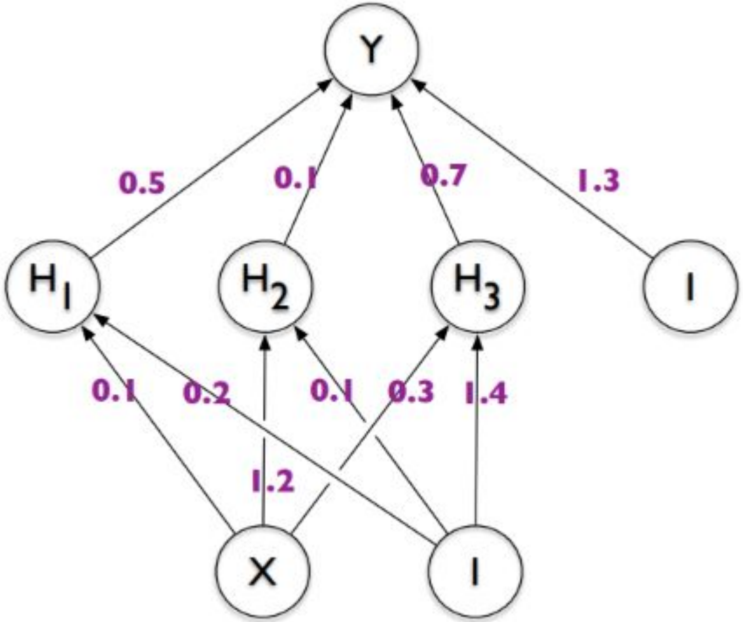
Motivación

- Las redes neuronales se utilizan para toma de decisiones pero tienden a estar mal calibradas

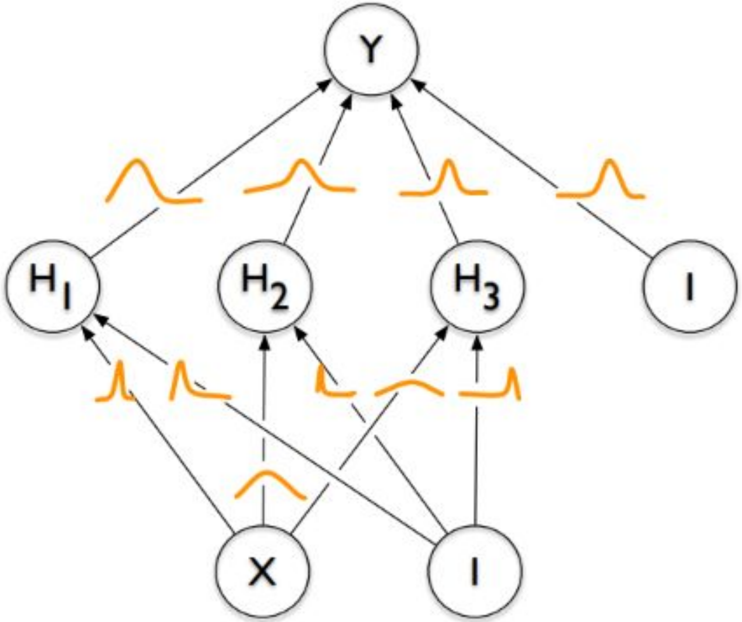
- No cuantifican la incerteza sobre los resultados

- Tienden a sobreajustar y mostrar un nivel de confianza excesivo en sus predicciones

Motivación



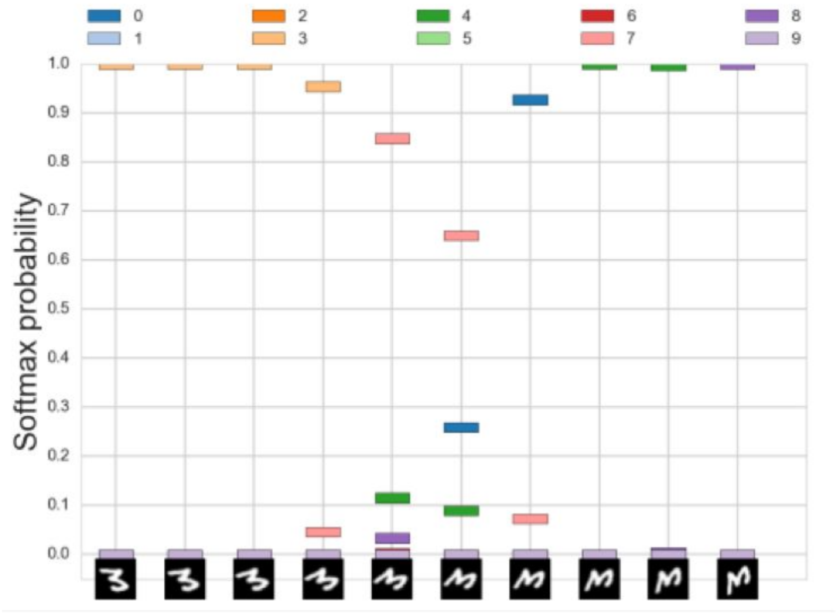
NN tradicional



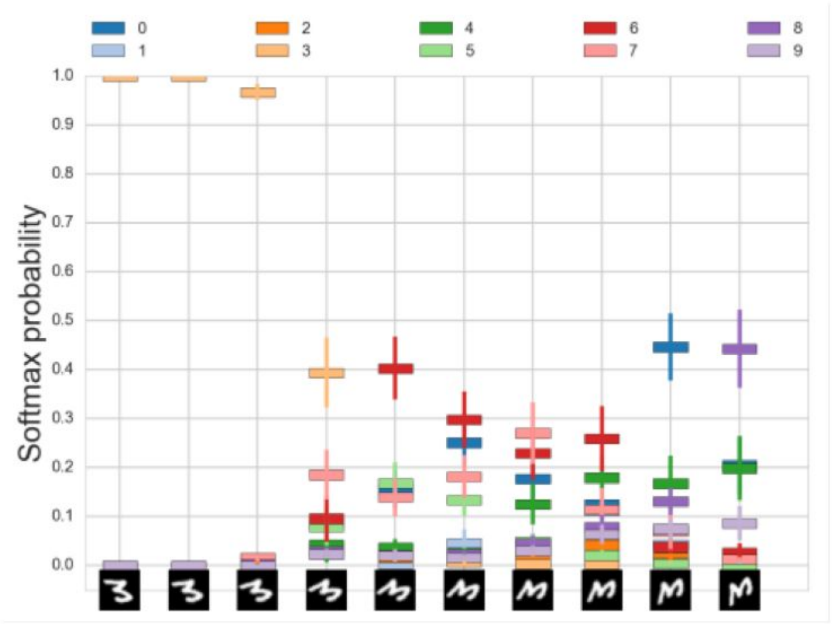
BNN

Imágen de Blundell et al. “Weight Uncertainty in Neural Networks”

Motivación

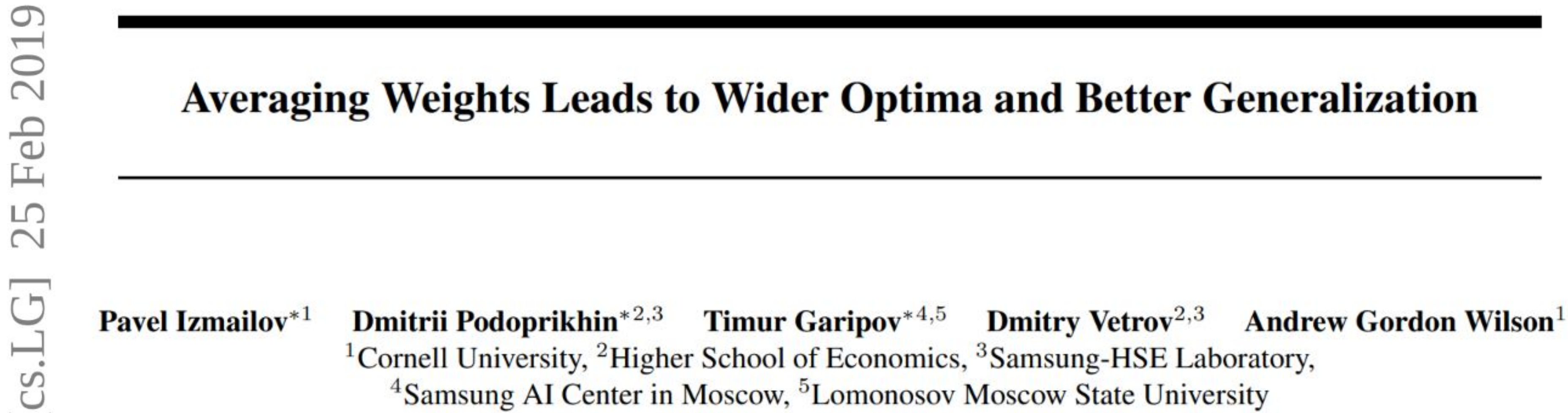


(a) LeNet con weight decay.



(b) LeNet con enfoque bayesiano.

Imágen de Louizos et al. “Multiplicative Normalizing Flows for Variational Bayesian Neural Networks



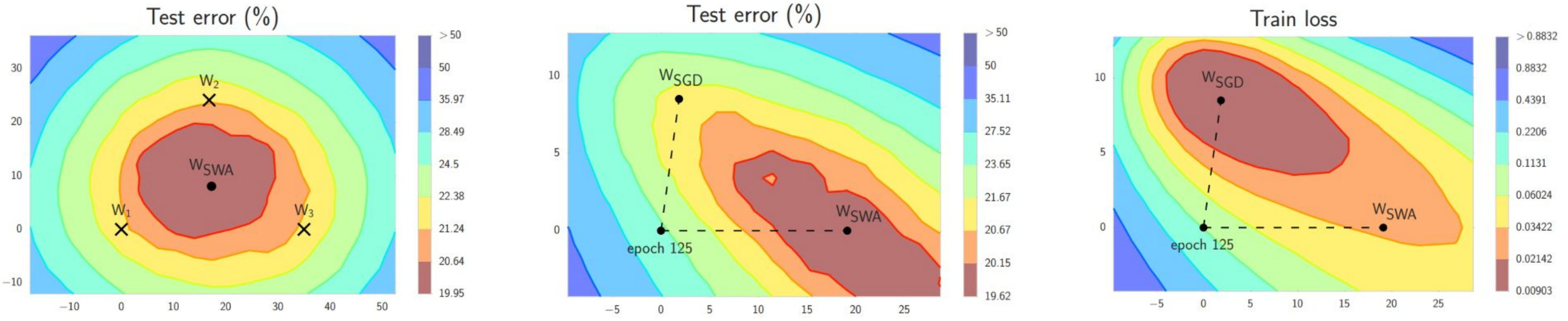
Averaging Weights Leads to Wider Optima and Better Generalization

Pavel Izmailov*[1] Dmitrii Podoprikhin*[2,3] Timur Garipov*[4,5] Dmitry Vetrov[2,3] Andrew Gordon Wilson[1]
[1]Cornell University, [2]Higher School of Economics, [3]Samsung-HSE Laboratory,
[4]Samsung AI Center in Moscow, [5]Lomonosov Moscow State University

SWA: Stochastic Weight Averaging

“Deep neural networks are typically trained by optimizing a loss function with an SGD variant, in conjunction with a decaying learning rate, until convergence. We show that simple averaging of multiple points along the trajectory of SGD, with a cyclical or constant learning rate, leads to better generalization than conventional training”

SWA: Stochastic Weight Averaging



Visualización 2D de la función de pérdida comparando entre SGD y SWA para una ResNet entrenada sobre el dataset CIFAR-100

Imágen de Izmailov et al. “Averaging Weights Leads to Wider Optima and Better Generalization”

Stochastic Gradient Descent as Approximate Bayesian Inference

Stephan Mandt STEPHAN.MANDT@GMAIL.COM
Data Science Institute
Department of Computer Science
Columbia University
New York, NY 10025, USA

Matthew D. Hoffman MATHOFFM@ADOBE.COM
Adobe Research
Adobe Systems Incorporated
601 Townsend Street
San Francisco, CA 94103, USA

David M. Blei DAVID.BLEI@COLUMBIA.EDU
Department of Statistics
Department of Computer Science
Columbia University
New York, NY 10025, USA

Editor: Manfred Opper

Abstract

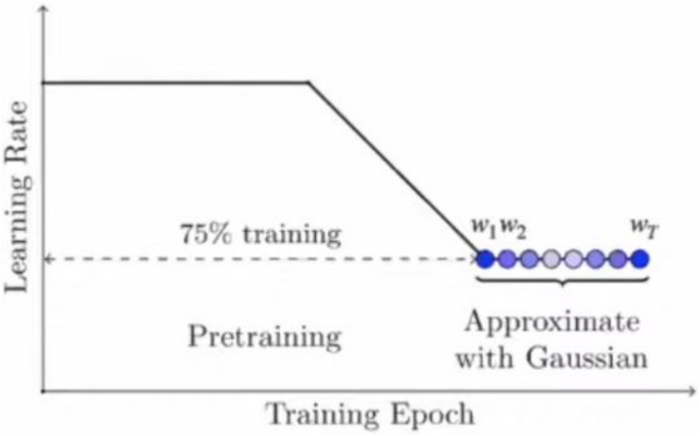
Stochastic Gradient Descent with a constant learning rate (constant SGD) simulates a Markov chain with a stationary distribution. With this perspective, we derive several new results. (1) We show that constant SGD can be used as an approximate Bayesian posterior inference algorithm. Specifically, we show how to adjust the tuning parameters of constant SGD to best match the stationary distribution to a posterior, minimizing the Kullback-Leibler divergence between these two distributions. (2) We demonstrate that constant SGD gives rise to a new variational EM algorithm that optimizes hyperparameters in complex probabilistic models. (3) We also show how to tune SGD with momentum for approximate sampling. (4) We analyze stochastic-gradient MCMC algorithms. For Stochastic-Gradient Langevin Dynamics and Stochastic-Gradient Fisher Scoring, we quantify the approximation errors due to finite learning rates. Finally (5), we use the stochastic process perspective to give a short proof of why Polyak averaging is optimal. Based on this idea, we propose a scalable approximate MCMC algorithm, the Averaged Stochastic Gradient Sampler.

Keywords: approximate Bayesian inference, variational inference, stochastic optimization, stochastic gradient MCMC, stochastic differential equations

SWAG

Idea del método

- SGD con learning rate constante se aproxima a samplear de una Gaussiana
- Usar los dos primeros momentos de SGD para construir una Gaussiana que aproxime la posterior de los pesos
- Samplear de esta distribución para realizar predicciones

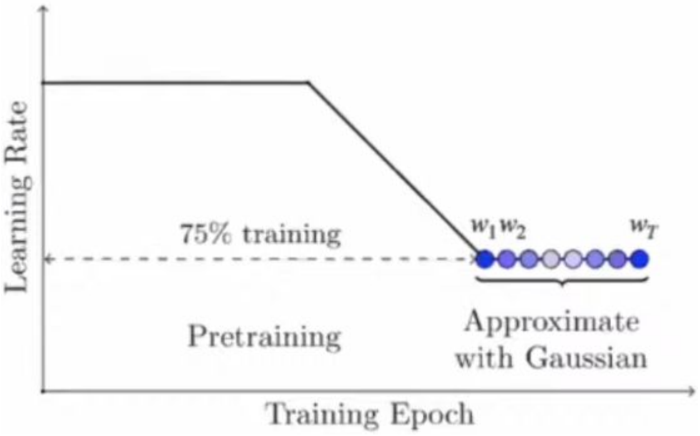


Idea del método

- SGD con learning rate constante se aproxima a samplear de una Gaussiana
- Usar los dos primeros momentos de SGD para construir una Gaussiana que aproxime la posterior de los pesos
- Samplear de esta distribución para realizar predicciones

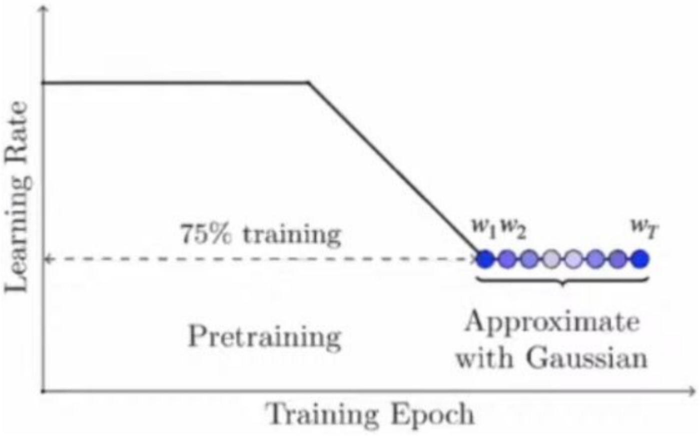
$$\theta_{SWA} = \frac{1}{T} \sum_{i=1}^{T} \theta_i$$

$$\overline{\theta^2} = \frac{1}{T} \sum_{i=1}^{T} \theta_i^2$$



Idea del método

- SGD con learning rate constante se aproxima a samplear de una Gaussiana
- Usar los dos primeros momentos de SGD para construir una Gaussiana que aproxime la posterior de los pesos
- Samplear de esta distribución para realizar predicciones



$$\theta_{SWA} = \frac{1}{T}\sum_{i=1}^{T}\theta_i$$

$$\overline{\theta^2} = \frac{1}{T}\sum_{i=1}^{T}\theta_i^2$$

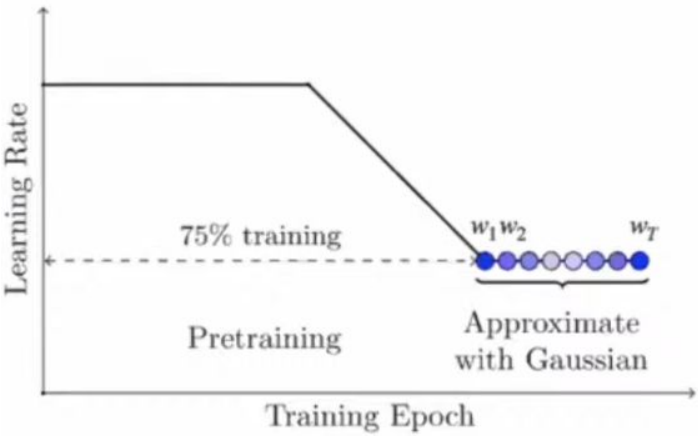
$$\Sigma_{diag} = diag(\overline{\theta^2} - \theta_{SWA}^2)$$

Idea del método

- SGD con learning rate constante se aproxima a samplear de una Gaussiana

- Usar los dos primeros momentos de SGD para construir una Gaussiana que aproxime la posterior de los pesos

- Samplear de esta distribución para realizar predicciones



$$\theta_{SWA} = \frac{1}{T}\sum_{i=1}^{T}\theta_i$$

$$\overline{\theta^2} = \frac{1}{T}\sum_{i=1}^{T}\theta_i^2$$

$$\Sigma_{diag} = diag(\overline{\theta^2} - \theta_{SWA}^2) \quad \mathcal{N}(\theta_{\text{SWA}}, \Sigma_{\text{Diag}})$$

SWAG Diagonal + Low-Rank

- Matriz de covarianza: $$\Sigma = \frac{1}{T}\sum_{i=1}^{T}(\theta_i - \theta_{SWA})(\theta_i - \theta_{SWA})^T$$

SWAG Diagonal + Low-Rank

- Matriz de covarianza: $\Sigma = \frac{1}{T} \sum_{i=1}^{T} (\theta_i - \theta_{SWA})(\theta_i - \theta_{SWA})^T$

- Aproximación: $\Sigma \approx \frac{1}{T-1} \sum_{i=1}^{T} (\theta_i - \bar{\theta}_i)(\theta_i - \bar{\theta}_i)^T = \frac{1}{T-1} DD^T$

SWAG Diagonal + Low-Rank

- Matriz de covarianza: $\Sigma = \frac{1}{T}\sum_{i=1}^{T}(\theta_i - \theta_{SWA})(\theta_i - \theta_{SWA})^T$

- Aproximación: $\Sigma \approx \frac{1}{T-1}\sum_{i=1}^{T}(\theta_i - \overline{\theta}_i)(\theta_i - \overline{\theta}_i)^T = \frac{1}{T-1}DD^T$

- Limitación del rango: $\Sigma_{low-rank} = \frac{1}{K-1}\hat{D}\hat{D}^T$

SWAG Diagonal + Low-Rank

- Matriz de covarianza: $\Sigma = \frac{1}{T}\sum_{i=1}^{T}(\theta_i - \theta_{SWA})(\theta_i - \theta_{SWA})^T$

- Aproximación: $\Sigma \approx \frac{1}{T-1}\sum_{i=1}^{T}(\theta_i - \overline{\theta}_i)(\theta_i - \overline{\theta}_i)^T = \frac{1}{T-1}DD^T$

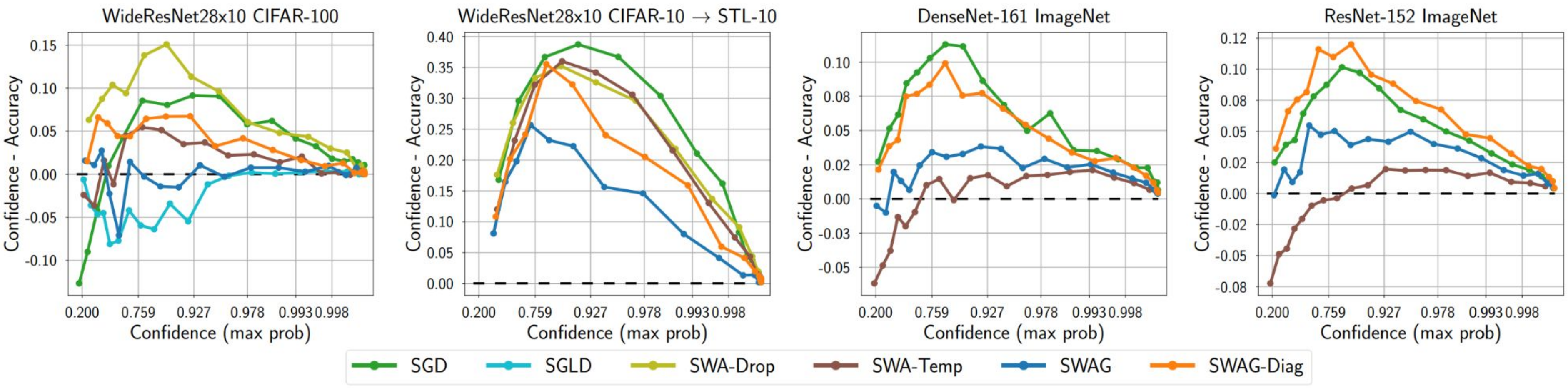
- Limitación del rango: $\Sigma_{low-rank} = \frac{1}{K-1}\hat{D}\hat{D}^T$

- Posterior: $p(\theta|\mathcal{D}) = \mathcal{N}\left(\theta_{SWA}, \frac{1}{2}(\Sigma_{diag} + \Sigma_{low-rank})\right)$

- Sampling: $\tilde{\theta} = \theta_{SWA} + \frac{1}{\sqrt{2}}\Sigma_{diag}^{\frac{1}{2}}z_1 + \frac{1}{\sqrt{2(K-1)}}\hat{D}z_2 \quad z_1 \sim \mathcal{N}(0, I_d) \;\; z_2 \sim \mathcal{N}(0, I_K)$

Aplicaciones

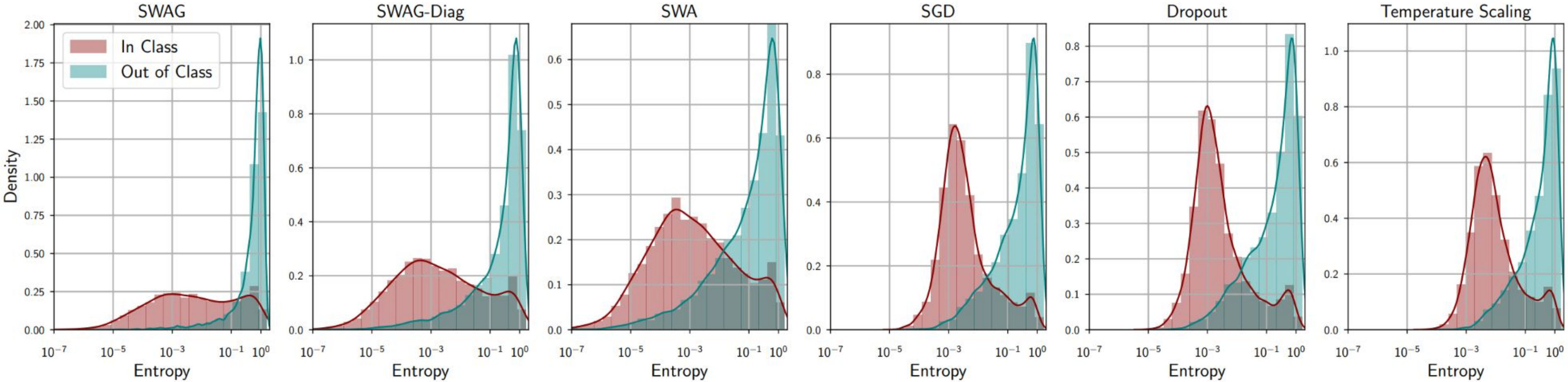
Calibración y Estimación de la Incerteza



Diagramas de fiabilidad para distintos métodos en una tarea de clasificación de imágenes

Imágen de Maddox et al. “A Simple Baseline for Bayesian Uncertainty in Deep Learning”

OOD Image Detection

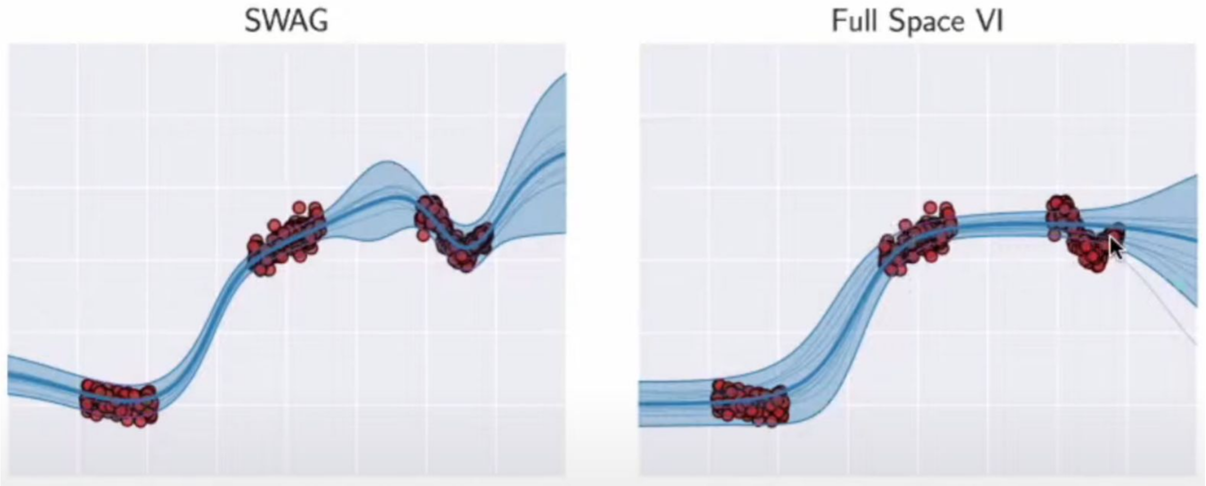


Distribución de la entropía para distintos métodos sobre imágenes in y out of sample

Imágen de Maddox et al. “A Simple Baseline for Bayesian Uncertainty in Deep Learning”

Otras tareas

- Problemas de regresión
- Tareas dentro del área de procesamiento del lenguaje



Ejemplo de visualización de la incerteza en una tarea de regresión comparando SWAG y VI

Referencias

- SWAG
- SWA
- Bayesian Deep Learning - Andrew Gordon Wilson
- Stochastic Gradient Descent as Approximate Bayesian Inference
- Loss Surfaces, Mode Connectivity, and Fast Ensembling of DNNs