Estimacion de la Incertidumbre en Redes Neuronales

Data Science Research Peru - NeurIPS 2020 Meetup

Dr. Matias Valdenegro Toro
matias.valdenegro@dfki.de
@mvaldenegro

Diciembre 2020

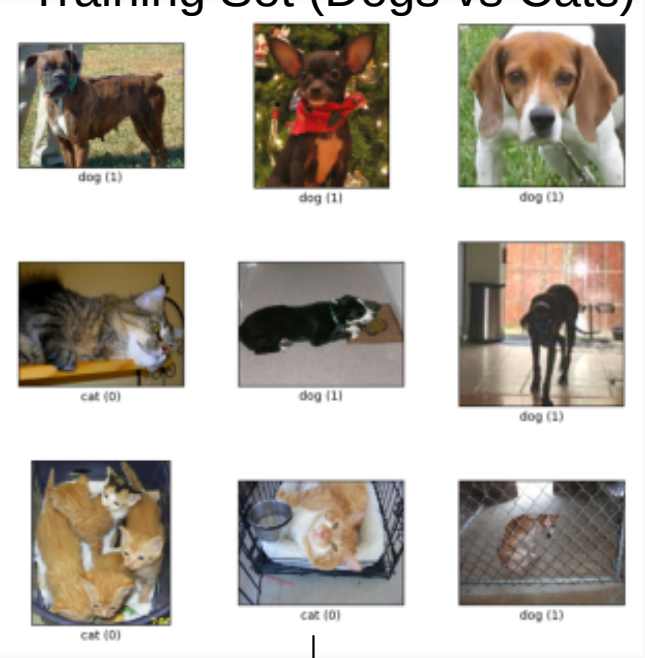
Robotics Innovation Center
German Research Center for Artificial Intelligence
Bremen, Germany

Contenidos

Introduccion a la Estimacion de la Incertidumbre

Que es Incertidumbre en Machine Learning?

Training Set (Dogs vs Cats)



What output probabilities make sense?

Human

Dog and Cat

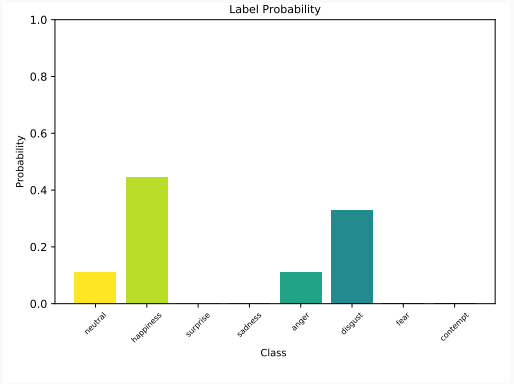
Trained Model

?

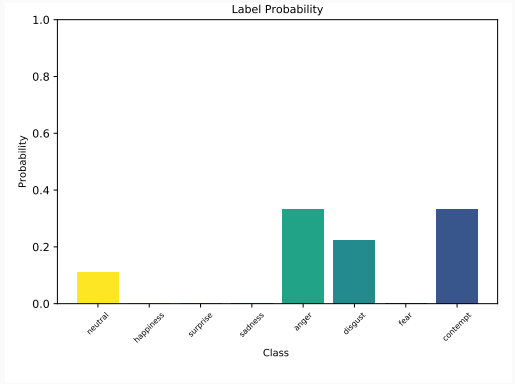
? ?

Que es Incertidumbre en Machine Learning?

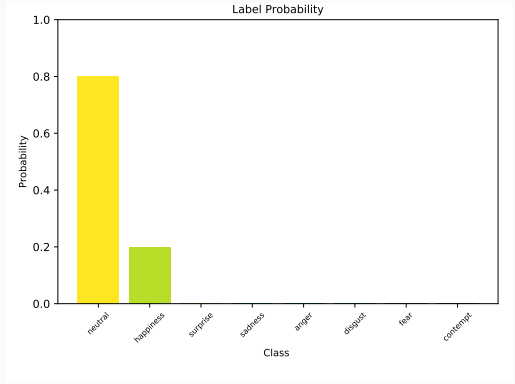
Happiness



Anger



Neutral



Dataset FER+ dataset, con etiquetas çrowd sourced”para reconocimiento de Emociones, sobre clases Neutral, Felicidad, Sorpresa, Tristeza, Enojo, Disgusto, Miedo, y Desprecio.

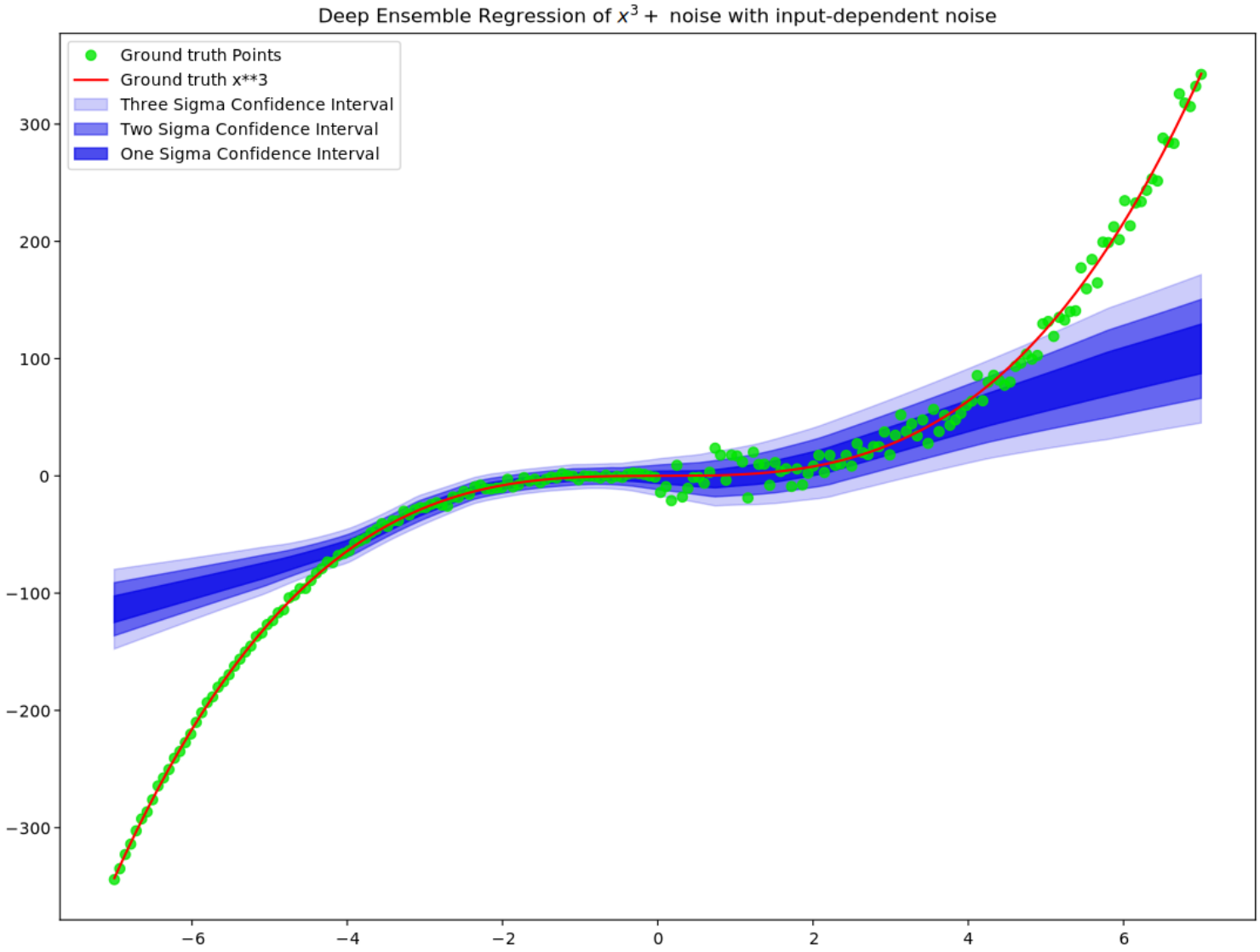
Que es Incertidumbre en Machine Learning?

- Los datasets del mundo real suelen estar desbalanceados, por lo que las confidencias de cada clase deberían ser diferentes, reflejando los datos y las capacidades del modelo.
- Datasets del mundo real pueden contener ruido, como etiquetas imprecisas, medidas ambiguas, o ruido del sensor. Un modelo debe considerar esto.
- La mayoría de las redes neuronales son demasiado confiadas, lo que significa que las confidencias de softmax no tienen una buena interpretación probabilística y podrían ser engañosas.

¿Qué les falta a los modelos clásicos?

- La mayoría de los modelos de machine learning no modelan explícitamente la incertidumbre en sus salidas.
- Ellos producen predicciones `puntuales`. Un modelo con incertidumbre produce una `distribucion de probabilidad` como salida.
- Una distribucion de probabilidad puede incluir mas informacion que una prediccion puntual, por ejemplo, media y varianza para una salida de regresion, en lugar de un valor puntual.
- Las redes neuronales suelen tener demasiada confianza (overconfident) y producen predicciones erróneas con mucha confidencia.

¿Qué les falta a los modelos clásicos?



Implicaciones Practicas de Confidencias

- Muchos profesionales prestan mucha atención al desempeño de la tarea (precisión/accuracy, error cuadrático, etc.), pero ignoran las confidencias que produce un modelo.
- Algunas aplicaciones utilizan los valores de confidencia de un modelo sin validarlos.
- Las confidencias tiene una relación con la interpretabilidad, ya que a menudo los ejemplos mal clasificados con alta incertidumbre pueden estar mal etiquetados o ser demasiado similares a otros ejemplos.

Aplicaciones Practicas de la Incertidumbre

- Se pueden usar estimaciones robustas de las confianzas para detectar ejemplos mal clasificados o cuando el modelo está extrapolando.
- Un modelo puede rechazar producir una salida si la incertidumbre es demasiado alta, por ejemplo, para requerir procesamiento humano en lugar de automatizado. Esto se denomina detección fuera de distribución (Out of distribution detection).
- La confidencia o incertidumbre de una predicción le dice al ser humano cuánto debe realmente confiar en la predicción.
- Se pueden tomar decisiones adicionales con una puntuación de confianza realista, que es muy importante para aplicaciones médicas y de interacción humana.

Tipos de Incertidumbre

Incertidumbre Aleatorica

Incertidumbre inherente a los datos, por ejemplo, ruido de sensor, procesos estocásticos.
No se puede reducir agregando más información.

Incertidumbre Epistemica

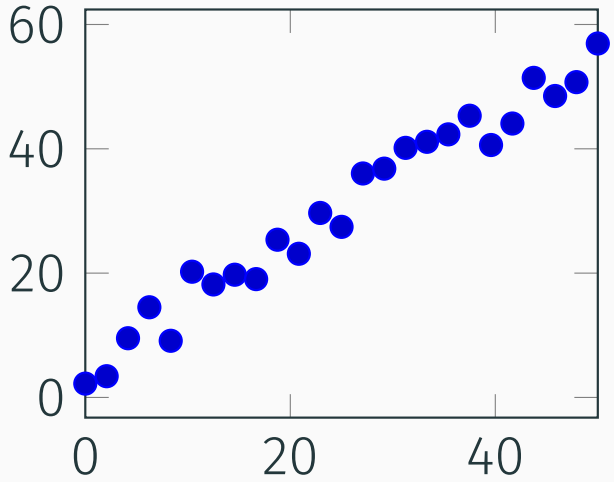
Incertidumbre producida por el modelo, por ejemplo, especificación incorrecta del modelo, desbalance de clases, falta de datos de entrenamiento.
Puede reducirse agregando más información al proceso de entrenamiento.

Incertidumbre Aleatorica

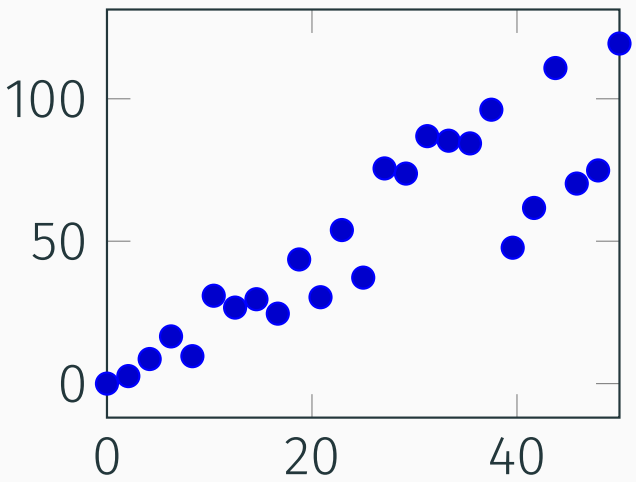
El ejemplo mas simple de IA son mediciones corrompidas por ruido aditivo, como $f(x) = x^3 + \epsilon$ Where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ donde x^3 es la funcion correcta.

Si σ^2 es constante, esto se llama ruido homoscedatico, si σ^2 es function de la entrada o una variable, entonces se denomina ruido heteroscedatico.

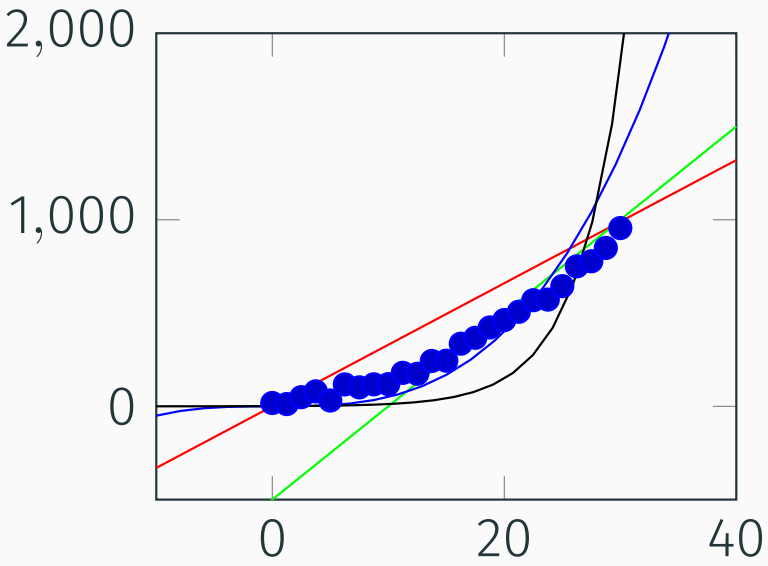
Homoscedatico



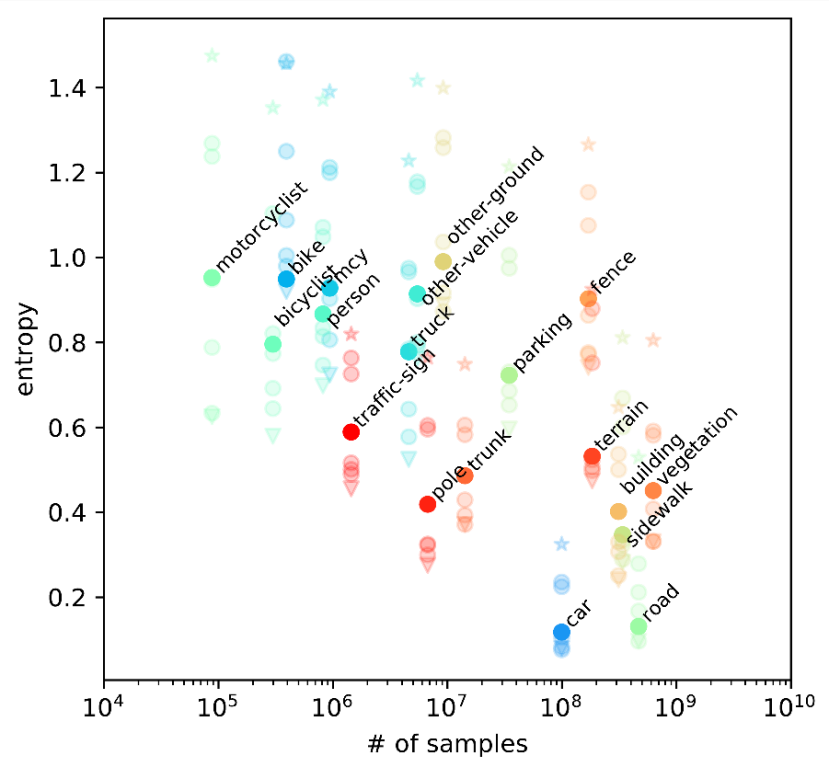
Heteroscedatico



Incertidumbre Epistemica



Especificacion incorrecta del modelo (Model Misspecification)



Variacion del numero de ejemplos en el set de entrenamiento

Formulacion Bayesiana

Una red neuronal bayesiana es aquella en la que los pesos (weights) son distribuciones de probabilidad, en lugar de estimaciones puntuales. Las distribuciones de peso codifican implícitamente la incertidumbre en la red.

Esto requiere algoritmos de inferencia radicalmente diferentes para aprender estas distribuciones a partir de los datos. La distribucion predictiva bayesiana para y desde entradas x y distribuciones de pesos θ es:

$$p(y \mid x) = \int_{\Theta} p(y \mid x, \theta) P(\theta \mid x) d\theta$$

Esto se denomina promediado del modelo bayesiano, ya que los pesos se muestrean a partir de las distribuciones de pesos aprendidas y se utilizan para producir estimaciones de salida, ponderadas por la probabilidad de cada peso. Esto hace que la estimación de la posterior completa sea computacionalmente muy costosa, por lo que rara vez se utiliza en la práctica.

Metodos para Estimacion de la Incertidumbre

Bayes by Backpropagation

- Una simplificacion comun es aproximar la distribucion real de pesos con una distribucion Gausiana.
- Con algunos trucos y Inferencia Variacional, se pueden estimar gradientes con respecto a la media μ y varianza σ^2 de la Gausiana que representa los pesos de una capa.
- Produce muy buenas estimaciones de incertidumbre epistemica, pero el problema es que generalmente falla en converger con redes minimamente complejas (los gradientes explotan).

Monte Carlo Dropout

- Dropout es una tecnica conocida para regularizar redes neuronales.
- Durante el entrenamiento, una mascara $m_i \sim \text{Bernoulli}(p)$ es muestreada y multiplicada con las activaciones de entrada, efectivamente haciendo que algunas se conviertan en ceros.
- Dropout también se puede habilitar en el momento de la inferencia/test, donde se ha comprobado que funciona como una aproximación de la distribución posterior predictiva.

Monte Carlo DropConnect

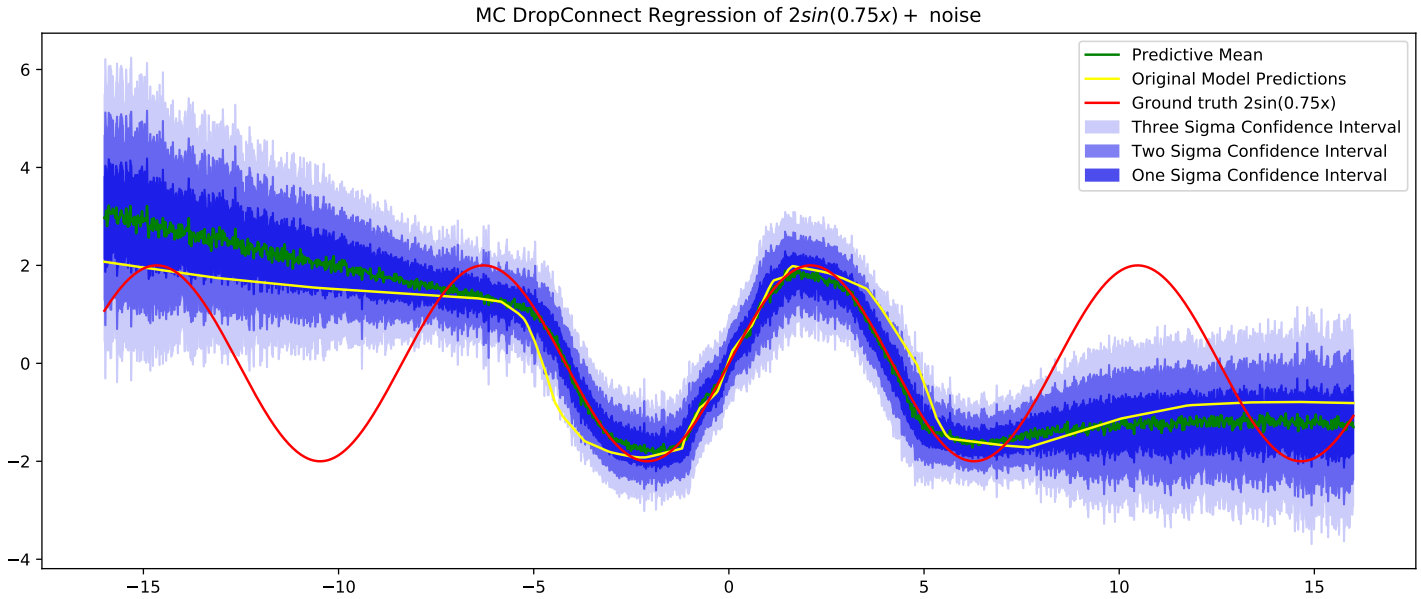
- DropConnect es una variacion de Dropout, donde la mascara se aplica a los `pesos` de una capa en lugar de aplicarse a las activaciones.
- Ha sido provado que tambien produce una aproximacion de la distribucion posterior predictiva. Requiere la implementacion de nuevas capas que usan DropConnect internamente.
- En algunos casos supera a MC Dropout tanto en metricas de tarea como en desempeño de incertidumbre, pero no siempre.

Que pasa con Monte Carlo?

- Habilitar Dropout o DropConnect en la inferencia transforma la red neuronal en un modelo estocástico.
- Esto significa que cada pasada hacia adelante (forward pass) produce un resultado diferente, correspondiente a una muestra de la distribución posterior predictiva.
- El modelo con incertidumbre se puede evaluar combinando las predicciones de M pasadas hacia adelante.
- Esta es la versión de Monte Carlo de la distribución posterior predictiva:

$$p(y \mid x) \sim M^{-1} \sum_{i}^{M} p(y \mid x, \theta_i) \quad \text{where} \quad \theta_i \sim \Theta$$

MC-DropConnect - Regresion de una Sinusoide



Ensembles

- Ensembles tambien tienen poderosas propiedades de estimacion de la incertidumbre.
- Ensamblado consiste en entrenar M instancias del mismo modelo, pero con diferentes pesos iniciales al azar, y luego combinar sus predicciones.
- Para regresion, cada miembro del ensemble tiene dos çabezas"de salida, una para la media $\mu_i(x)$ y otra para la varianza $\sigma_i^2(x)$, y una loss special se usa para el entrenamiento:

$$-\log p(y_n|\mathbf{x}_n) = \frac{\log \sigma_i^2(\mathbf{x}_n)}{2} + \frac{(\mu_i(\mathbf{x}_n) - y_n)^2}{2\sigma_i^2(\mathbf{x}_n)} + C$$

- Esta loss es una funcion de log-verosimilitud negativa (negative log-likelihood) con varianza heretoscedatica, el modelo predice una varianza para cada entrada.

Ensembles - Combinacion

Donde p_i is la salida del i-esimo miembro del ensemble.

Clasificacion

La salida del Ensemble es la media de las probabilidades:

$$p_e(y \mid \mathbf{x}) = M^{-1} \sum_i p_i(y \mid \mathbf{x})$$

Regresion

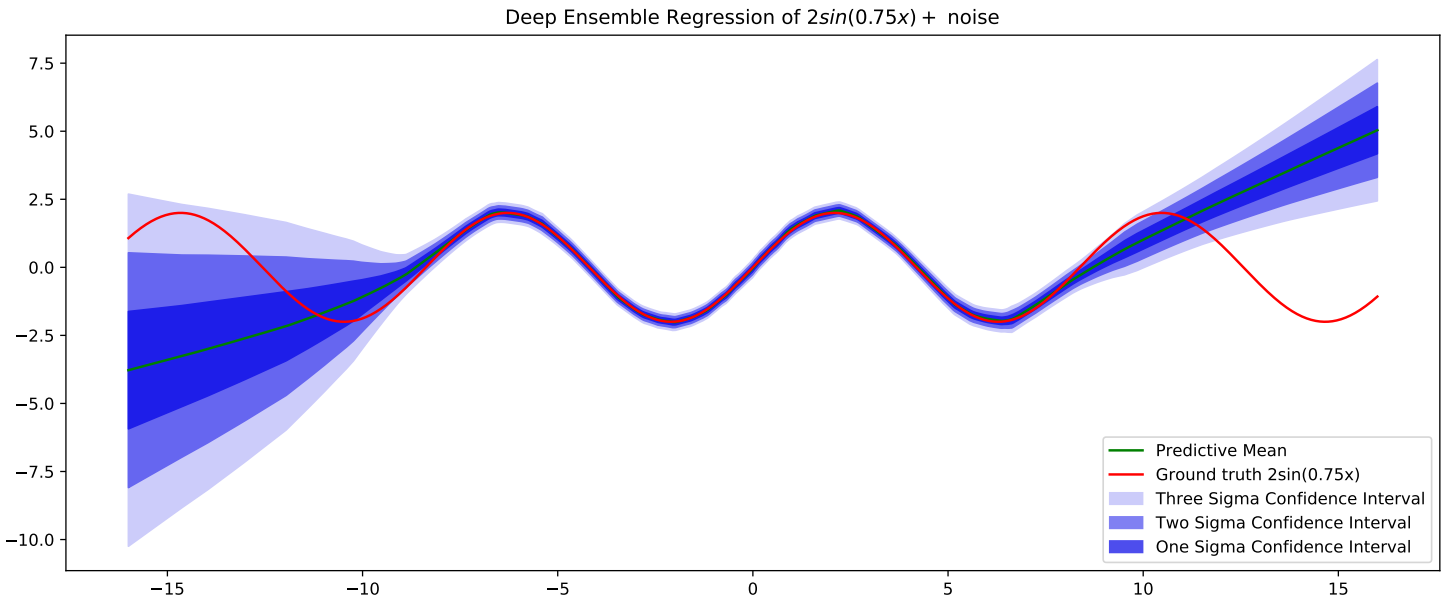
La salida del Ensemble es un modelo de mezcla gaussiano (Gaussian mixture model):

$$p_e(y \mid \mathbf{x}) \sim \mathcal{N}(\mu_*(\mathbf{x}), \sigma_*^2(\mathbf{x}))$$

$$\mu_*(\mathbf{x}) = M^{-1} \sum_i \mu_i(\mathbf{x})$$

$$\sigma_*^2(\mathbf{x}) = M^{-1} \sum_i (\sigma_i^2(\mathbf{x}) + \mu_i^2(\mathbf{x})) - \mu_*^2(\mathbf{x})$$

Ensembles - Regresion de una Sinusoide



Otros Metodos

- Diferentes formas de inferencia variacional, generalmente cambiando la representacion y como se estima la distribucion sobre los pesos.
- Regresion de Cuantiles (Quantile Regression).
- Ecuaciones Diferenciales Neuronales (Neural ODEs).
- Redes RBF (Radial Basis Function).
- Test-time Augmentation.

Evaluacion de la Incertidumbre

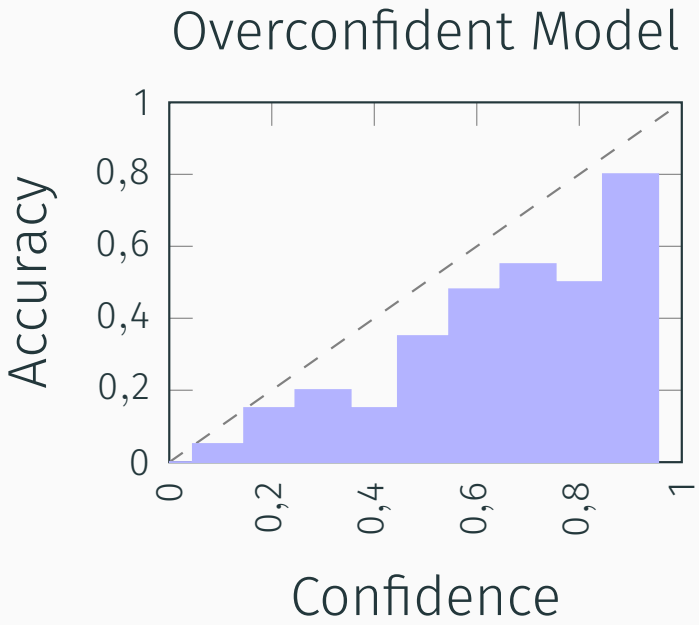
Calibración

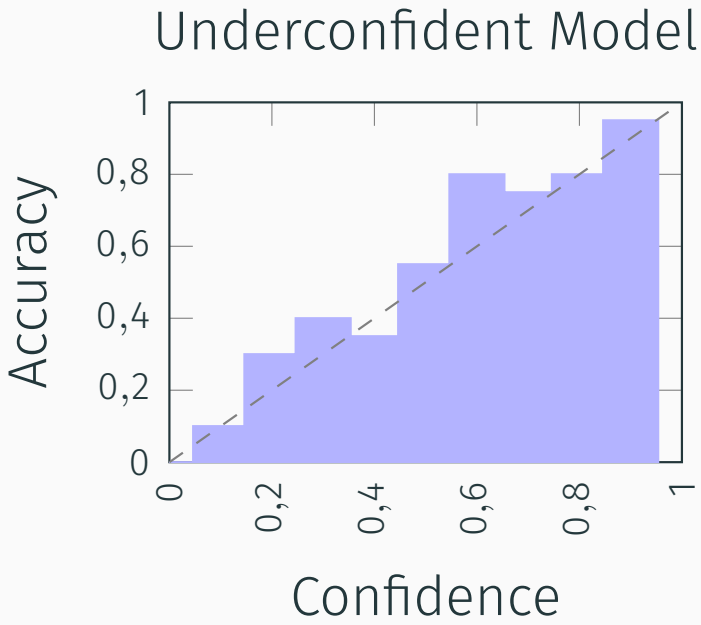
- Hablamos de un concepto que indica cuánto podemos confiar en las confidencias producidas por un modelo.
- Esto se puede formalizar comparando el desempeño de la tarea (como la precisión/accuracy) a medida que cambia la confidencia de las predicciones.
- Por ejemplo, si una prediccion es hecha con 10 % de confidencia, entonces esperamos que esas predicciones seras correctas un 10 % del tiempo.
- Y correspondientemente, si una prediccion es hecha con 90 % de confidencia, entonces solo 10 % de esas predicciones seran incorrectas.

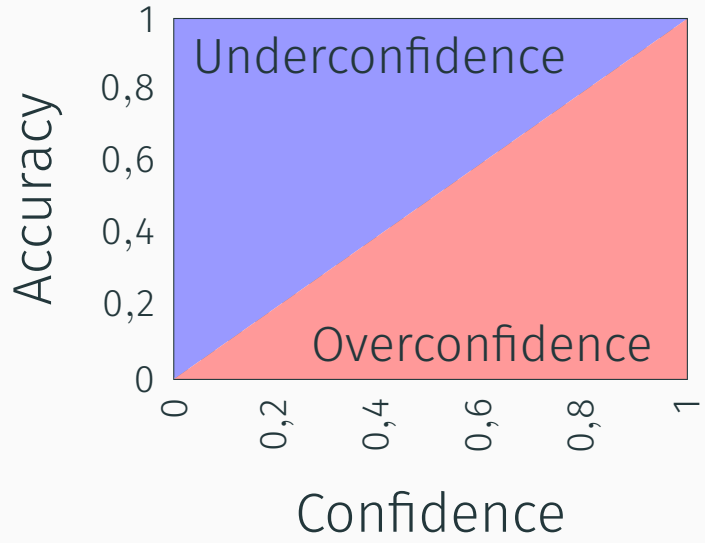
Calibración - Grafico de Reliabilidad

- La calibración se puede observar haciendo un gráfico de confiabilidad.
- Tomamos las predicciones de un modelo sobre un dataset, dividimos las predicciones por valores de confidencia $\text{conf}(B_i)$ en bins B_i, por cada bin se calcula la precision (accuracy) $\text{acc}(B_i)$, y luego se grafican los valores $(\text{conf}(B_i), \text{acc}(B_i))$.
- Regiones donde $\text{conf}(B_i) < \text{acc}(B_i)$ indican que el modelo es subconfidente (underconfident), mientras que las regiones $\text{conf}(B_i) > \text{acc}(B_i)$ indican sobreconfidencia (overconfidence).
- La linea $\text{conf}(B_i) = \text{acc}(B_i)$ indica calibracion perfecta.

Calibración - Grafico de Reliabilidad







Calibración - Metricas

Error de Calibración

$$\mathrm{CE} = \sum_i |\mathrm{acc}(B_i) - \mathrm{conf}(B_i)|$$

Error de Calibración Esperado (Expected Calibration Error)

$$\mathrm{ECE} = \sum_i N^{-1} |B_i| \, |\mathrm{acc}(B_i) - \mathrm{conf}(B_i)|$$

Maximo Error de Calibración (Maximum Calibration Error)

$$\mathrm{MCE} = \max_i |\mathrm{acc}(B_i) - \mathrm{conf}(B_i)|$$

Calibración en Regresion

- El concepto de calibracion tambien se puede extender a problemas de regresion.
- Para esto se usan intervalos de confianza, se define un intervalo de confianza $1 - \alpha$ y se estima una precision (accuracy) considerando que porcentaje de los intervalos producidos sobre un dataset contienen el valor real.
- Se varia el valor de α (que funciona como confidencia) y con esto se puede producir un grafico de reliabilidad de forma similar, tambien usando bins de histograma.

Calibracion - Grafico de Reliabilidad con MC-Dropout en MNIST

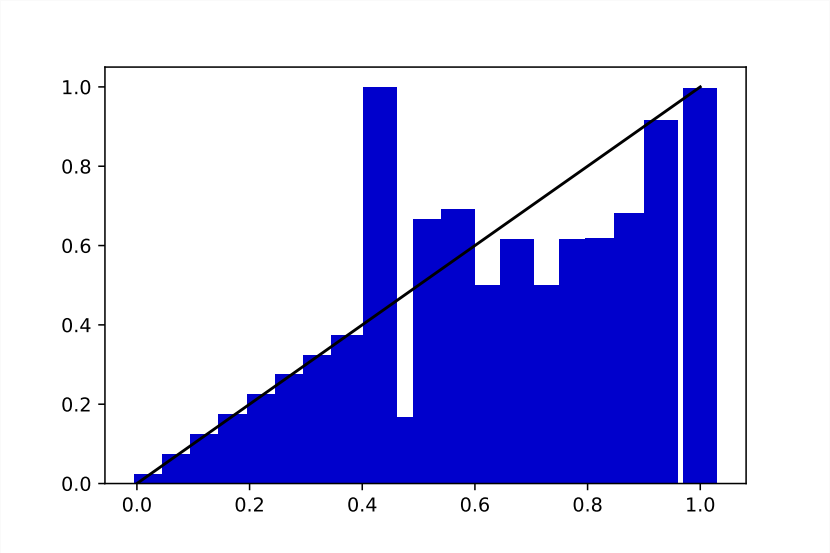


Figura 1: Red Neuronal Clasica, Error de Calibration es 0.18

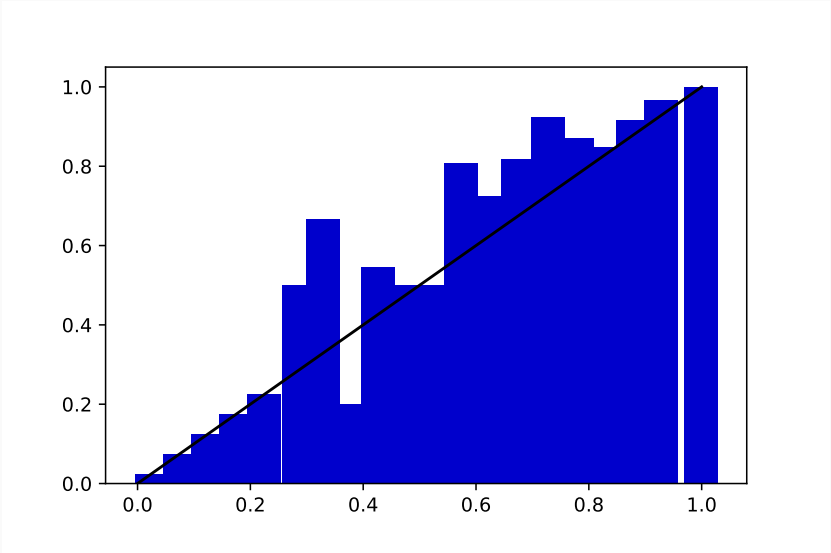


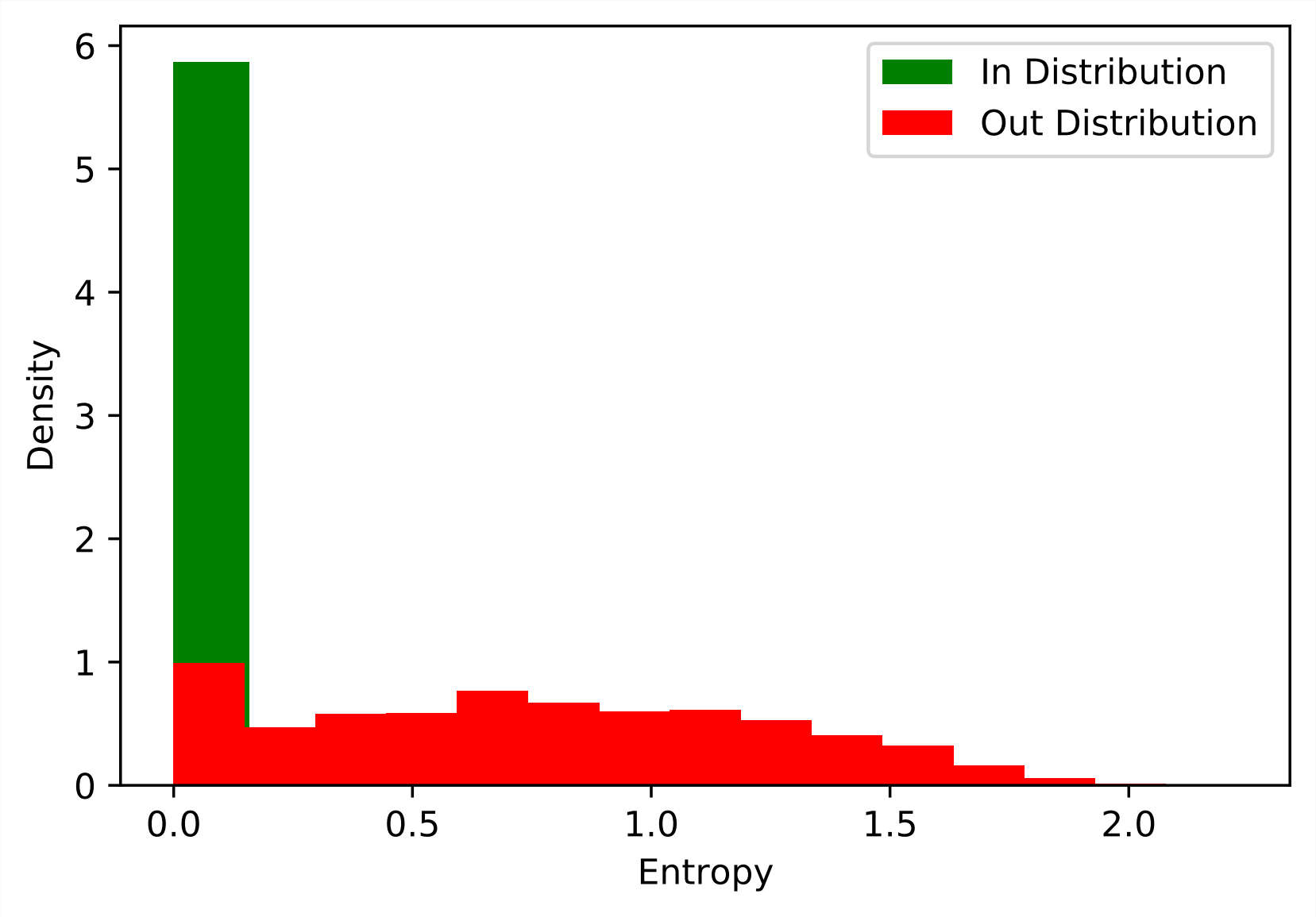
Figura 2: Red Neuronal Bayesiana con MC-Dropout, Error de Calibration es 0.11

Detección fuera de distribución (Out of Distribution Detection, OOD)

- Es la tarea de detectar cuándo la entrada al modelo está fuera de la distribución del dataset de entrenamiento utilizado para entrenar el modelo.
- Esto corresponde al rechazo del modelo de proporcionar una salida si no está seguro de ello.
- Hacer esto es simple, rechace considerar la salida de un modelo si la incertidumbre es demasiado grande. El truco consiste en seleccionar un umbral/threshold apropiado.
- Para la regresión, se puede utilizar la desviación estándar de la salida. Para la clasificación, se prefiere la entropía H:

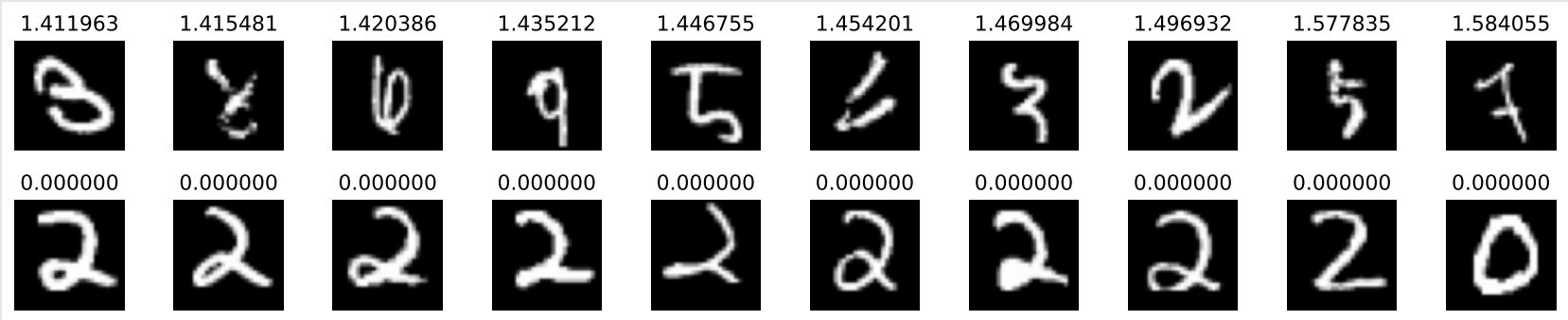
$$H(p(x)) = \sum_i p(x)_i \log p(x)_i$$

Detección fuera de distribución (OOD) - MNIST vs Fashion MNIST

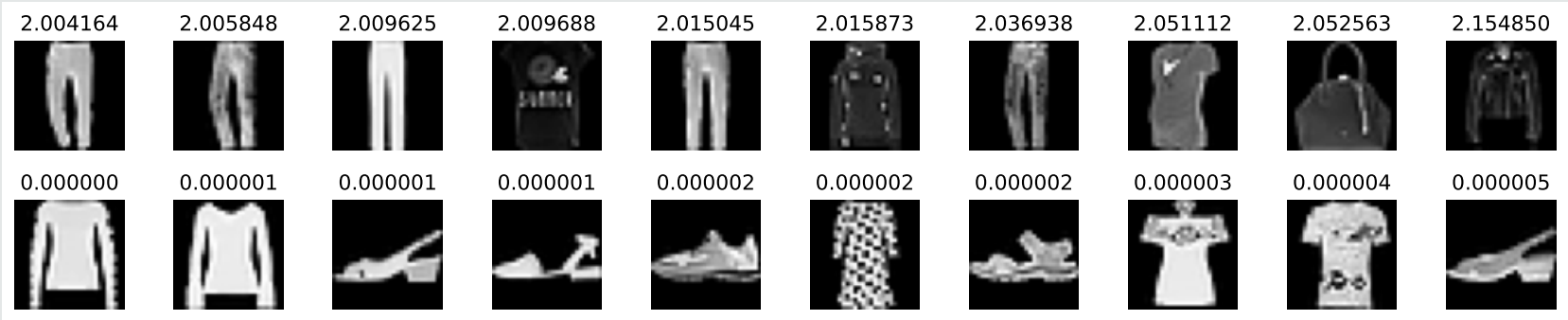


Detección fuera de distribución (OOD) - MNIST vs Fashion MNIST

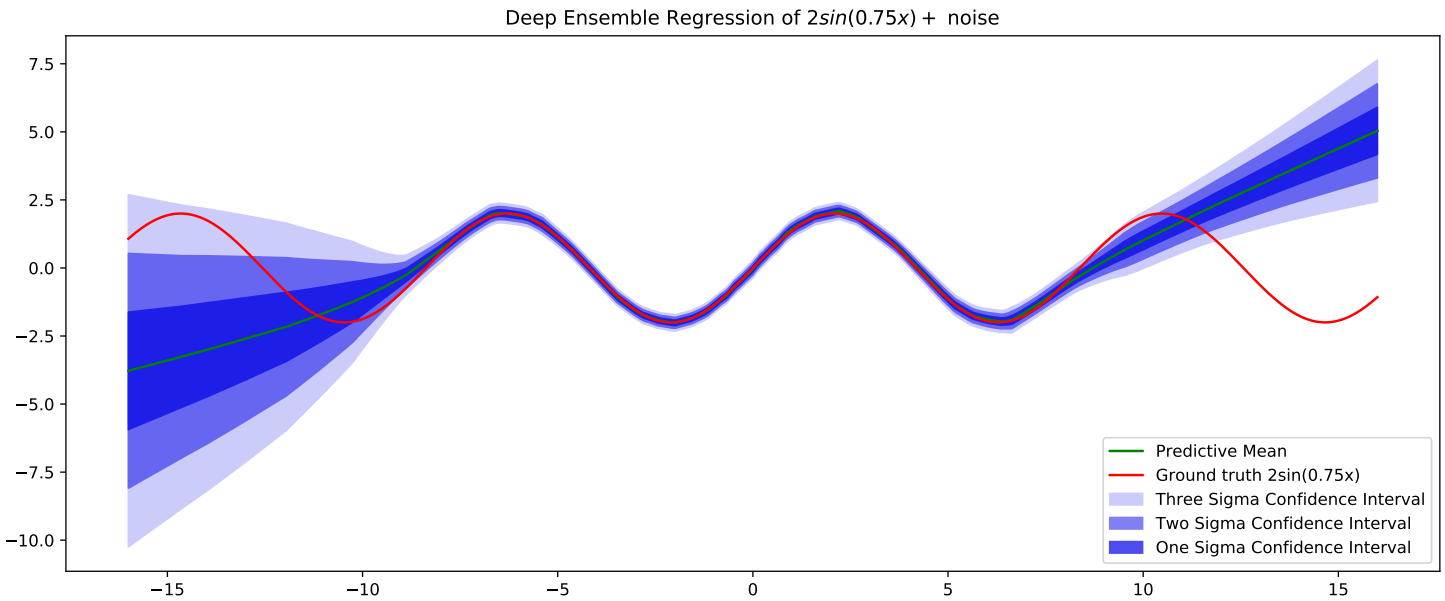
MNIST



Fashion MNIST



Detección fuera de distribución (OOD) - Regresion de una Sinusoide con Ensembles



En este ejemplo, el set de entrenamiento es $x \in [-8, 8]$, Se puede observar visualmente que fuera de este rango la desviación estándar de la salida (incertidumbre) aumenta considerablemente, y aumenta con la distancia a ese rango.

Detección fuera de distribución (OOD) - Dificultades

- No es fácil separar completamente los ejemplos de ID y OOD, ya que algunos ejemplos ID todavía tienen una alta incertidumbre y, a veces, los ejemplos de OOD tienen poca incertidumbre. Esto se debe a variabilidad en las clases.
- Elegir un umbral no es fácil, ya que se deben realizar mucho análisis.
- Desafortunadamente, no hay garantías sobre el rendimiento de OOD y se conocen casos de efectos negativos. (Ver Ovadia et al.)
- La incertidumbre debe utilizarse como información adicional a partir de la cual se puede decidir el análisis humano adicional, en lugar de permitir un procesamiento completamente automático.

Mi Investigacion en Incertidumbre

Sub-Ensembles

- Un gran problema con usar Ensembles es que aumenta el costo computacional por un factor igual al numero de miembros en el ensemble.
- Una pregunta basica es si es necesario que todos los miembros del ensemble sean independientes, usando la idea de ”weight sharing”.
- Resulta que no es necesario que todas las capas del modelo participen en el ensemble, se puede compartir una serie de capas (desde la entrada) y hacer ensemble de un cierto numero de capas (desde la salida de la red), y esto funciona como una aproximacion al ensemble completo.

Sub-Ensembles

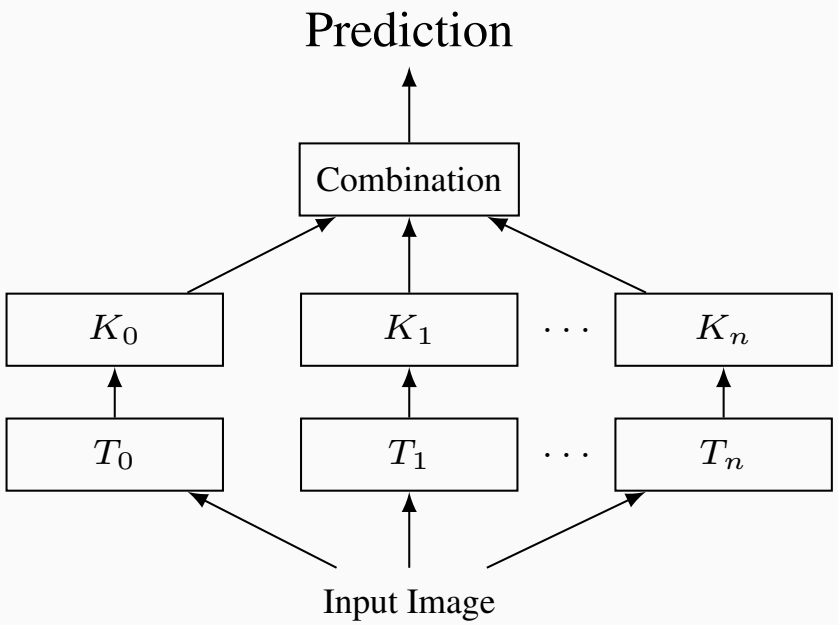


Figura 3: Ensemble

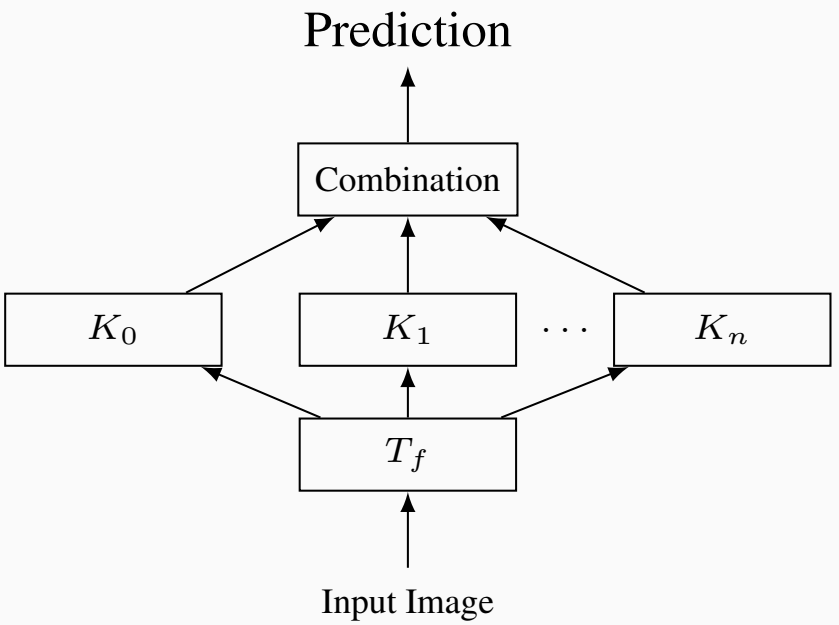
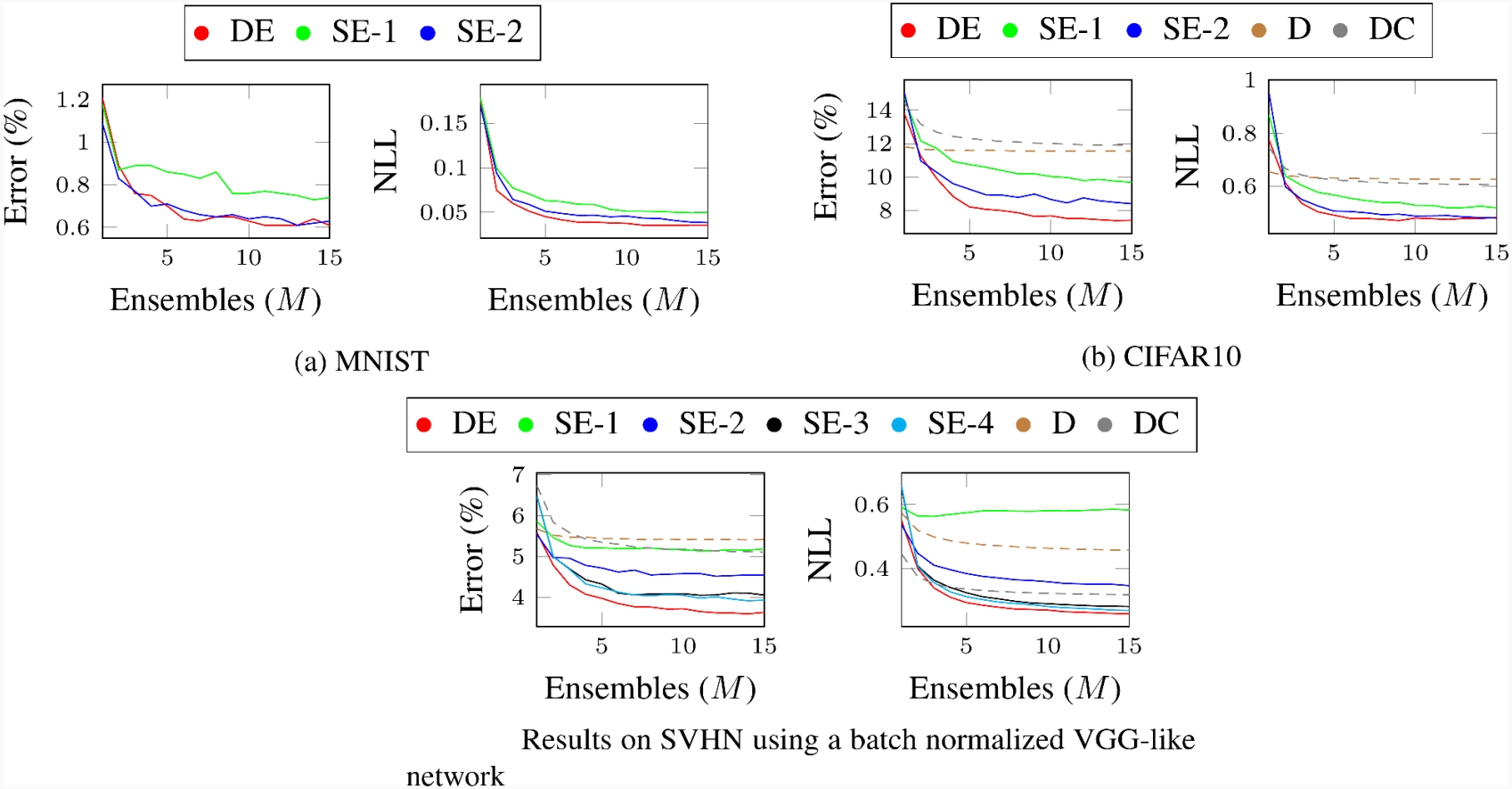


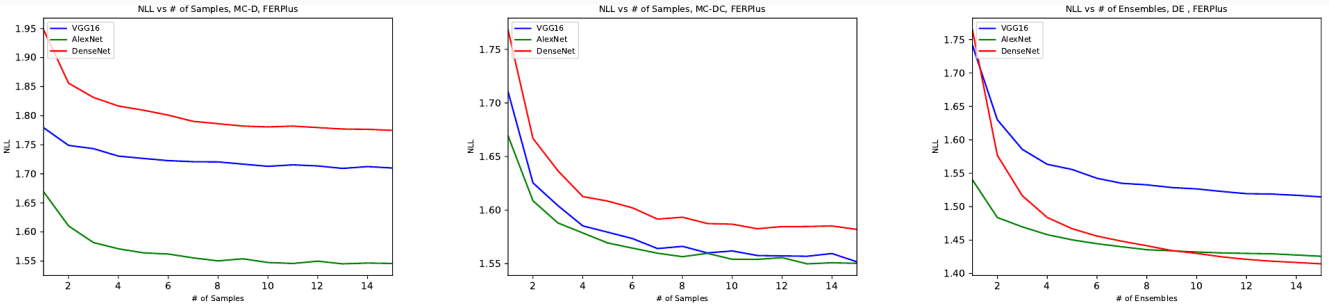
Figura 4: Sub-Ensemble

Sub-Ensembles - Rendimiento



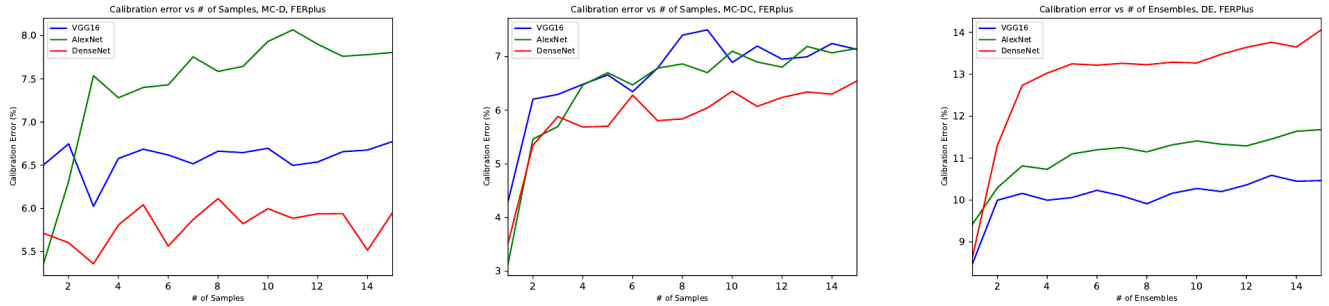
Paper fue presentado en el Bayesian Deep Learning Workshop @ NeurIPS 2019.

Incertidumbre en Clasificacion de Emociones



(a) MC-D (b) MC-DC (c) DE

Fig. 2: NLL as a function of # of samples/ensembles for all methods in different models on FERPlus dataset



(a) MC-D (b) MC-DC (c) DE

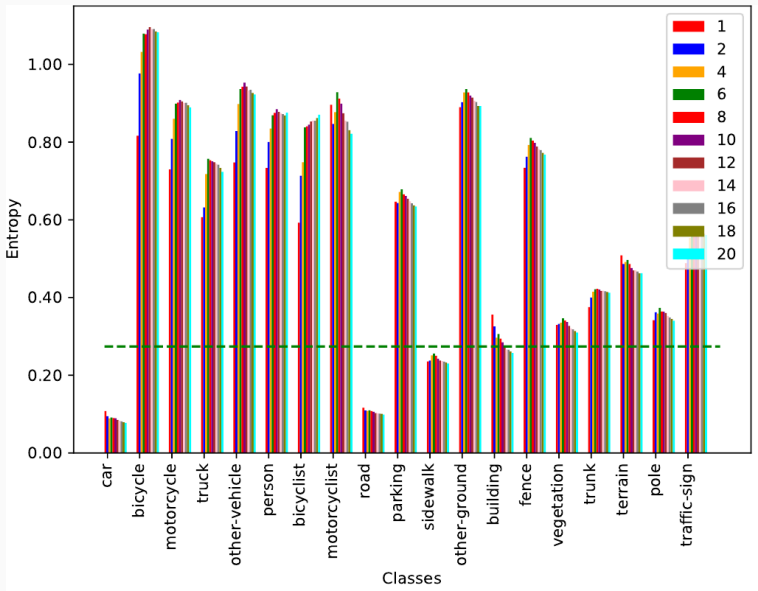
Fig. 3: Calibration error as a function of # of samples/ensembles for all methods in different models on FERPlus dataset

Paper fue presentado en Women in Computer Vision @ ECCV 2020.

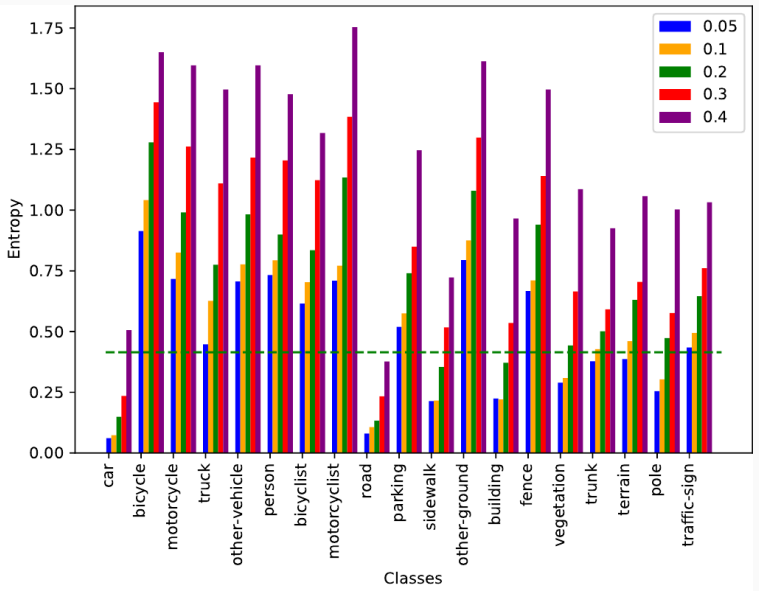
Incertidumbre en Clasificacion de Emociones

Fig. 4: Five most uncertain images based on DenseNet model and Deep Ensembles with # of ensembles and a plot of predictive probabilities using 1, 5, 10 and 15 ensembles. The first column represents the image, and the second its ground truth label distribution. Under each probability plot, the predicted class is presented.

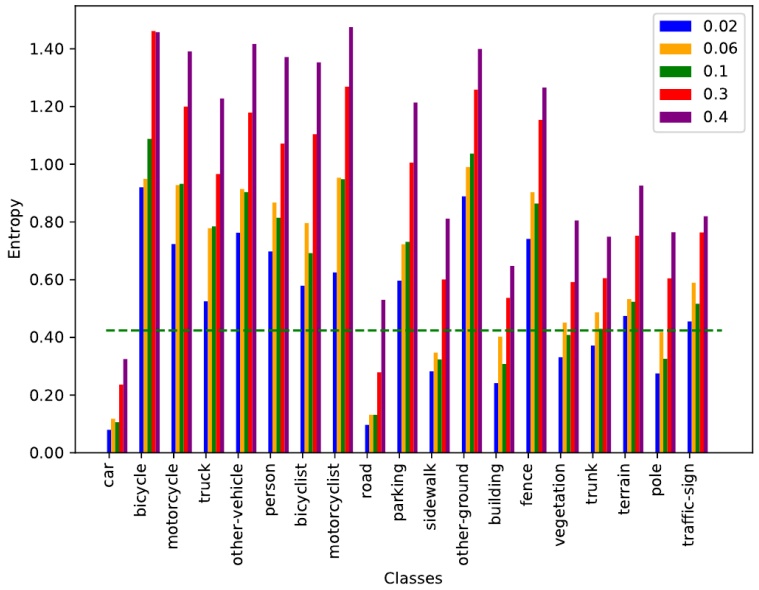
Incertidumbre en Segmentacion de Point Clouds (en KITTI)



(a) Deep Ensembles

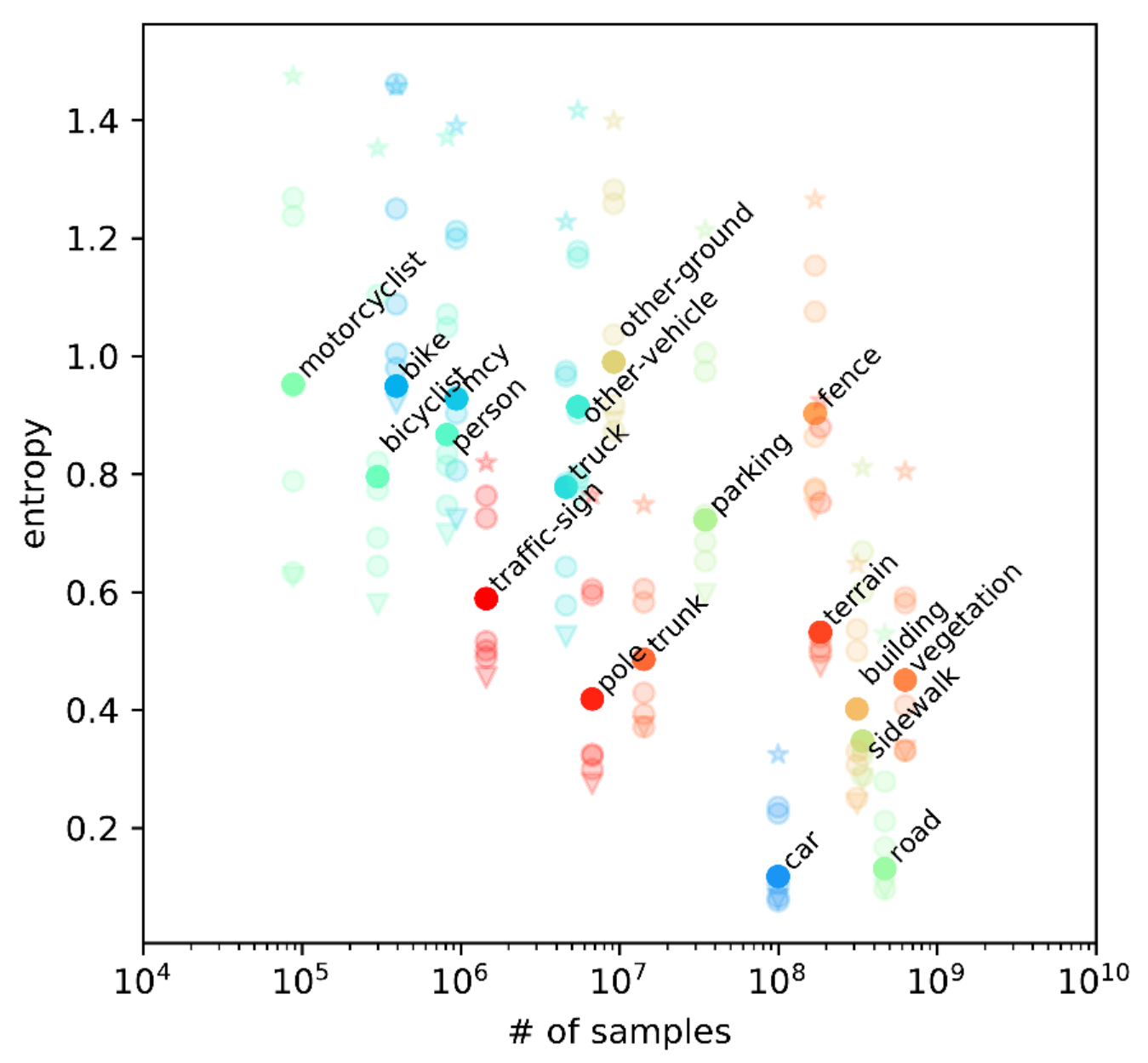


(b) MC-Dropout

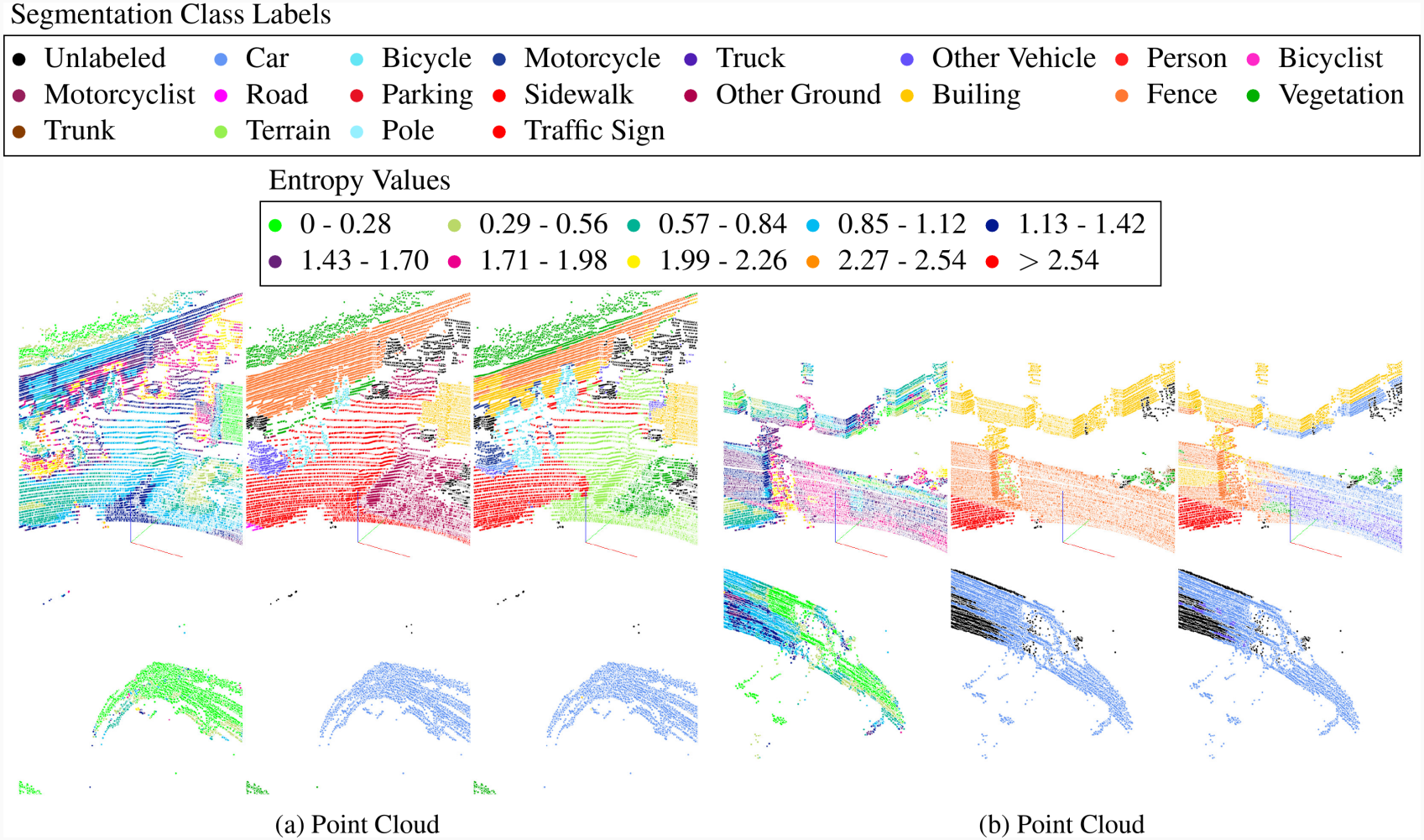


(c) MC-DropConnect

Incertidumbre en Segmentacion de Point Clouds (en KITTI)

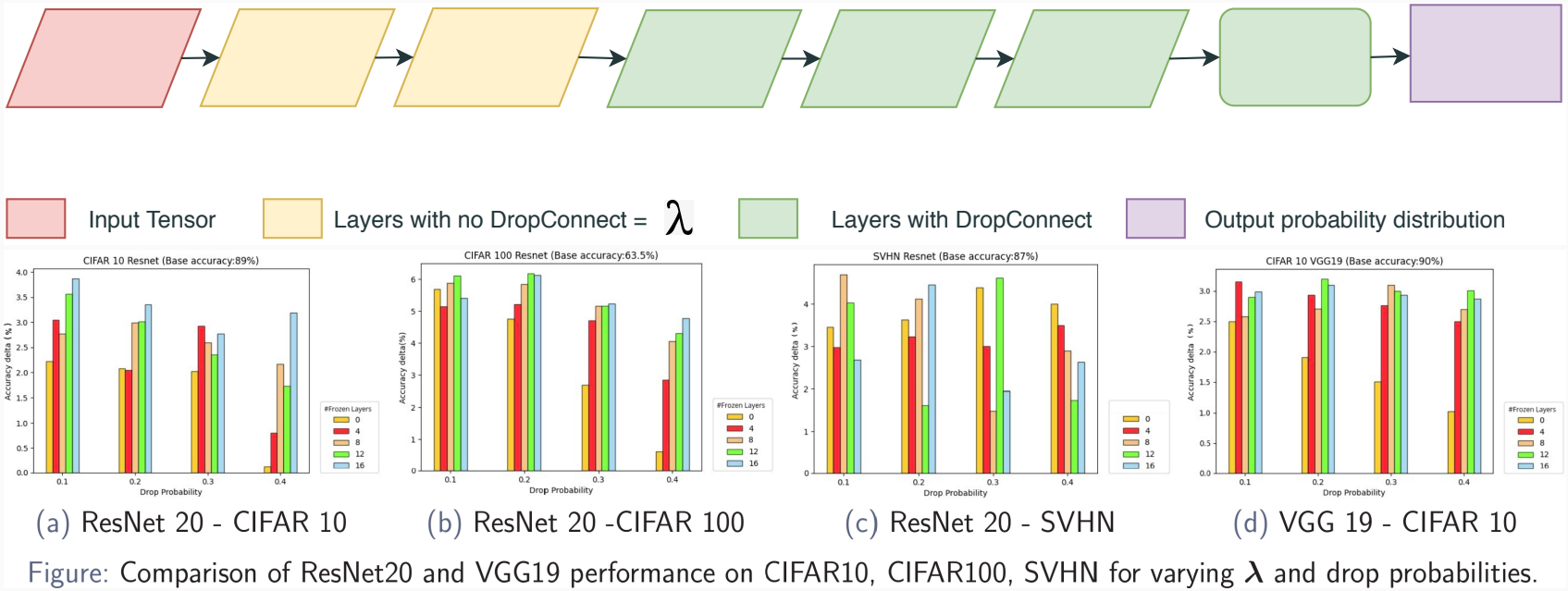


Incertidumbre en Segmentacion de Point Clouds (en KITTI)



En cada grupo, se presenta Entropia (izq), Segmentacion Correcta (centro), Segmentacion Predecida (derecha). Presentado en el Workshop on Uncertainty and Robustness @ ICML 2020.

SelectDC - DropConnect en Capas Seleccionadas



(a) ResNet 20 - CIFAR 10 (b) ResNet 20 -CIFAR 100 (c) ResNet 20 - SVHN (d) VGG 19 - CIFAR 10

Figure: Comparison of ResNet20 and VGG19 performance on CIFAR10, CIFAR100, SVHN for varying λ and drop probabilities.

Resultados de accuracy varian bastante, en general el rendimiento computacional mejora. Resultados en OOD indican que no hay perdida significativa de rendimiento.

Presentado mañana en el Ï Can't Believe Its Not Better"Workshop @ NeurIPS 2020.

Unsupervised Difficulty Estimation

Idea: Mirar como evoluciona el loss por cada ejemplo en el set de entrenamiento o validacion, acumulando el loss por cada sample. La hipotesis es que ejemplos mas dificiles acumulan mas loss que los ejemplos faciles.

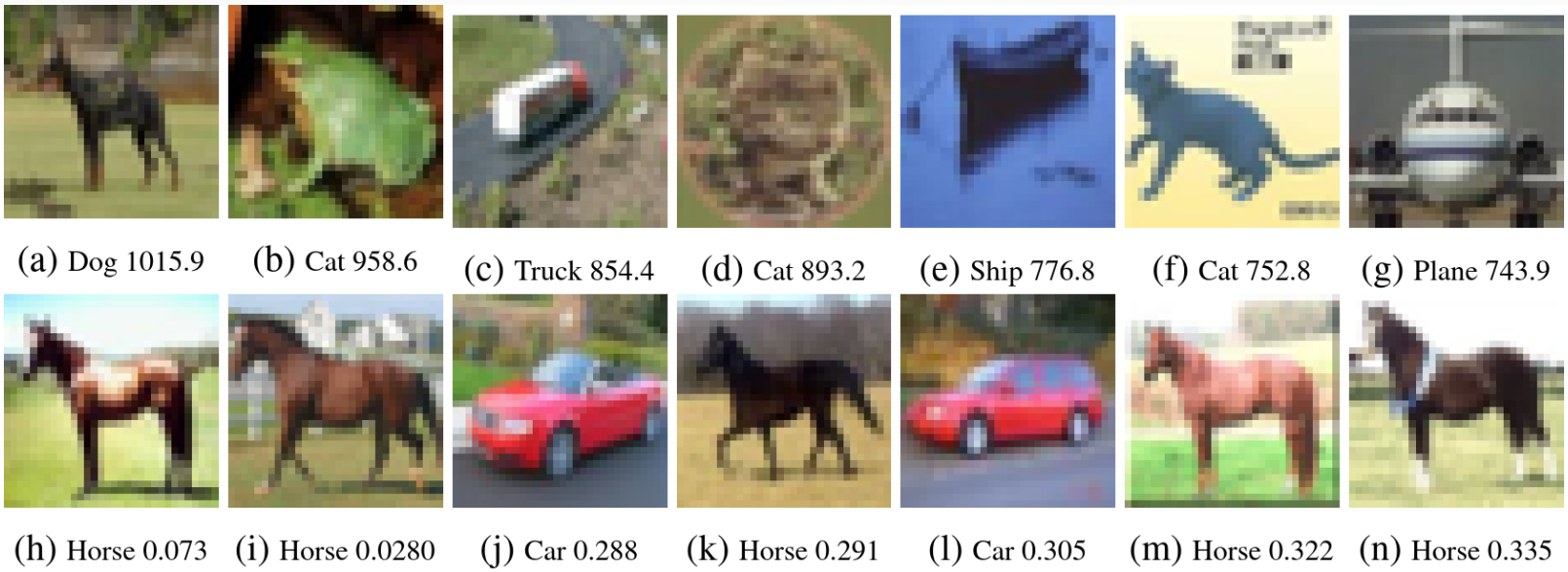
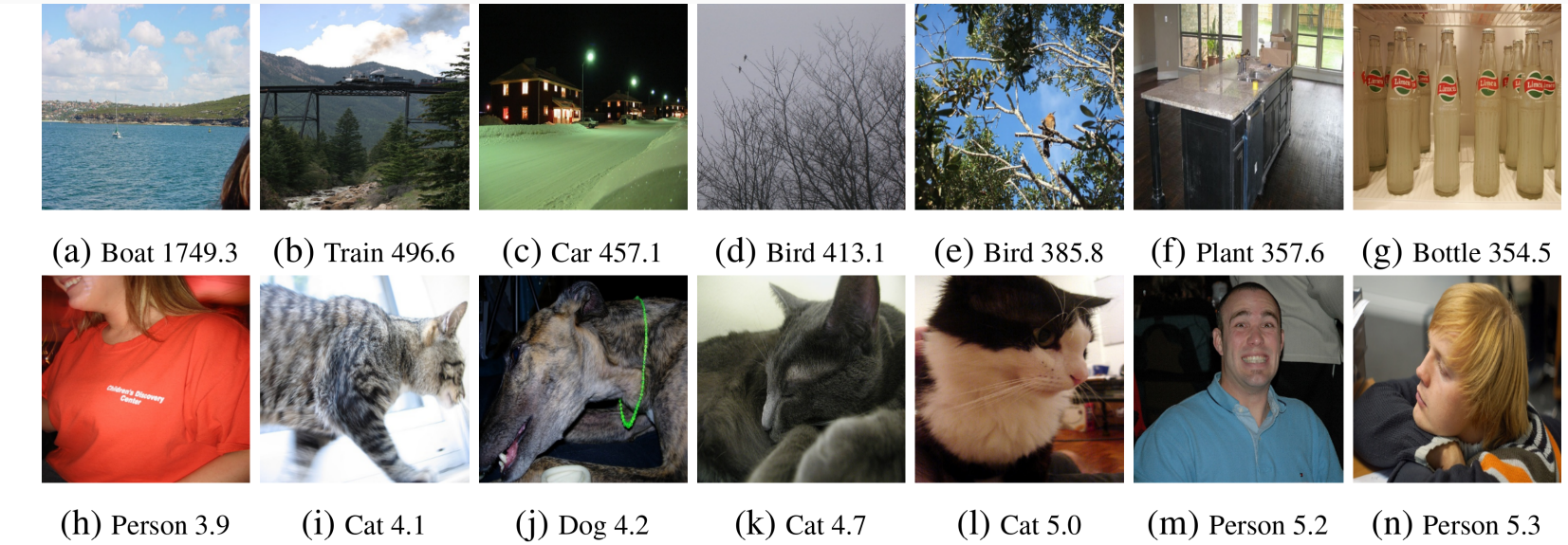


Figure 1: Most difficult (top-row) and easiest examples (bottom-row) in CIFAR10. Our proposed *action score* is displayed below each image as well as the true label.

Unsupervised Difficulty Estimation



(a) Boat 1749.3 (b) Train 496.6 (c) Car 457.1 (d) Bird 413.1 (e) Bird 385.8 (f) Plant 357.6 (g) Bottle 354.5

(h) Person 3.9 (i) Cat 4.1 (j) Dog 4.2 (k) Cat 4.7 (l) Cat 5.0 (m) Person 5.2 (n) Person 5.3

Figure 2: Most difficult (top-row) and easiest examples (bottom-row) in the VOC 2007-VAL with the SSD localization loss. The *action scores* are displayed below each image as well as the true label.

Creemos que el difficulty estimado con el action score tiene una relacion con la incertidumbre del modelo, pero eso quedara para un proximo paper :)

Cierre y Outlook

Aplicaciones

Medicina

Practicamente todas las aplicaciones medicas requieren estimar la incertidumbre correctamente para ser usadas con humanos/animales, conseguir aprobacion regulatoria, y ser utiles para que doctores medicos practicantes tomen decisiones.

Robotica

Generalmente en Robotica no se modela incertidumbres que pueden ser utiles, por ejemplo: incertidumbres en sistemas dinamicos (parametros), percepcion (deteccion de objetos), estimar cuando las capacidades del robot son extrapoladas. El mejor ejemplo de esto es Autonomous Driving.

Aplicaciones

Aprendizaje Reforzado (Reinforcement Learning)

De la misma forma, es muy importante tener policies aprendidas con RL que pueda estimar su propia incertidumbre y no decidir una accion cuando el entorno es muy diferente al entrenamiento.

- RL en robots o mecanismos reales, con consideraciones de seguridad (Safe RL).
- RL en entornos no estacionarios (por ejemplo, obstaculos dinamicos y impredecibles).
- Reducir el numero de ejemplos requeridos para entrenar a travez de Active Learning y Exploracion.

Conclusiones

- La incertidumbre es una medida útil para detectar ejemplos mal clasificados y fuera de distribución.
- Las Redes Neuronales Bayesianas no se utilizan a menudo en la práctica y muchas aplicaciones se beneficiarían de ellos. El rendimiento computacional es una gran razón.
- Es importante difundir estas tecnicas y sus posibles aplicaciones, especialmente ahora que ML se usa en aplicaciones reales que requieren estimar los limites del modelo.
- En general hay una gran brecha entre los investigadores que crean nuevas tecnicas y las aplicaciones mas concretas.
- Yo espero un mayor uso de estas técnicas en la práctica.

Otros Recursos

Hay tutoriales previos sobre este tema, que generalmente son mas teoricos y profundos ($\sim$ 2 Hrs):

NeurIPS 2019 Deep Learning with Bayesian Principles por Emtiyaz Khan.
`https://slideslive.com/38923183/deep-learning-with-bayesian-principles`

NeurIPS 2020 Practical Uncertainty Estimation and Out-of-Distribution Robustness in Deep Learning por Dustin Tran, Balaji Lakshminarayanan, y Jasper Snoek.
`https://neurips.cc/virtual/2020/protected/tutorial_0f190e6e164eafe66f011073b4486975.html`

Ambos estan disponibles en linea en SlidesLive.

Otros Recursos

Di una charla en PyData Global 2020 sobre Incertidumbre en Redes Neuronales usando la libreria Keras-Uncertainty (que yo desarrollé). El video esta disponible en:
`https://www.youtube.com/watch?v=ZTjquWQu-F8`

Tambien las slides y codigo en Jupyter notebook esta disponible en:
`https://github.com/mvaldenegro/talk-uncertainty-pydata-global-2020`

Keras-uncertainty esta disponible en:
`https://github.com/mvaldenegro/keras-uncertainty`

Bibliografia

- Charles Blundell y col. “Weight Uncertainty in Neural Network”. En: *International Conference on Machine Learning*. 2015, págs. 1613-1622.
- Yarin Gal y Zoubin Ghahramani. “Dropout as a bayesian approximation: Representing model uncertainty in deep learning”. En: *International Conference on Machine Learning*. 2016, págs. 1050-1059.
- Chuan Guo y col. “On calibration of modern neural networks”. En: *arXiv preprint arXiv:1706.04599* (2017).
- Fredrik K Gustafsson, Martin Danelljan y Thomas B Schon. “Evaluating scalable bayesian deep learning methods for robust computer vision”. En: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 2020, págs. 318-319.
- Alex Kendall y Yarin Gal. “What uncertainties do we need in bayesian deep learning for computer vision?” En: *Advances in Neural Information Processing Systems*. 2017, págs. 5574-5584.

Bibliografia (cont.)

- Balaji Lakshminarayanan, Alexander Pritzel y Charles Blundell. “Simple and scalable predictive uncertainty estimation using deep ensembles”. En: *Advances in Neural Information Processing Systems*. 2017, págs. 6402-6413.
- Maryam Matin y Matias Valdenegro-Toro. “Hey Human, If your Facial Emotions are Uncertain, You Should Use Bayesian Neural Networks!” En: *arXiv preprint arXiv:2008.07426* (2020).
- Aryan Mobiny y col. “DropConnect Is Effective in Modeling Uncertainty of Bayesian Deep Networks”. En: *arXiv preprint arXiv:1906.04569* (2019).
- Yaniv Ovadia y col. “Can you trust your model’s uncertainty? Evaluating predictive uncertainty under dataset shift”. En: *Advances in Neural Information Processing Systems*. 2019, págs. 13991-14002.