Tackling the Disagreement Problem with (Un)Certainty

Angel Bujalance, Iason Skylitsis, Zoe Tzifa-Kratira

UNIVERSITY OF AMSTERDAM

Problem

Disagreement Problem: For the same data point, different attribution methods predict different features as being the most important.

Motivation

- A model can be uncertain in its predictions even with a high softmax output. To determine this uncertainty, we need to examine the posterior distribution of the neural network.
- This can be approximated with Monte Carlo Dropout, a tractable alternative to Bayesian inference in deep Gaussian processes.
- MC Dropout during inference simulates an Ensemble Network.
- High uncertainty points to example difficulty.
- What can be said about the disagreement of the attribution methods when the model is very uncertain about the output?

Implications

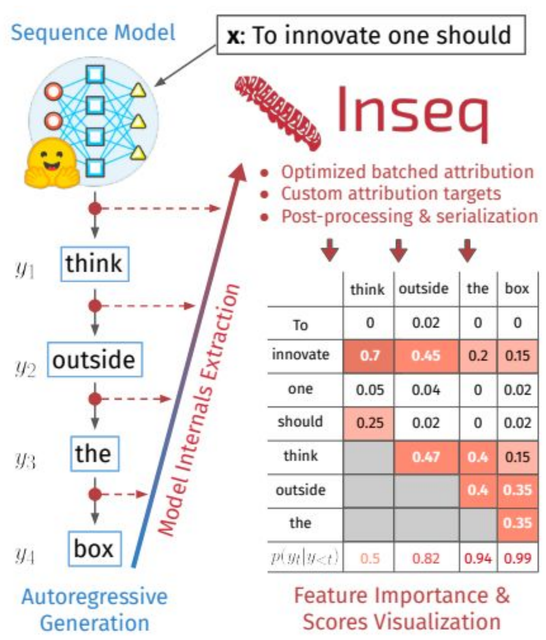
- Aiding interpretability through case-based reasoning.
- Explaining the Attribution Problem and suggesting an improvement.

Our Work

Is there higher disagreement between attribution methods for sentences that are difficult (high uncertainty) than for sentences that are easy (low uncertainty)?

Method

- **Inseq library**
- **Attribution Methods:**
  - **Input x Gradient**
  - **LIME**
  - **Gradient SHAP**
- **Uncertainty Estimation:**
  - **Monte Carlo Dropout**
- **Dataset**
  - **200 sentences from gsarti/wmt_vat: multilingual dataset (we used German)**
- **Translation Model:**
  - **facebook/nllb-200-distilled-600M**



	think	outside	the	box
To	0	0.02	0	0
innovate	0.7	0.45	0.2	0.15
one	0.05	0.04	0	0.02
should	0.25	0.02	0	0.02
think		0.47	0.4	0.15
outside			0.4	0.35
the				0.35
$p(y_t \mid y_{<t})$	0.5	0.82	0.94	0.99

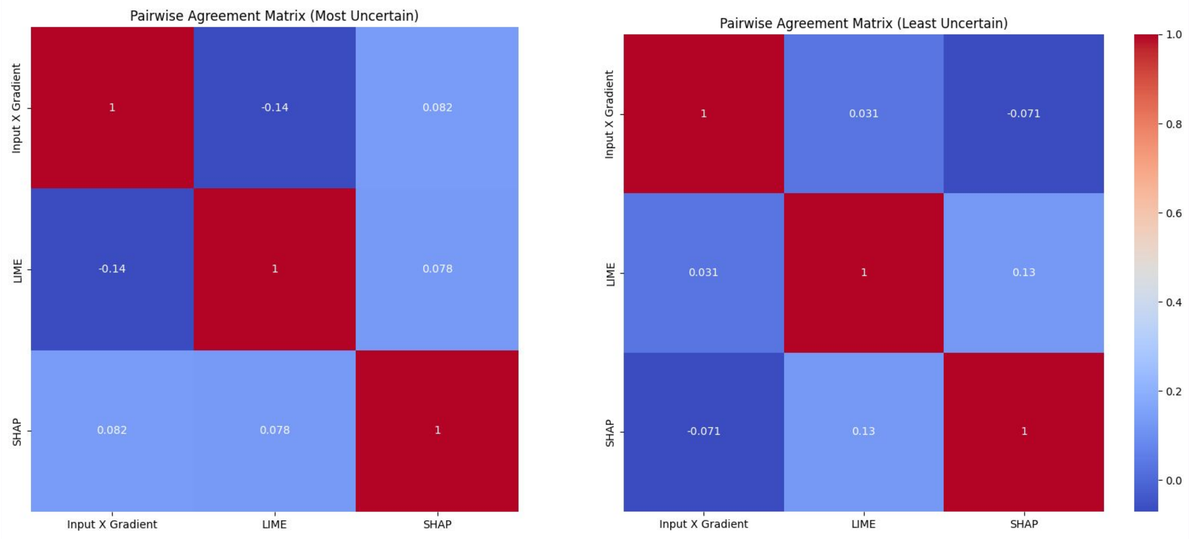
Source Saliency Heatmap
x: Generated tokens, y: Attributed tokens

	__El	__gerente	__llamó	__a	__la	__secretaria	__a	__la	__oficina	.	</s>
__The	0.098	0.052	0.044	0.037	0.037	0.019	0.023	0.029	0.02	0.031	0.074
__manager	0.363	0.465	0.099	0.09	0.078	0.04	0.061	0.071	0.074	0.072	0.113
__called	0.11	0.068	0.195	0.058	0.072	0.042	0.083	0.068	0.042	0.053	0.118
__the	0.063	0.059	0.077	0.094	0.076	0.07	0.038	0.037	0.031	0.032	0.057
__secretary	0.128	0.129	0.182	0.411	0.369	0.47	0.135	0.12	0.127	0.109	0.162
__into	0.044	0.034	0.097	0.05	0.068	0.051	0.114	0.056	0.06	0.047	0.051
__the	0.035	0.034	0.046	0.064	0.051	0.05	0.046	0.055	0.052	0.031	0.04
__office	0.072	0.055	0.072	0.079	0.074	0.056	0.118	0.173	0.292	0.115	0.059
.	0.04	0.034	0.042	0.041	0.035	0.018	0.03	0.033	0.033	0.06	0.059
</s>	0.046	0.047	0.039	0.061	0.049	0.039	0.027	0.028	0.036	0.034	0.042
probability	0.776	0.516	0.714	0.703	0.884	0.831	0.534	0.869	0.936	0.889	0.895

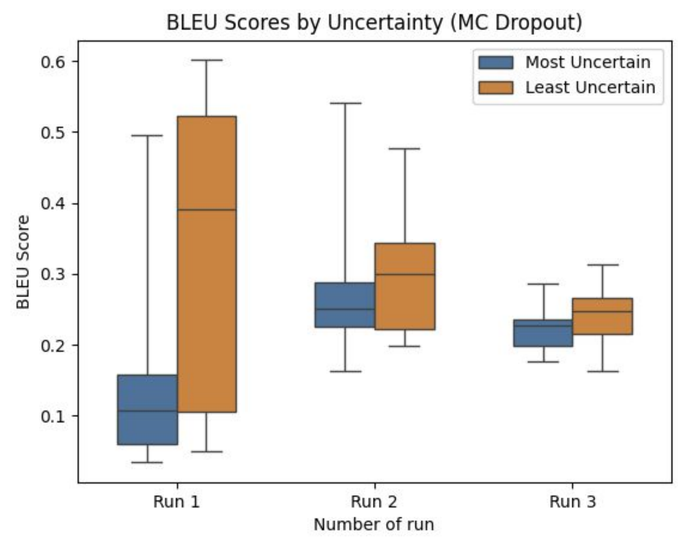
- **Evaluation Methods:**
  - **BLEU**
  - **Rank Correlation Matrix (Spearman's correlation)**

Results

Attribution Methods Agreement Matrix



Translation Quality Quantitative Evaluation



Attributing a Difficult Example

Source Saliency Heatmap
x: Generated tokens, y: Attributed tokens

	</s>eng_Latn	__And	__also	__the	__Albachtener	__machten	__damals	__to	__their	__Unmut	__Luft.	</s>
<unk>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
__Und	0.114	0.245	0.179	0.069	0.068	0.08	0.067	0.057	0.04	0.043	0.092	0.067
__auch	0.141	0.172	0.162	0.141	0.072	0.076	0.064	0.058	0.048	0.056	0.069	0.081
__die	0.111	0.088	0.085	0.276	0.065	0.064	0.056	0.042	0.037	0.039	0.091	0.044
__Albachtener	0.122	0.096	0.072	0.142	0.261	0.07	0.064	0.058	0.051	0.044	0.098	0.069
__machten	0.165	0.065	0.088	0.095	0.116	0.256	0.109	0.091	0.081	0.041	0.05	0.084
__damals	0.105	0.053	0.054	0.052	0.048	0.136	0.242	0.08	0.135	0.049	0.045	0.071
__ihrem	0.079	0.08	0.09	0.055	0.069	0.059	0.139	0.292	0.198	0.091	0.069	0.045
__Unmut	0.077	0.068	0.039	0.038	0.033	0.037	0.057	0.064	0.062	0.435	0.109	0.069
__Luft.	0.086	0.08	0.064	0.06	0.05	0.056	0.067	0.067	0.049	0.114	0.229	0.095
</s>	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
mc_dropout_prob_avg	0.551	3.581	-0.127	1.139	-0.63	0.184	-4.722	0.864	1.164	0.214	0.386	-0.988

Target: "*And also the people of Albachten demonstrated their anger*."

Key Findings

- Uncertainty correlates with the quality of the translation. If the model is uncertain about the output translation, it is more likely that the translation is of lower quality.
- The Disagreement Problem is more prominent when there is High Uncertainty in the model's outputs.
- Different Attribution methods tend to agree more when the model's outputs have been produced with certainty.
- When it comes to difficult examples, we observe that the attribution methods are failing to attend to the surrounding relevant tokens.

Limitations

- We used 5 iteration steps in MC Dropout.
- More Agreement metrics could have been explored given small values.

References

- Gal, Y., & Ghahramani, Z. (2016, June). **Dropout as a bayesian approximation: Representing model uncertainty in deep learning.** In international conference on machine learning (pp. 1050-1059). PMLR.
- Sarti, G., Feldhus, N., Sickert, L., Van Der Wal, O., Nissim, M., & Bisazza, A. (2023). **Inseq: An interpretability toolkit for sequence generation models.** arXiv preprint arXiv:2302.13942.
- Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). **The disagreement problem in explainable machine learning: A practitioner's perspective.** arXiv preprint arXiv:2202.01602.