

# Uncertainty estimation in BERT-based Named Entity Recognition

Łukasz Rączkowski, Riccardo Belluzzo, Paweł Olszewski, Piotr Zieliński, Paweł Zawistowski

06.11.2022

# Agenda

- Named Entity Recognition in Allegro
- Uncertainty estimation
  - Why do we need uncertainty estimation?
  - Bayesian Neural Networks
  - Variational dropout
- Results
  - NER calibration
  - Misclassification detection
  - Out-of-distribution detection
- Conclusions



# Named Entity Recognition

**allegro**[search many](#)[All categories](#)[SEARCH](#)

**Parameters**

Condition	New
Invoice	With VAT invoice
Sleeve Size	Big boi
Manufacturer	Some Company
Name	Super Product Pro
Packaging Status	original
Color	multicolor, other

**Description**

Our product is the best of its kind. Its qualities are truly extraordinary and mindblowing. It has 128 GB of RAM, imagine that.

Also, it can't be forgotten that the Super Product makes literally everyone happy on first sight. People tend to dance in joy when they hold it in their hands. It's made for all ages, but its effects vary between age groups.

Oh, and its catalogue number is BUY-M3-4S4P. Order now!

# Named Entity Recognition

**allegro**[search many](#)[All categories](#)[SEARCH](#)

**Parameters**

Condition	New
Invoice	With VAT invoice
Sleeve Size	Big boi
Manufacturer	Some Company
Name	Super Product Pro
Packaging Status	original
Color	multicolor, other

**Description**

Our product is the best of its kind. Its qualities are truly extraordinary and mindblowing. It has 128 GB of RAM, imagine that.

Also, it can't be forgotten that the Super Product makes literally everyone happy on first sight. People tend to dance in joy when they hold it in their hands. It's made for all ages, but its effects vary between age groups.

Oh, and its catalogue number is BUY-M3-4S4P. Order now!

## Named Entity Recognition

Our product is the best of its kind. Its qualities are truly extraordinary and mindblowing. It has 128 GB of RAM, imagine that.

Also, it can't be forgotten that the Super Product makes literally everyone happy on first sight. People tend to dance in joy when they hold it in their hands. It's made for all ages, but its effects vary between age groups.

Oh, and its catalogue number is BUY-M3-4S4P. Order now!

# Named Entity Recognition

Our product is the best of its kind. Its qualities are truly extraordinary and mindblowing. It has 128 GB of RAM, imagine that.

Entity:

Memory size

Also, it can't be forgotten that the Super Product makes literally everyone happy on first sight. People tend to dance in job when they hold it in their hands. It's made for all ages, but its effects vary between age groups.

Model name

Oh, and its catalogue number is BUY-M3-4S4P Order now!

Catalogue number

# Named Entity Recognition

Our product is the best of its kind. Its qualities are truly extraordinary and mindblowing. It has **128 GB** of RAM, imagine that.

Also, it can't be forgotten that the **Super Product** makes literally everyone happy on first sight. People tend to dance in job when they hold it in their hands. It's made for all ages, but its effects vary between age groups.

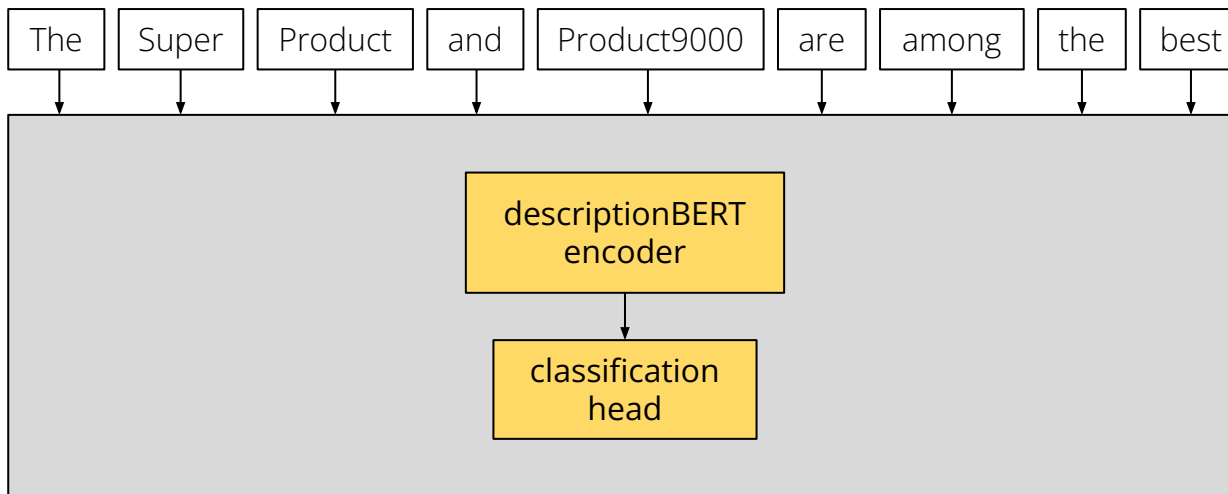
Oh, and its catalogue number is **BUY-M3-4S4P** Order now!

## Parameters

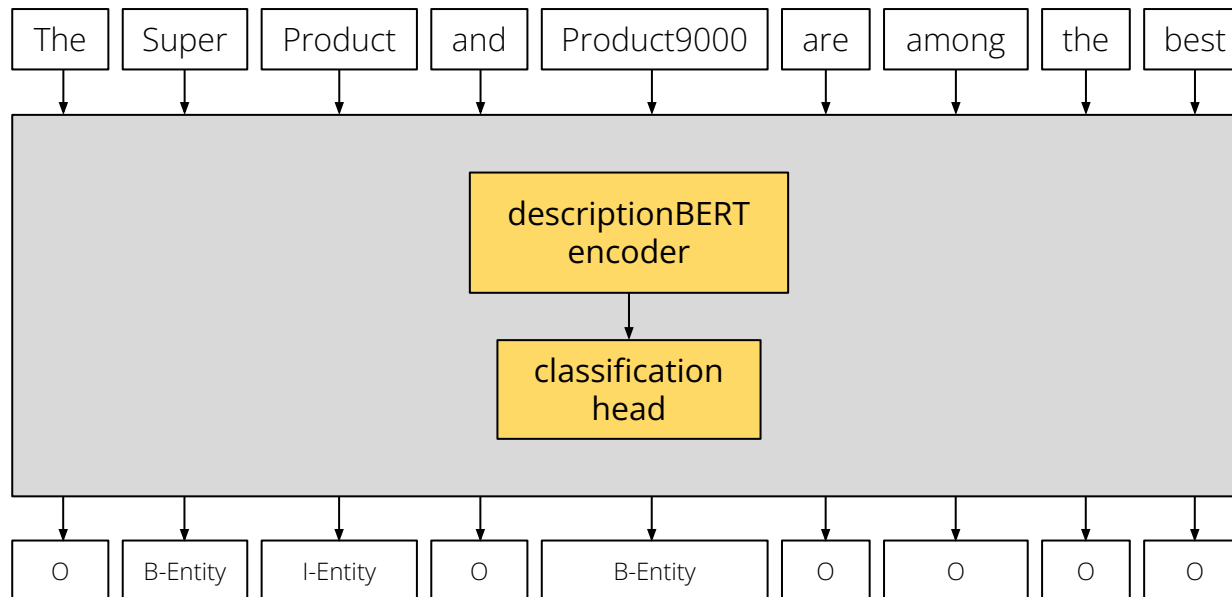
Condition	New
Invoice	With VAT invoice
Sleeve Size	Big boi
Manufacturer	Some Company
Name	Super Product Pro
Packaging Status	original
Color	multicolor, other
Memory	128 GB
Catalogue number	BUY-M3-4S4P



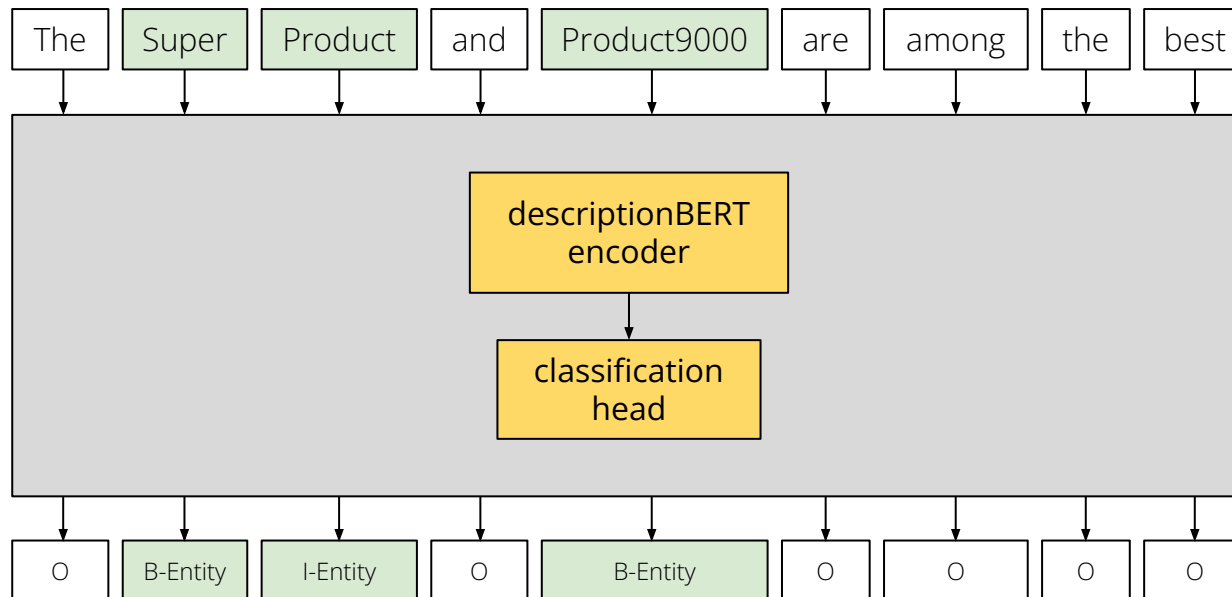
## Sequence Tagging Scheme



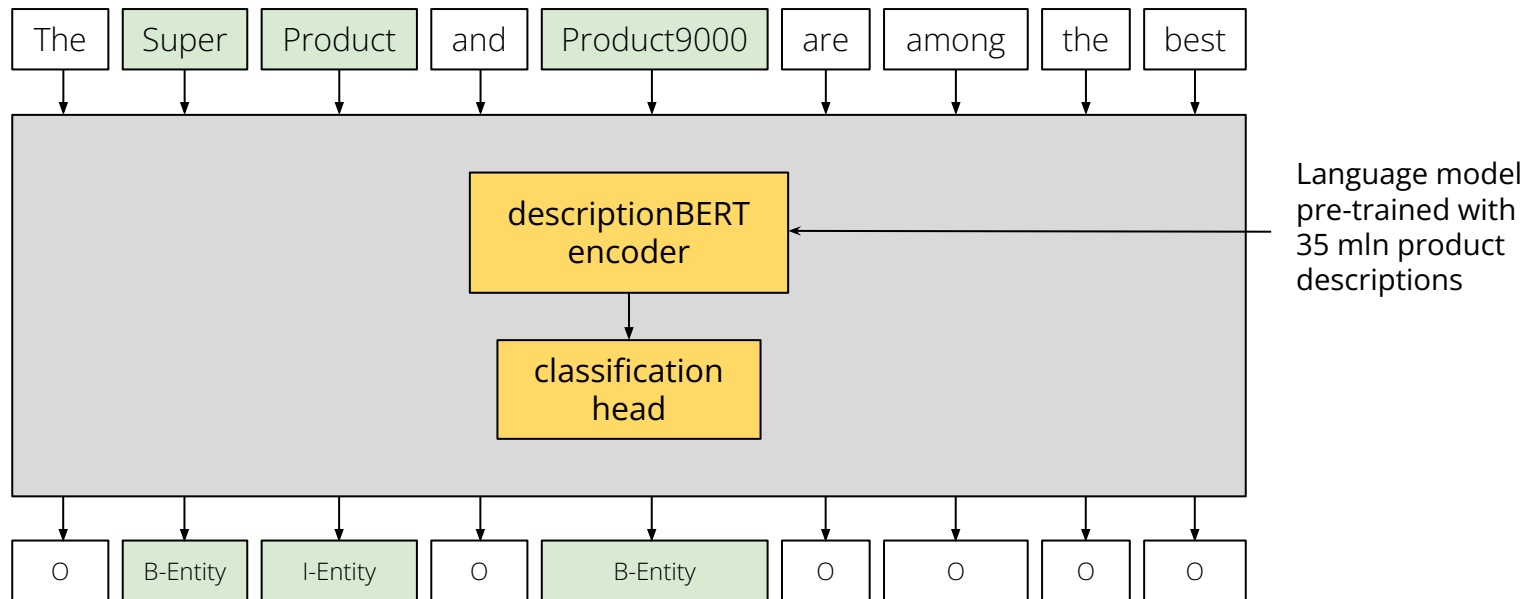
## Sequence Tagging Scheme



## Sequence Tagging Scheme



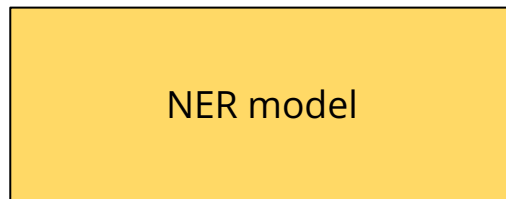
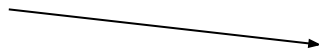
## Sequence Tagging Scheme





## Why do we need uncertainty estimation?

The Super Product and Product9000 are among the best.



Model name: Super Product

90%

Model name: Product9000

92%



## Why do we need uncertainty estimation?

The Super Product and Product9000 are among the best.

This really is a great product, I like it a lot.

NER model

Model name: Super Product

90%

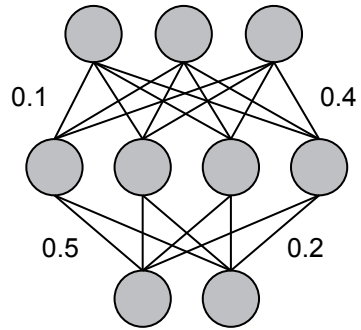
Model name: Product9000

92%

Model name: great product

??%

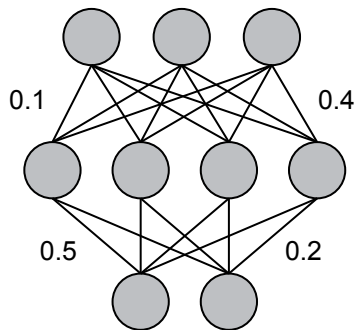
# Bayesian Neural Networks



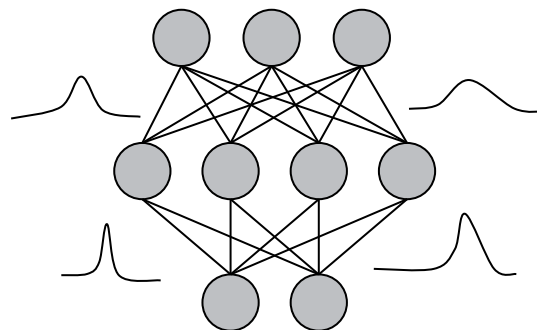
Standard Neural Network



# Bayesian Neural Networks



Standard Neural Network



Bayesian Neural Network

# Variational inference

Predictive distribution:

$$P(y'|x', D_{tr})$$

## Variational inference

Predictive distribution:

$$P(y'|x', D_{tr}) = \int P(y'|x', \omega) P(\omega|D_{tr}) d\omega$$

## Variational inference

Predictive distribution:

$$P(y'|x', D_{tr}) = \int P(y'|x', \omega) \underbrace{P(\omega|D_{tr})}_{\text{Intractable posterior}} d\omega$$

## Variational inference

Predictive distribution:

$$P(y'|x', D_{tr}) = \int P(y'|x', \omega) \underbrace{P(\omega|D_{tr})}_{\text{Intractable posterior}} d\omega$$

We approximate the posterior with a variational distribution  $q(\omega)$ .

## Variational inference

Predictive distribution:

$$P(y'|x', D_{tr}) = \int P(y'|x', \omega) \underbrace{P(\omega|D_{tr})}_{\text{Intractable posterior}} d\omega$$

We approximate the posterior with a variational distribution  $q(\omega)$ .

We want to minimize  $KL(q(\omega) || P(\omega|D_{tr}))$  which is equivalent to optimizing the variational lower bound:

$$\mathcal{L}_{VI} = \int q(\omega) \log P(D_{tr}|\omega) d\omega - KL(q(\omega) || P(\omega))$$

## Variational dropout

$$\omega = [W_i]_{i=1}^L$$

## Variational dropout

$$\omega = [W_i]_{i=1}^L$$

$$z_{i,j} \sim \text{Bernoulli}(p_i)$$

$$W_i = M_i \cdot \text{diag}\left([z_{i,j}]_{j=1}^{K_i}\right)$$

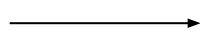


## Variational dropout

$$\omega = [W_i]_{i=1}^L$$

$$z_{i,j} \sim \text{Bernoulli}(p_i)$$

$$W_i = M_i \cdot \text{diag}\left([z_{i,j}]_{j=1}^{K_i}\right)$$



$q(\omega)$  is parameterized by  $M_i$

## Variational dropout

$$\omega = [W_i]_{i=1}^L$$

$$z_{i,j} \sim \text{Bernoulli}(p_i)$$

$$W_i = M_i \cdot \text{diag}\left([z_{i,j}]_{j=1}^{K_i}\right) \longrightarrow q(\omega) \text{ is parameterized by } M_i$$

$$\mathcal{L}_{VI} = \int q(\omega) \log P(D_{tr}|\omega) d\omega - KL(q(\omega) || P(\omega))$$

## Variational dropout

$$\omega = [W_i]_{i=1}^L$$

$$\begin{aligned} z_{i,j} &\sim \text{Bernoulli}(p_i) \\ W_i &= M_i \cdot \text{diag}([z_{i,j}]_{j=1}^{K_i}) \end{aligned} \longrightarrow q(\omega) \text{ is parameterized by } M_i$$

$$\mathcal{L}_{VI} = \int q(\omega) \log P(D_{tr}|\omega) d\omega - KL(q(\omega) || P(\omega))$$

$$\hat{\mathcal{L}}_{VI} = \frac{1}{N} \sum_{i=1}^N E(y_i, \hat{f}(x_i, \hat{\omega}_i)) - KL(q(\omega) || P(\omega)) \longrightarrow \hat{\omega}_i \sim q(\omega)$$

## Variational dropout

$$\omega = [W_i]_{i=1}^L$$

$$\begin{aligned} z_{i,j} &\sim \text{Bernoulli}(p_i) \\ W_i &= M_i \cdot \text{diag}([z_{i,j}]_{j=1}^{K_i}) \end{aligned} \longrightarrow q(\omega) \text{ is parameterized by } M_i$$

$$\mathcal{L}_{VI} = \int q(\omega) \log P(D_{tr}|\omega) d\omega - KL(q(\omega) || P(\omega))$$

$$\begin{aligned} \hat{\mathcal{L}}_{VI} &= \frac{1}{N} \sum_{i=1}^N E(y_i, \hat{f}(x_i, \hat{\omega}_i)) - KL(q(\omega) || P(\omega)) \longrightarrow \hat{\omega}_i \sim q(\omega) \\ &\approx \frac{1}{N} \sum_{i=1}^N E(y_i, \hat{f}(x_i, \hat{\omega}_i)) + \lambda \sum ||W_i||^2 \end{aligned}$$

## Variational dropout

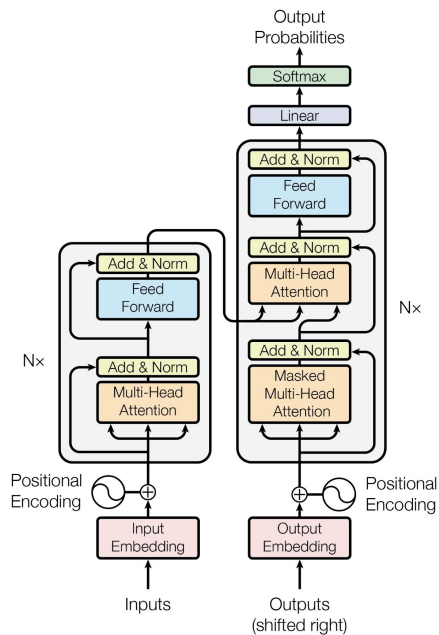
$$P(y'|x', D_{tr}) = \int P(y'|x', \omega) P(\omega|D_{tr}) d\omega$$

## Variational dropout

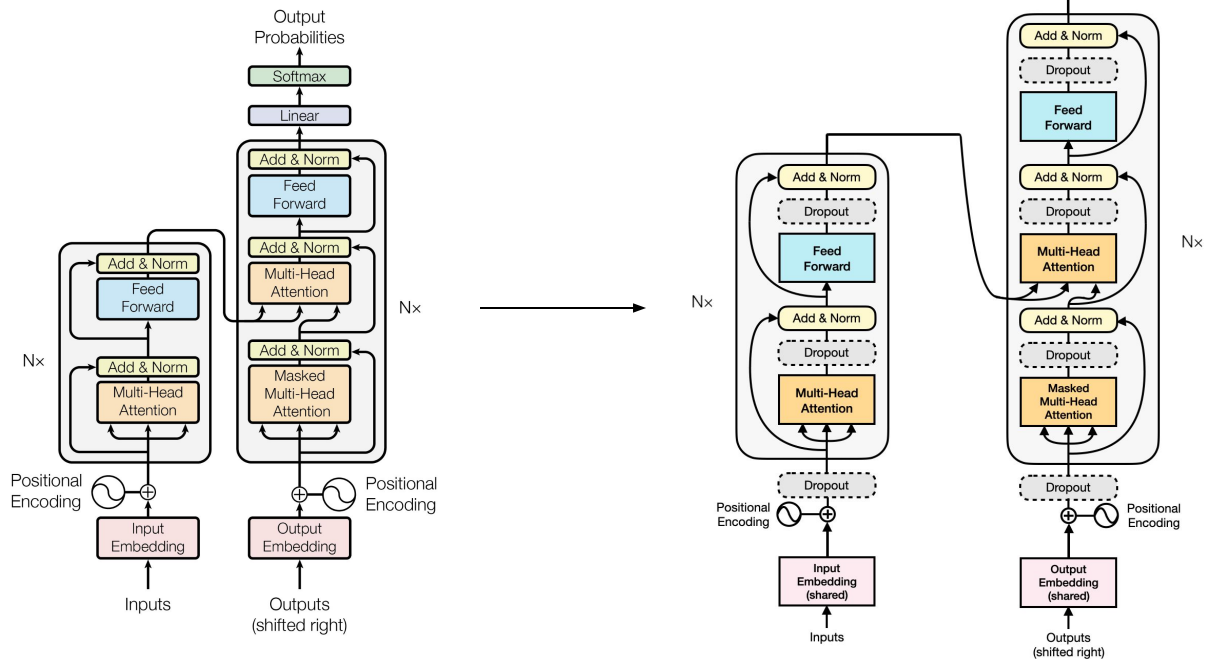
$$\begin{aligned} P(y'|x', D_{tr}) &= \int P(y'|x', \omega) P(\omega|D_{tr}) d\omega \\ &\approx \int P(y'|x', \omega) q(\omega) \approx \frac{1}{T} \sum_{t=1}^T P(y'|x', \hat{\omega}_t) \longrightarrow \hat{\omega}_t \sim q(\omega) \end{aligned}$$

$T$  - number of variational dropout calls

# Variational dropout in BERT

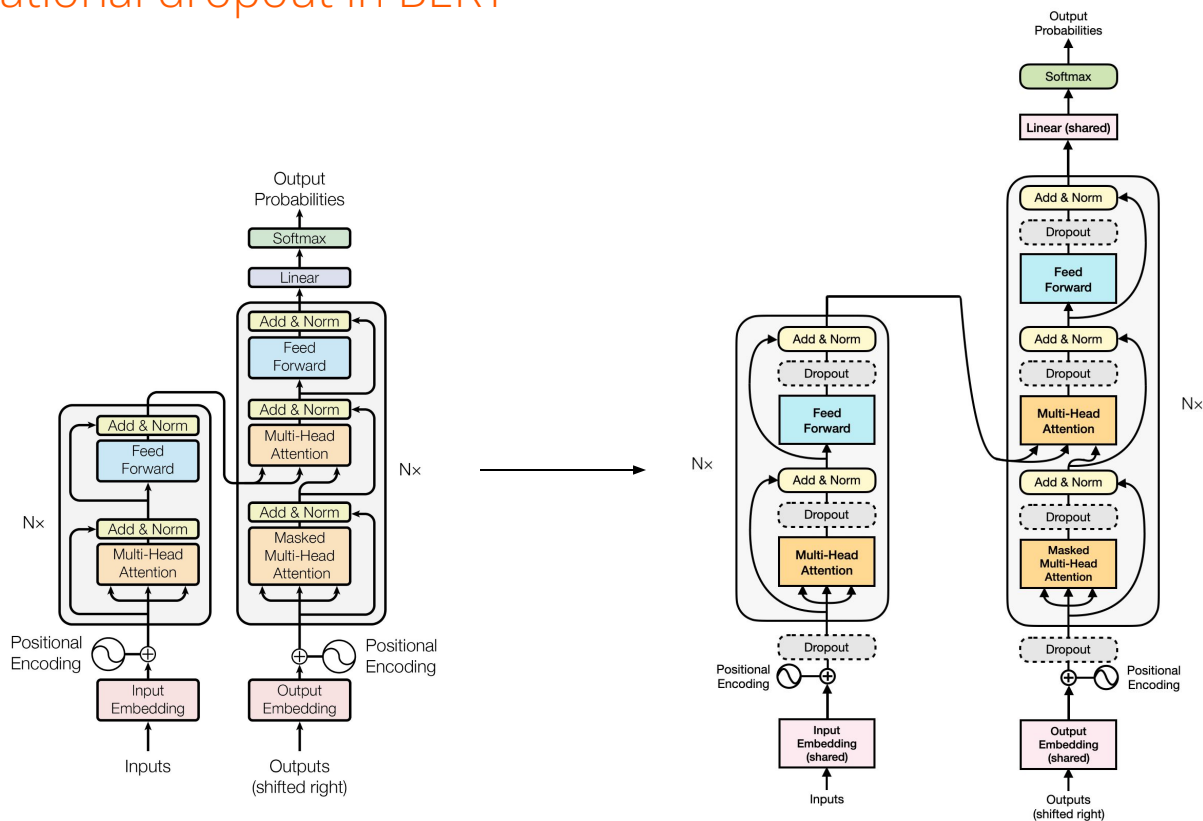


# Variational dropout in BERT





# Variational dropout in BERT



dropout rate = 0.5  
 $T = 20$

## Uncertainty measures

$$u_{SMP} = 1 - \max_{c \in C} \frac{1}{T} \sum_{t=1}^T p_t^c$$

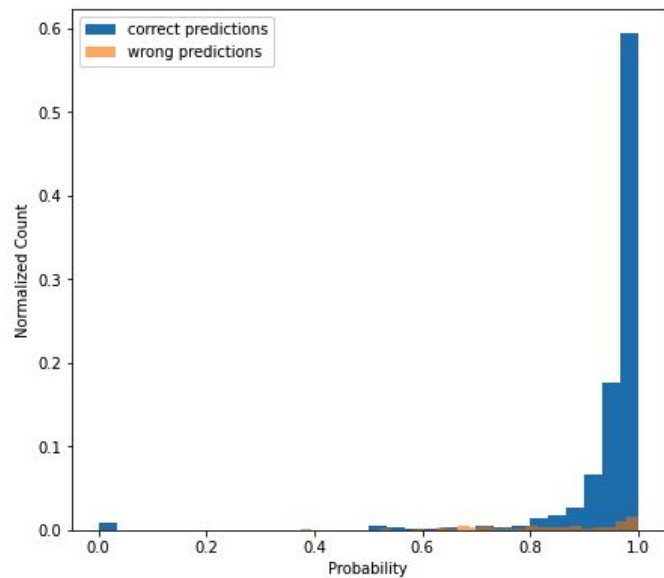
sampled maximum probability

$$u_H = \frac{1}{T} \sum_{c,t} p_t^c \log p_t^c$$

predictive entropy

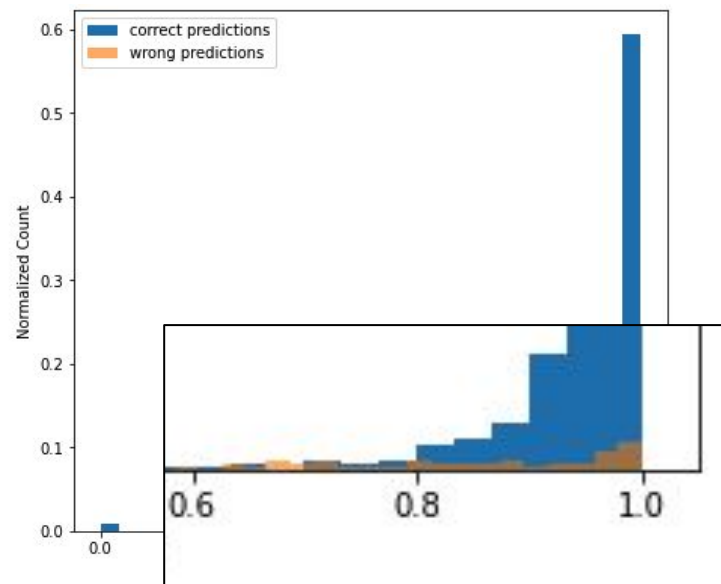


## Variational dropout calibrates NER



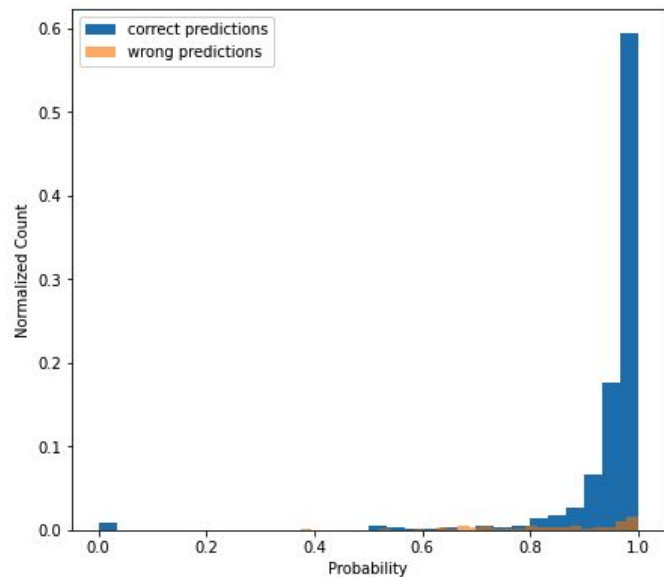
Variational dropout disabled

## Variational dropout calibrates NER

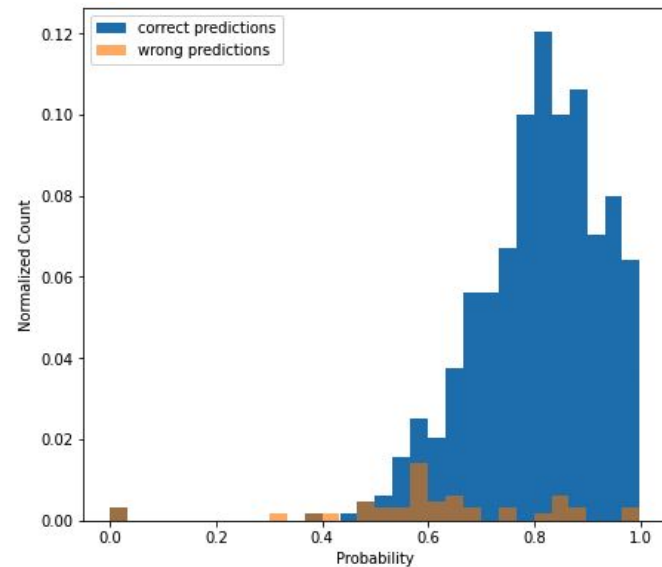


Variational dropout disabled

## Variational dropout calibrates NER

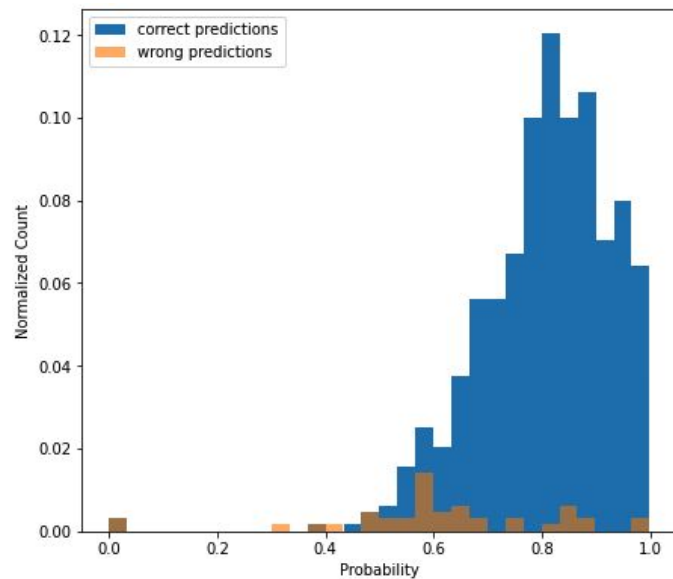


Variational dropout disabled

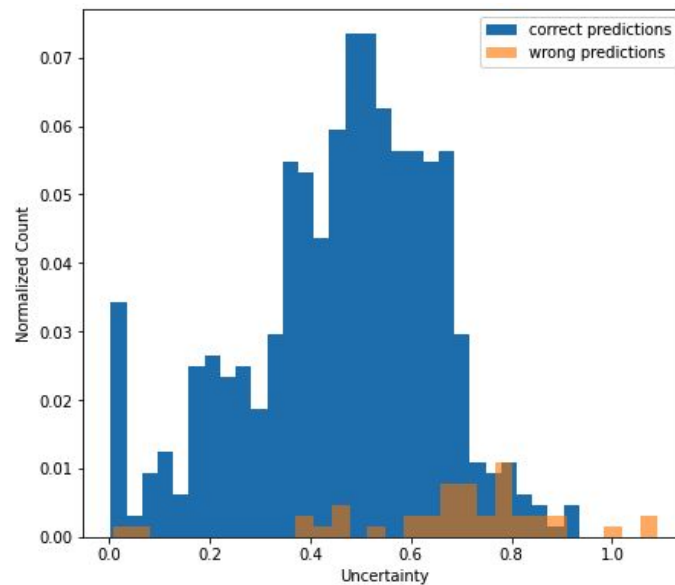


Variational dropout enabled

## Uncertainty identifies misclassified examples

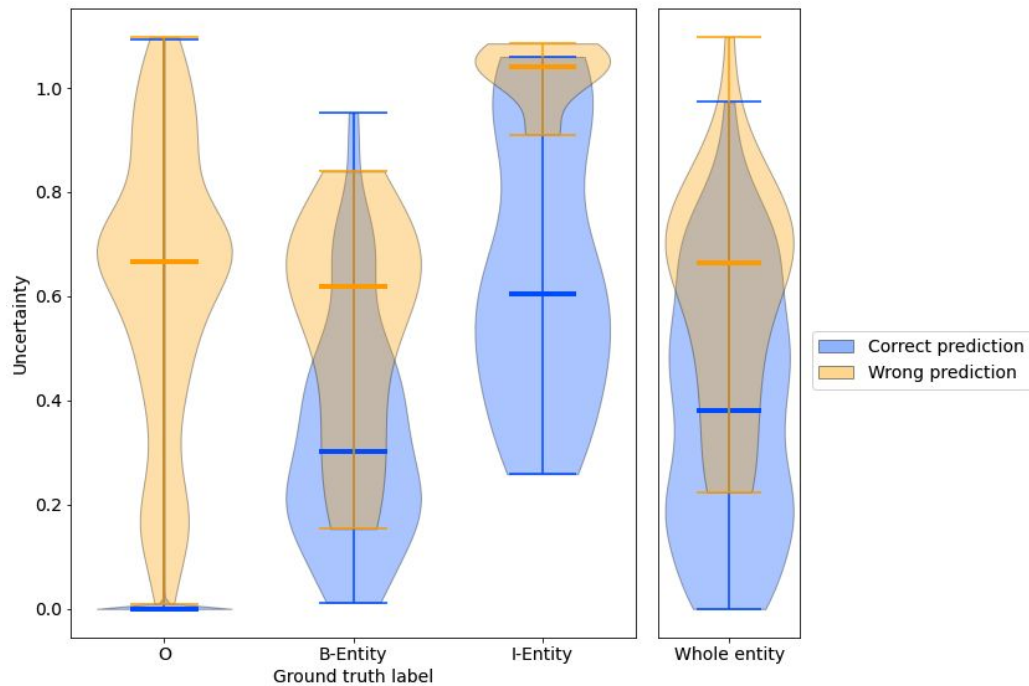


$1 - u_{SMP}$



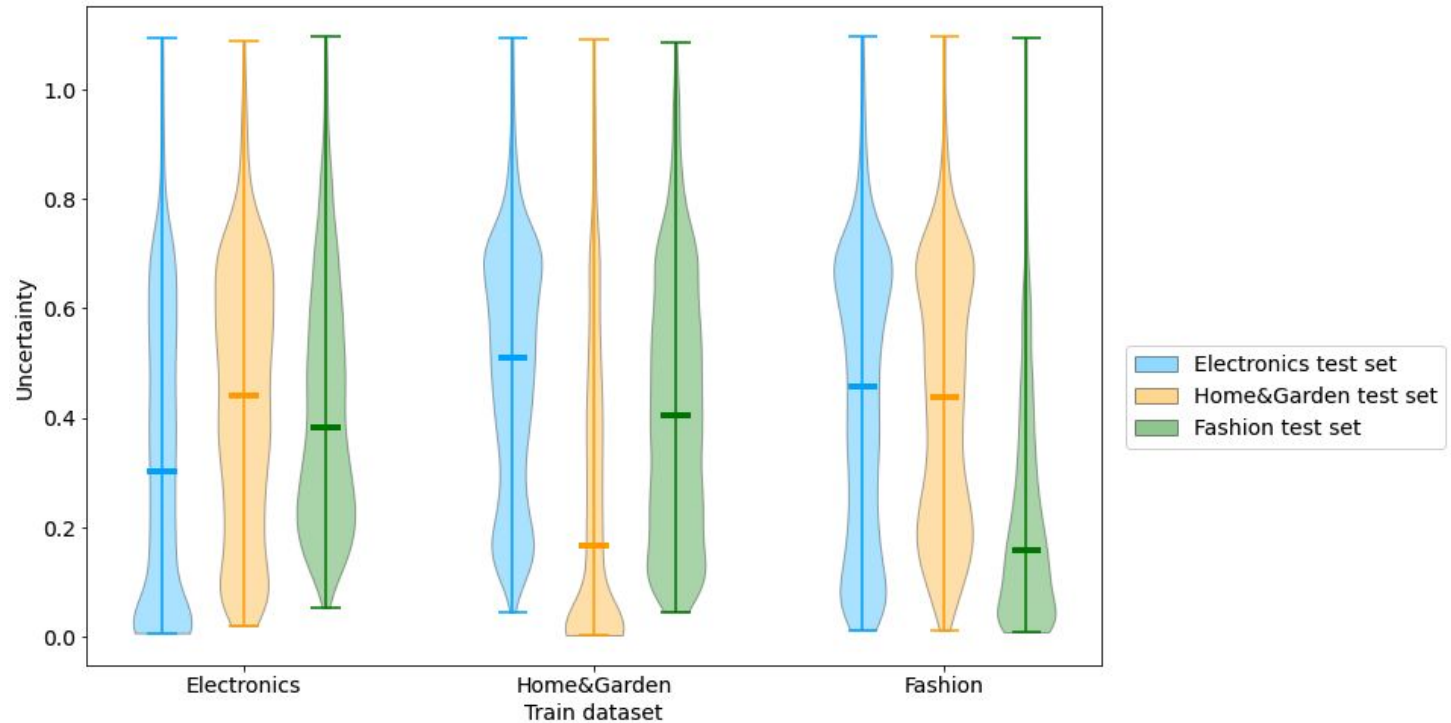
predictive entropy  $u_H$

## Uncertainty identifies misclassified examples





## Uncertainty detects out of distribution examples







# Conclusions

- 
- Named Entity Recognition is an important problem in the e-commerce domain


# Conclusions

- 
- Named Entity Recognition is an important problem in the e-commerce domain
  - Variational dropout enables uncertainty estimation in neural networks


# Conclusions

- 
- Named Entity Recognition is an important problem in the e-commerce domain
  - Variational dropout enables uncertainty estimation in neural networks
  - Variational dropout can be easily utilized with BERT-based models

# Conclusions

- 
- Named Entity Recognition is an important problem in the e-commerce domain
  - Variational dropout enables uncertainty estimation in neural networks
  - Variational dropout can be easily utilized with BERT-based models
  - It improves model calibration in NER

# Conclusions

- 
- Named Entity Recognition is an important problem in the e-commerce domain
  - Variational dropout enables uncertainty estimation in neural networks
  - Variational dropout can be easily utilized with BERT-based models
  - It improves model calibration in NER
  - It allows for misclassification detection and out-of-distribution detection in NER

# Team

**allegro** ML Research

[Our projects](#) [Talks](#) [Blog](#) [Open Source](#) [Publications](#) [Jobs](#)

## About us

Machine Learning Research is Allegro's R&D lab created to develop and apply state-of-the-art machine learning methods, helping Allegro grow and innovate with artificial intelligence. Beyond bringing AI to production, we are committed to advance the understanding of machine learning through open collaboration with the scientific community.

ml.allegro.tech



# Team





Questions?