Incertidumbre en Redes Neuronales

Profesor: Felipe Tobar
Auxiliares: Mauricio Araneda, Alejandro Cuevas, Mauricio Romero
Alumnos: Felipe Cordova, Mario Vicuña, Miguel Videla
Fecha: 11 de Julio de 2019

I. Introducción

Las redes neuronales han sido ampliamente utilizadas en el último tiempo para diversas tareas de regresión y clasificación, sin embargo, este modelo presenta ciertas desventajas, tales como ser altamente propenso a sufrir sobreajuste y la ignorancia de la incertidumbre en el proceso de inferencia debido a su naturaleza determinista. Un ejemplo de la importancia del modelamiento de incertidumbre se presenta en tareas de clasificación en el área de la salud, donde es preferible que el modelo declare su ignorancia frente a casos médicos distintos a los previamente observados, los cuales pueden ser derivados con médicos especialistas para su resolución, a que el modelo establezca un diagnóstico sobreconfiado erróneo que pueda poner en riesgo la integridad de un paciente.

En el presente trabajo se abordará el problema del modelamiento de incertidumbre en redes neuronales en base al estudio de dos métodos relevantes en la literatura actual, correspondientes al método *MC Dropout* [1] como aproximador Bayesiano de redes neuronales y al método *Bayes by Backprop* [2] para modelamiento de incertidumbre en los pesos de las redes, donde se estudiarán los fundamentos teóricos de ambos métodos, aplicándolos sobre problemas ejemplificativos de regresión y clasificación, realizando un análisis y una comparativa de los modelos mencionados en base a los resultados obtenidos.

El documento se estructura de la siguiente manera: En la sección II se presenta, de manera breve y concisa, la revisión teórica de rigor de ambos métodos de modelamiento de incertidumbre en redes neuronales. En la sección III se explica en detalle la metodología experimental a desarrollar, los problemas de regresión y clasificación a estudiar y los respectivos conjuntos de datos a utilizar. En la sección IV se presentan los resultados obtenidos y se desarrolla un acabado análisis y discusión de los mismos, y finalmente, en la sección V se elabora un resumen de los aspectos importantes y las conclusiones del trabajo realizado, clarificando el aporte y el impacto del mismo, así como también, se establecen los lineamientos principales de posibles futuras extensiones de este.

II. Marco Teórico

II-A. MC Dropout

Usualmente para medir incertidumbre en redes neuronales se utiliza el enfoque probabilístico que da la función de salida softmax, en general, no existe ninguna garantía que dicha respuesta del modelo realmente describa la distribución en predicción del sistema, es usual que modelos con altos valores para la función softmax posean alta varianza en predicción al contrario de lo que la intuición indica. Con la intención de formalizar la estructura de una red neuronal capaz de capturar la incertidumbre del sistema se procede a hacer un estudio estocástico de las distribuciones inducidas por el método Dropout a las redes neuronales y mediante estimaciones de Monte Carlo aproximar los momentos de primer y segundo orden del valor predicho.

El primer paso para alcanzar una aproximación a los momentos de primer y segundo orden es mostrar que una red neuronal con una cantidad arbitraria de capas y funciones de activación no lineales las cuales se les aplica Dropout antes de cada capa de pesos, es matemáticamente equivalente a una aproximación estadística de un proceso Gaussiano. La demostración de dicho enunciado se encuentra ampliamente estudiada en [3], de la cual es importante destacar que el resultado es extrapolable a cualquier tipo de red neuronal y que no es necesario ningún supuesto simplificatorio para llegar a dicho resultado.

Otro resultado importante de esta derivación implica que el método de Dropout lo que realmente realiza es minimizar la divergencia de Kullback-Leibler (KL) entre la distribución del modelo a la distribución posterior del proceso Gaussiano, métrica generalmente utilizada como medida entre 2 distribuciones probabilísticas bajo el mismo espacio de medida. A continuación se resume la derivación de dicho resultado.

Dado el setting usual de una red neuronal de L capas escondidas asociadas a pesos W_i con dimensión $K_i \times K_{i-1}$, una función de perdida denotada como $E(\cdot, \cdot)$, regularizada por norma L_2 y un set de datos $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$. Luego la función de perdida regularizada viene dada por:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} E(y_i, \hat{y}_i) + \lambda \sum_{i=1}^{L} (||W_i||_2^2 + ||b_i||_2^2) \qquad (1)$$

Por otro lado un proceso Gaussiano que sigue una estructura de pesos dada por la distribución probabilística $p(w)$ en general implica que la distribución a posteriori $p(w|D)$, que nos permite estimar la distribución de predicción $p(y|x, D)$, resulte ser intrabable. Con el objetivo de hacer frente a dicho problema de cálculo, se decide aproximar la distribución intratable a partir de una distribución $q(w)$ la cual representa el comportamiento de matrices en donde las columnas se activan o permanecen iguales con probabilidad p y se vuelven 0 con probabilidad $1 - p$. Luego como criterio de aproximación se utiliza la divergencia KL entre $q(w)$ y la distribución posterior $p(w|D)$ obteniendo la siguiente función objetivo:

$$\mathcal{J} = -\sum_{i=1}^{N} \int q(w) \log p(y_i|x_i, w) dw + KL(q(w)||p(w)) \quad (2)$$

Luego realizando una integración por Monte Carlo sobre muestras de pesos $\hat{w}_n$, similar al muestreo de pesos por dropout, y escogiendo inteligentemente la precisión τ sobre el proceso Gaussiano es posible recuperar la expresión (1). Y por tanto se concluye que una red con Dropout aproxima las estadísticas de un proceso Gaussiano.

Ahora la derivación de la aproximación de los momentos de primer y segundo orden, pueden ser encontrados en [3] en donde se muestra que dadas las similitudes entre una NN Dropout y un proceso Gaussiano profundo es posible estimar el valor esperado de predicción como:

$$\mathbb{E}_{q(y|x)}(y) \approx \frac{1}{T}\sum_{t=1}^{T}\hat{y}(x, W_1^t, \ldots, W_n^t) \tag{3}$$

Donde T son la cantidad de realizaciones de una pasada hacia delante de la red NN con Dropout y W_i^t denota un sampling realizado por Dropout a la capa i-esima en la realización t.

Por otro lado la estimación del momento de segundo orden se calcula como:

$$\begin{aligned} VAR_{q(y|x)}(y) \approx & \frac{1}{T}\sum_{t=1}^{T}\hat{y}(x, W_1^t, \ldots, W_n^t)^\dagger \hat{y}(x, W_1^t, \ldots, W_n^t) \\ & - \mathbb{E}_{q(y|x)}(y)^\dagger \mathbb{E}_{q(y|x)}(y) \end{aligned} \tag{4}$$

Finalmente se llama método *MC Dropout* [1] a la estimación del valor esperado y la varianza de una NN Dropout por medio del calculo de el promedio y la varianza sobre T realizaciones de la red utilizando el muestre de pesos Dropout en cada pasada hacia delante.

Si bien, la conclusión de este método parece muy intuitiva, lo importante de [1] fue mostrar que por medio de un análisis estadístico profundo se puede demostrar matemáticamente que calcular los estadísticos sobre T realizaciones de una red converge a la verdadera esperanza y varianza del modelo, lo que al final del día permite estimar la incertidumbre de cualquier red neuronal a un costo computacional excepcionalmente bajo.

II-B. Bayes by Backprop

El algoritmo *Bayes by Backprop* consiste en un algoritmo eficiente, y compatible con el algoritmo *backpropagation*, para el aprendizaje de probabilidades de distribución de los pesos de una red neuronal, regularizando los pesos de la red mediante minimización del costo de compresión, también conocido como energía libre variacional. Los autores proponen entrenar un ensamble de redes neuronales donde los pesos de cada una provienen de una distribución de probabilidad Gaussiana compartida, doblando el numero de parámetros del modelo y permitiendo entrenar un ensamble infinito usando estimación insesgada de Monte Carlo de sus gradientes. Como en general, la inferencia Bayesiana exacta de los pesos de una red neuronal es intratable debido al gran número de parámetros de la misma y a que la forma del funcional de la red neuronal no posee una expresión cerrada para su integración, se propone utilizar una aproximación variacional para la actualización de las distribuciones mencionadas.

Considerando a las redes neuronales como un modelo probabilístico $P(y|x)$, donde dada una entrada $x \in \mathbb{R}^p$, el modelo asigna una probabilidad a cada posible salida $y \in \mathbf{Y}$ mediante un conjunto de parámetros w. Sea $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ un conjunto de datos de entrenamiento, la inferencia Bayesiana en redes neuronales calcula la distribución a posteriori $P(w|\mathcal{D})$, la cual es utilizada para realizar predicciones $\hat{y}$ sobre nuevos datos $\hat{x}$ mediante el calculo del valor esperado sobre la distribución a posteriori de los pesos $P(\hat{y}|\hat{x}) = \mathbb{E}_{P(w|\mathcal{D})}P(\hat{y}|\hat{x}, w)$, donde se observa que dicha expresión es equivalente a utilizar un ensamble infinito de redes neuronales. Desafortunadamente, la distribución a posteriori $P(w|\mathcal{D})$ resulta intratable debido al gran numero de parámetros de las redes neuronales, por tanto, se busca aproximar dicha distribución mediante una distribución variacional $q(w|\theta)$ de forma conocida, mediante la minimización de la divergencia de Kullback-Leibler (KL) entre $P(w|\mathcal{D})$ y $q(w|\theta)$, con respecto al parámetro de la distribución variacional θ, es decir, $\theta^* = \arg\min \mathrm{KL}(P(w|\mathcal{D}) \,||\, q(w|\theta))$, la cual, en virtud del teorema de Bayes, es equivalente a minimizar el funcional:

$$\mathcal{F}(\mathcal{D}, \theta) = \mathrm{KL}(q(w|\theta) \,||\, P(w)) - \mathbb{E}_{q(w|\theta)} \log P(\mathcal{D}|w) \tag{5}$$

Conocido como energía libre variacional.

La expresión (5) puede ser aproximada mediante un muestreo de Monte Carlo de los pesos w_i de la distribución variacional posterior $q(w_i|\theta)$. De este modo, el funcional aproximado resulta:

$$\mathcal{F}(\mathcal{D}, \theta) \approx \frac{1}{N}\sum_{i=1}^{N} \log q(w_i|\theta) - \log P(w_i) - \log P(\mathcal{D}|w_i) \tag{6}$$

Los autores mostraron que, dados una variable aleatoria $\epsilon \sim q(\epsilon)$ y $w = t(\theta, \epsilon)$ con $t(\theta, \epsilon)$ una función determinística, suponiendo una distribución marginal de probabilidad $q(w|\theta)$ de w tal que cumpla $q(\epsilon)d\epsilon = q(w|\theta)dw$, entonces para una función f con derivadas en w se cumple:

$$\frac{\partial}{\partial\theta}\mathbb{E}_{q(w|\theta)}[f(w, \theta)] = \mathbb{E}_{q(\epsilon)}\left[\frac{\partial f(w,\theta)}{\partial w}\frac{\partial w}{\partial \theta} + \frac{\partial f(w,\theta)}{\partial \theta}\right] \tag{7}$$

Luego, suponiendo una distribución variacional posterior $q(w_i|\theta)$ Gaussiana diagonal, una muestra w de dicha distribución puede ser obtenida como $w = t(\theta, \epsilon) = \mu + \log(1 + \exp(\rho)) \circ \epsilon$, con $\theta = (\mu, \rho)$ el vector de medias y de desviaciones estándar parametrizadas ($\sigma = \log(1 + \exp(\rho))$) de la distribución variacional posterior, respectivamente, ϵ un parámetro (libre) de ruido y $\circ$ el producto punto a punto. De este modo, aplicando la proposición (7) en el problema de optimización (5), considerando $f(w, \theta) = \log q(w|\theta) - \log P(w)P(\mathcal{D}|w)$, los gradientes de los parámetros $\theta = (\mu, \rho)$ de la distribución variacional posterior pueden ser calculados como:

$$\begin{aligned} \Delta_\mu &= \frac{\partial f(w,\theta)}{\partial w} + \frac{\partial f(w,\theta)}{\partial \theta} \\ \Delta_\rho &= \frac{\partial f(w,\theta)}{\partial w}\frac{\epsilon}{1 + \exp(-\rho)} + \frac{\partial f(w,\theta)}{\partial \rho} \end{aligned} \tag{8}$$

En consecuencia, los parámetros μ y ρ pueden ser actualizados mediante gradiente descendente, notando que el término $\frac{\partial f(w,\theta)}{\partial w}$ corresponde al gradiente determinado por el algoritmo *backpropation* en redes neuronales. De este modo, es posible

entrenar los parámetros $\theta = (\mu, \rho)$ de las distribuciones de los pesos w de la red neuronal de manera tratable.

Finalmente, los autores proponen utilizar una mezcla escalada de dos Gaussianas como distribución a priori fija $P(w)$:

$$P(w) = \prod_j \pi \mathcal{N}(w_j; 0, \sigma_1^2) + (1-\pi)\mathcal{N}(w_j; 0, \sigma_2^2) \quad (9)$$

Con $\sigma_1 > \sigma_2$, $\sigma_2 \ll 1$ y $\pi \in [0, 1]$.

III. Metodología Experimental

En la presente sección se detalla metodología experimental utilizada, donde se presentan dos casos ilustrativos y sencillos de problemas de regresión y clasificación para el análisis del modelamiento de incertidumbre de los métodos propuestos.

III-A. Problema de Regresión

Se procedió a medir la incertidumbre en las predicciones de los métodos anteriormente presentados sobre un problema de regresión. Para ello, se generó un conjunto de entrenamiento de 500 muestras de una función sinusoidal perturbada:

$$y = 3x + 10\sin(2\pi(x+\epsilon)) + 10\sin(4\pi(x+\epsilon)) + \epsilon \quad (10)$$

Con $\epsilon \sim \mathcal{N}(0, 0.02)$, mientras que el conjunto de evaluación consistió en 1000 muestras equiespaciadas entre -0.5 y 1.5.

Ambos modelos neuronales constaron de 2 capas ocultas de 75 y 50 neuronas, respectivamente, una tasa de aprendizaje $\mu = 0.03$, un optimizador de gradiente descendente estocástico, un tamaño de batch de 500, un muestreo de Monte Carlo de 500 muestras por cada evaluación y 2500 epócas de entrenamiento. Para el método *Bayes By Backprop*, se seleccionó la distribución a priori recomendada en (9), con $\sigma_1 = 1.0$, $\sigma_2 = 0.1$ y $\pi = 0.2$. Para el método *MC Dropout* se utilizó una probabilidad de supresión de unidades neuronales $p = 0.2$.

III-B. Problema de Clasificación

Se procedió a entrenar un modelo neuronal simple *MLP*, un modelo neuronal con *MC Dropout* y un modelo neuronal con *Bayes by Backprop* sobre el conjunto de datos de dígitos manuscritos MNIST. Se seleccionó un conjunto de entrenamiento aleatorio de 60000 muestras y un conjunto de prueba de 10000 muestras, donde cada imagen de dimensión 28x28 fue transformada a dimensión 1x784, mediante estiramiento.

Todos modelos neuronales constaron de 2 capas ocultas de 100 y 100 neuronas, respectivamente, una tasa de aprendizaje $\mu = 0.001$, un optimizador de gradiente descendente estocástico, un tamaño de batch de 1000, un muestreo de Monte Carlo de 200 muestras por cada evaluación y 10 épocas de entrenamiento. Para el método *Bayes By Backprop*, se seleccionó la distribución a priori recomendada en (9), con $\sigma_1 = 1$, $\sigma_2 = e^{-6}$ y $\pi = 0.5$. Para el método *MC Dropout* se utilizó una probabilidad de supresión de unidades neuronales $p = 0.2$.

Para medir la incertidumbre de clasificación, se evaluaron los modelos entrenados sobre muestras del conjunto de letras manuscritas japonesas KMNIST, del mismo formato que el conjunto de datos MNIST.

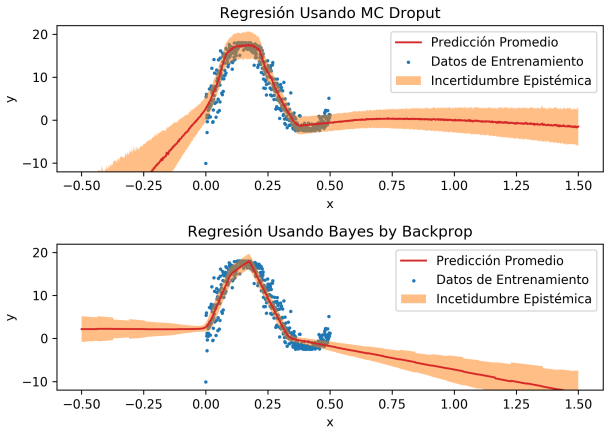

Figura 1. Predicción promedio e incertidumbre epistémica ($\mu \pm 2\sigma$) de métodos de modelamiento de incertidumbre en redes neuronales *MC Dropout* y *Bayes by Backprop*.

IV. Resultados y Discusión

En la presente sección se muestran los resultados obtenidos de los experimentos especificados en la sección III, junto la elaboración de los respectivos análisis y discusiones.

IV-A. Problema de Regresión

En la figura 1 se presentan los resultados de regresión de los distintos modelos neuronales sobre el conjunto de datos sinusoidal perturbado.

Para ambos modelos es posible observar como la predicción sobre el conjunto de entrenamiento describe de buena manera la distribución de los datos, naturalmente se verifica que dado los datos de entrenamiento brindados ambos modelos convergen a una distribución de parámetros óptima.

Es importante mencionar que para los puntos que distan a la región de entrenamiento el comportamiento predicho no es equivalente para ambos modelos, este resultado es esperado puesto que el grado de generalización depende explícitamente de la representatividad del conjunto de entrenamiento con respecto al fenómeno a modelar, por lo tanto aquellas regiones alejadas del setting de entrenamiento no se encuentran representadas en el. Ahora lo que si es de interés desde el punto de vista de análisis es la incertidumbre epistémica medida en ambas redes, para ambos casos las regiones de entrenamiento poseen en general baja incertidumbre (la franja naranja representa el área donde el 95 % de las predicciones bajo hipótesis Gaussiana). Por otro lado se observa que en ambos casos un crecimiento constante en la incertidumbre proporcional a la distancia del punto evaluado con respecto a los datos de entrenamiento, este resultado puede parecer intuitivo pero en realidad es muy importante puesto que permite validar una hipótesis que se asume en las redes de predicción de forma rigurosa, tanto para los métodos de *MC Dropout* y *Bayes By Backprop*.

IV-B. Problema de Clasificación

En la tabla I se presentan los resultados de clasificación de los tres modelos neuronales especificados sobre conjunto

Tabla I
RESULTADOS DE CLASIFICACIÓN DE MODELOS NEURONALES SOBRE CONJUNTO DE DATOS MNIST.

Modelo	Accuracy
MLP	0.9707
MC Dropout	0.9722
Bayes by Backprop	0.9273

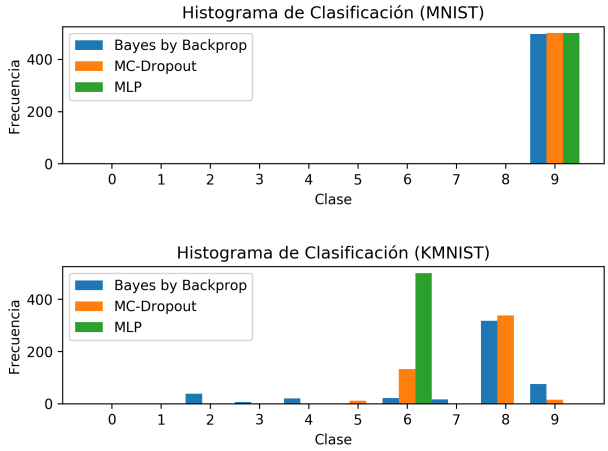

Figura 2. Histograma de clasificación de modelos neuronales *MLP*, *MC Dropout* y *Bayes by Backprop*, sobre una muestra del conjunto de datos MNIST ('9') y KMNIST ('や').

de datos MNIST. En la figura 2 se presentan los histogramas de clasificación mediante muestreo de Monte Carlo sobre una muestra conocida (MNIST) y una desconocida (KMNIST).

Similarmente al caso de regresión, es apreciable en la figura 2 que los tres modelos son capaces de predecir correctamente una muestra del conjunto de datos MNIST conocido (dígito '9'), donde tanto el modelo *MC Dropout* como el *Bayes by Backprop* predicen la clase correcta con $100\,\%$ de certeza, es decir, entregan la misma predicción en cada uno de los muestreos de Monte Carlo. En cambio, se aprecia que al intentar predecir una muestra radicalmente distinta a las presentadas en entrenamiento, perteneciente al conjunto de datos KMNIST, los modelos que *MC Droput* y *Bayes by Backprop* incrementan notablemente su incertidumbre, donde el histograma de predicción de cada muestra de Montecarlo, tiende a distribuirse por distintos símbolos, en vez de concentrarse en uno sólo como el caso anterior, mientras que, de acuerdo a lo esperado, el modelo *MLP* predice con total certeza una predicción errónea del nuevo símbolo observado ya que es incapaz de modelar la incertidumbre en sus predicciones, quedando de manifiesto las ventajas de los métodos de modelamiento de incertidumbre en redes neuronales.

En la tabla I se aprecia que la certeza de predicción sobre el conjunto MNIST de prueba es superior en el modelo *Bayes by Backprop* que el modelo *MLP*, lo cual se explica por la inherente mejora de predicción al modelar la incertidumbre, considerando además que el muestreo de Monte Carlo puede entenderse como un ensamble de múltiples redes neuronales, por tanto, es esperable que el ensamble presente mejor desempeño que sólo un modelo, notando que el ensamble induce una auto-regularización del modelo, por tanto, el modelo de ensamble resulta robusto al sobreajuste, mientras que el modelo *MLP* resulta muy susceptible a sufrir este fenómeno. Sin embargo, se observa que, a pesar que el modelo *Bayes by Backprop* resulta similar al *MC Dropout*, en el sentido que ambos son modelos probabilísticos capaces de modelar la incertidumbre y pueden ser entendidos como un ensamble de redes neuronales, este presente un considerable peor desempeño en comparación a los demás, lo cual puede ser explicado por la inadecuada elección de los parámetros de su distribución a priori (Mezcla ponderada de Gaussianas), el cual pudo inducir una regularización demasiado fuerte impidiendo que el modelo logre ajustarse de mejor manera a los datos, además de considerar que este modelo, a diferencia de los demás, dobla la cantidad de parámetros (Media y varianza por cada peso), por lo cual, se espera que requiera una mayor cantidad de épocas para alcanzar la convergencia.

V. CONCLUSIONES

En el presente trabajo se abordado de manera concisa y completa el tópico de modelamiento de incertidumbre en redes neuronales estudiando dos métodos relevantes, *MC Dropout* y *Bayes by Backprop*, entregando la motivación del modelamiento de incertidumbre, presentando los fundamentos teóricos de cada uno y comprobando su efectividad mediante dos experimentos ilustrativos abocados a tareas de regresión y clasificación.

En las experiencias desarrolladas se comprobó la capacidad de modelamiento de incertidumbres de los métodos aludidos, comprobando la relación de proporcionalidad entre el nivel de incertidumbre de las predicciones de estos modelos y el nivel de desconocimiento de la respectiva observación, donde se observó que si bien, ambos modelos resultan aproximaciones tratables de redes Bayesianas, el método *MC Dropout* destaca por su simpleza de implementación, entendimiento de técnicas otrora heurísticas (Dropout), y por su eficiencia, no añadiendo parámetros extras al modelo. En cambio el modelo *Bayes by Backprop*, dobla el numero de parámetros de una red neuronal tradicional, además de añadir hiper-parámetros relativos a la distribución a priori, de alta injerencia en el desempeño del modelo, aumentando la complejidad de diseño del mismo frente a un problema específico.

El trabajo desarrollado expone una revisión clara y didáctica del enfoque probabilístico Bayesiano en redes neuronales con fines pedagógicos, siendo un tópico de vanguardia en el área de aprendizaje de máquinas, el cual es extensamente estudiado por la academia, pero vagamente aplicado en la industria, por desconocimiento y por su complejidad teórica inherente, por tanto, esperamos que este trabajo ayude a facilitar y difundir esta importante temática.

Como posible futura extensión del presente trabajo, se propone estudiar el modelamiento de incertidumbre en arquitecturas neuronales de mayor complejidad, tales como, las redes convolucionales y las redes recurrentes, analizando las ventajas y dificultades de la incorporación del enfoque probabilístico en ellas.

REFERENCIAS

[1] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.
[2] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural networks," *arXiv preprint arXiv:1505.05424*, 2015.
[3] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Appendix," 2015.