Attention Meets Post-hoc Interpretability: A Mathematical Perspective

Gianluigi Lopardo,[1] Frederic Precioso,[1] Damien Garreau[2]

[1]Université Côte d'Azur - Inria
[2]Julius-Maximilians Universität Würzburg - CAIDAS

January 8, 2025

Where is Würzburg?

- **Figure:** (left panel) Würzburg is located in Northern Bavaria (right panel) Festung Marienberg (credits google maps / wikipedia)

Outline

1. Transformers

2. Post-hoc interpretability
   - Gradient-based approaches
   - Perturbation-based approaches
   - Attention weights

3. Analysis on a simple model

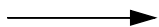
4. Conclusion

1. Transformers

Introduction

- **Context = natural language processing:** from text input $x \in \mathcal{X}$, predict $y \in \mathcal{Y}$ as $f(x)$
- **Running example:** sentiment analysis:

the selection on the menu is great
and so is the food the service
is not bad prices are fine



positive

- $f = f_\theta$ corresponds to some architecture choice, and $\theta \in \Theta$ to the parameters
- $\theta^\star$ = good parameter learned from data
- **State-of-the-art toady:** f_θ = attention[1]-based model (a transformer[2])
- **Goal of this section:** describe a simple transformer architecture

[1] Bahdanau, Cho, Bengio, *Neural machine translation by jointly learning to align and translate*, ICLR, 2025
[2] Vaswani et al., *Attention is all you need*, NeurIPS, 2017

Tokens

- **Notation:** $\xi \in \mathcal{X}$ is a document
- encoded for the computer as a sequence of tokens
- we identify tokens with elements of $\{1, \ldots, D\} = [D]$
- **Several possibilities in practice:**
  - individual letters ($D = 100$)
  - words ($D = 100.000$)
  - in-between (*e.g.*, BERT[3] uses WordPiece,[4] $D = 30.000$)
- **Example:**

$$\text{"DATAIA"} \quad \longmapsto \quad [4, 1, 20, 1, 9, 1]$$

- **Special tokens:** `<UNK>`, `<CLS>`, etc.

[3] Devlin et al., *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, ACL Proc., 2019

[4] Wu et al., *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, preprint, 2016

Token embeddings

- **Summary from last slide:** $\xi =$ ordered sequence of tokens $\xi_1, \ldots, \xi_T \in [D]$
- **Next step:** vector representation of each token
- for each $t \in [T]$, token $\xi_t = j$ is embedded as

$$e_t := (W_e)_{:,j} + W_p(t) \in \mathbb{R}^{d_e},$$

where:

- $W_e \in \mathbb{R}^{d_e \times D}$ matrix containing embeddings of all tokens
- $W_p : \mathbb{N} \to \mathbb{R}^{d_e}$ positional embedding

- **Typically,** W_e is learned and W_p is set to something arbitrary
- **Example:**

$$\begin{cases} W_p(t)_{2i} & = \cos(t/T_{\max}^{2i/d_e}) \\ W_p(t)_{2i-1} & = \sin(t/T_{\max}^{2i/d_e}). \end{cases}$$

Keys, queries, values

- **Note:** I am following

  Phuong and Hutter, *Formal Algorithms for Transformers*, preprint, 2022

- **Padding** until $T_{\max}$ with <UNK> token, embedded to $h + W_p(t)$
- **Next step:** for all $t \in [t]$, map embeddings to

$$\begin{cases} k_t & := W_k e_t + b_k \in \mathbb{R}^{d_{\text{att}}} & \text{(key)} \\ q_t & := W_q e_t + b_q \in \mathbb{R}^{d_{\text{att}}} & \text{(query)} \\ v_t & := W_v e_t + b_v \in \mathbb{R}^{d_{\text{out}}} & \text{(value)} \end{cases}$$

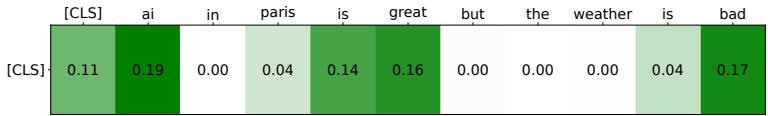
- matrices $W_k, W_q \in \mathbb{R}^{d_{\text{att}} \times d_e}$, $W_v \in \mathbb{R}^{d_{\text{out}} \times d_e}$ learned parameters
- bias vectors $b_k, b_q \in \mathbb{R}^{d_{\text{att}}}$, $b_v \in \mathbb{R}^{d_{\text{out}}}$ also learnable, set to zero for simplicity

Attention mechanism

- for a given query $q \in \mathbb{R}^{d_{\text{att}}}$, index t receives *attention*

$$\forall t \in T_{\max}, \qquad \alpha_t := \frac{\exp\left(q^\top k_t / \sqrt{d_{\text{att}}}\right)}{\sum_{u=1}^{T_{\max}} \exp\left(q^\top k_u / \sqrt{d_{\text{att}}}\right)},$$

  softmax of the vector $(q^\top k_1, \ldots, q^\top k_{T_{\max}})^\top$
- **Intuition:** if query "matches" with k_t, then α_t large
- this mechanism is at the core of the transformer architecture



Final output

- for a given query q, output

$$\tilde{v} := \sum_{s=1}^{T_{\max}} \alpha_s v_s \in \mathbb{R}^{d_{\text{out}}} .$$

- **In our simplified setting**, q corresponds to the <CLS> token, $W_\ell \in \mathbb{R}^{1 \times d_{\text{out}}}$, and

$$\boxed{f(x) := W_\ell \tilde{v} \in \mathbb{R} .}$$

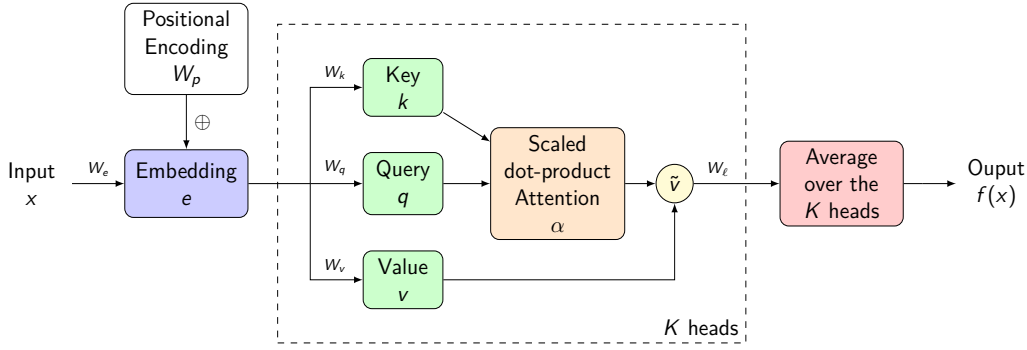
- **Remark:** we can deal with several heads:

$$\forall i \in [K], \qquad \tilde{v}^{(i)} := \sum_{s=1}^{T_{\max}} \alpha_s^{(i)} v_s^{(i)} \in \mathbb{R}^{d_{\text{out}}} ,$$

and

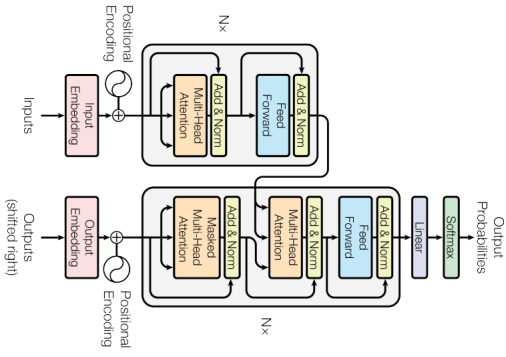
$$f(x) := \frac{1}{K} \sum_{i=1}^{K} W_\ell^{(i)} \tilde{v}^{(i)} \in \mathbb{R} .$$

Our model, in pictures



In practice

- **In reality,** self-attention, meaning each token associated to value $\tilde{v}_t$
- then layer-norm + small perceptron, several layers

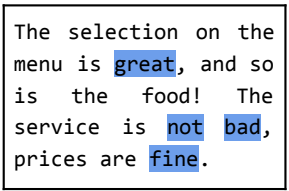


- **Figure:** transformer architecture from Vaswani et al. (2017)

2. Post-hoc interpretability

A (very brief) introduction to interpretability

- **Context:** automated systems are reaching human-level in many applications
- **Problem:** sometime performance is not the only metric (especially in critical applications)
- **Interpretability methods:** give insights to why a specific decision was taken
- this talk = local, post-hoc (single example, model already trained)
- **Example:** sentiment analysis, outline words that are important for the decision



- **Figure:** Anchors outlining words supporting a positive prediction

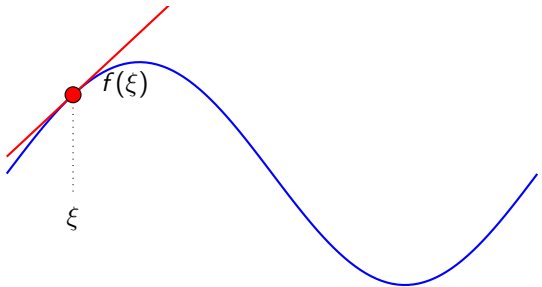
2.1. Gradient-based approaches

Gradients

- **Simple idea:** take the gradient of prediction with respect to input
- **Intuition:** if feature is (positively) important, positive partial derivative
- **Problem:** documents are *discrete* objects
- **Solution:** rewrite f as a function of the embeddings, *i.e.*,

$$f(x) = g(e_1(x), e_2(x), \ldots, e_{T_{\max}}(x)) .$$

- gradient-based approaches compute, for each token, $\nabla_{e_t} g \in \mathbb{R}^{d_e}$



Token-level measure of importance

- **Problem:** still complicated object, need to come back to token level
- several competing approaches:
  - take the mean (G-AVG)[5]
  - take the L^1 norm (G-L1)[6]
  - take the L^2 norm (G-L2)[7]
  - take the dot-product with e_t (G×I)[8]
  - ...
- **Example:** L^2 norm of the token gradients:

attention based explanations are popular but questionable

[5] Atanasova et al., *A diagnostic study of explainability techniques for text classification*, EMNLP, 2020

[6] Li et al., *Visualizing and understanding neural models in NLP*, ACL Proc., 2016

[7] Poerner et al., *Evaluating neural network explanation methods using hybrid documents and morphosyntactic agreement*, ACL Proc., 2018

[8] Denil et al., *Extraction of salient sentences from labelled documents*, preprint, 2014

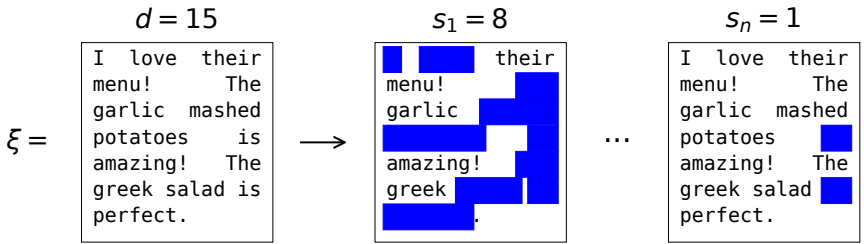
2.2. Perturbation-based approaches

Perturbation-based approaches

- **Idea:** remove words at random and look at how the prediction varies
- **Example:** LIME[9]
- recall $\xi = (\xi_1, \ldots, \xi_T) \in [D]^T$ document, f our model
- local dictionary $[d]$ with $d < D$
- **Step 1:** create n *perturbed samples* $X_1, \ldots, X_n$ by removing s words at random
- s follows uniform distribution on $[d]$
- the subset S of words to be removed is chosen uniformly
- words are removed with repetition

[9] Ribeiro et al., *"Why should i trust you?" Explaining the predictions of any classifier*, SIGKDD, 2016

Sampling



- **Figure:** LIME sampling on a small example

Weights

- **Step 2:** give positive weights to the samples
- define $Z_i \in \{0,1\}^T$ as presence / absence of words
- $\mathbb{1}$ corresponds to the original document
- weights are defined by

$$\forall i \in [n], \qquad \pi_i := \exp\left(\frac{-\delta(\mathbb{1}, Z_i)^2}{2\nu^2}\right),$$

with $\nu > 0$ bandwidth parameter and δ the *cosine distance*

$$\delta(a,b) := 1 - \frac{a^\top b}{\|a\| \cdot \|b\|}.$$

- **Intuition:** if perturbed sample far from ξ, $\delta(\mathbb{1}, X_i) \approx 1$, assign small weight

Local surrogate model

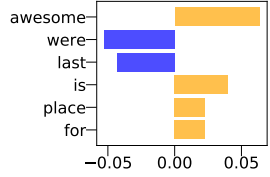
- **Step 3:** local surrogate model
- fit linear model on absence / presence of words:

$$\hat{\beta}_n \in \underset{\beta \in \mathbb{R}^d}{\arg\min} \left\{ \sum_{i=1}^{n} \pi_i (Y_i - \beta^\top Z_i)^2 + \lambda \|\beta\|^2 \right\},$$

where $Y_i = f(X_i)$ and $\lambda > 0$

Explaining a prediction with LIME

Update! Went back last night for dinner, this place is still awesome. I had the Las Vegas Rolls, they were pure deep fried goodness.



- **Figure:** visualizing LIME output

2.3. Attention weights

Attention weights

- **Another idea:** look directly at attention weights[10]
- **In our setting,** attention wrt <CLS> token $\Rightarrow$ look at attention weight of individual tokens
- **What happens with several heads?**
  - either take the average

$$\alpha\text{-avg}_t := \frac{1}{K}\sum_{i=1}^{K}\alpha_t^{(i)},$$

  - or the max[11]

$$\alpha\text{-max}_t := \max_{i\in[K]}\alpha_t^{(i)}.$$

- if several layers, further aggregation scheme required[12]

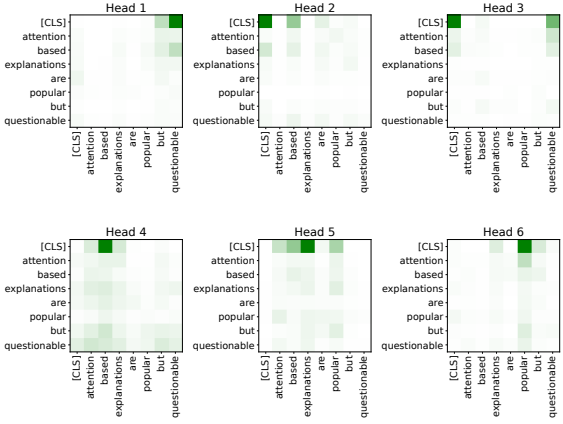
---

[10] Clark et al., *What does BERT look at? An analysis of BERT's attention*, Blackbox @ EMNLP, 2019

[11] Schwenke and Atzmueller, *Show me what you're looking for: visualizing abstracted transformer attention for enhancing their local interpretability on time series data*, FLAIRS, 2021

[12] Mylonas et al., *An attention matrix for every decision: faithfulness-based arbitration among multiple attention-based interpretations of transformers in text classifcation*, Data Mining and Knowledge Discovery, 2024

Post-hoc interpretability III: attention, in pictures



- **Figure:** attention patterns inside a single-layer transformer

Is attention explanation?

- tempting to rely on these coefficients: they are *really* used by the model
- **But,** some dissident voices:[13]
  - if attention is explanation, attention coefs should correlate with feature importance
  - counterfactual attention weight configuration should change prediction
- the debate is not settled
- there are criticisms regarding experimental setting of Jain and Wallace[14]
- not many theoretical contributions
- related work show that single-layer attention models can get near-perfect accuracy with un-informative attention pattern[15,16]

---

[13] Jain and Wallace, *Attention is not explanation*, NAACL Proc., 2019

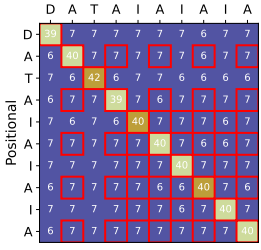
[14] Wiegreffe and Pinter, *Attention is not not Explanation*, EMNLP, 2019

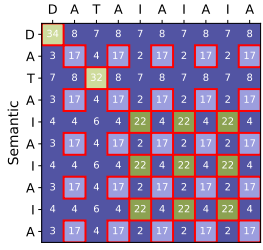
[15] Wen et al., *Transformers are uninterpretable with myopic methods: a case study with bounded Dyck grammars*, NeurIPS, 2024

[16] Cui, Behrens, Krzakala, Zdeborová, *A phase transition between positional and semantic learning in a solvable model of dot-product attention*, preprint, 2024

Is attention explanation?, ctd.

- **Histogram task:** count number of times token appears in the sequence[17]
- **Example:** "DATAIAIAIA" $\longmapsto$ $[1, 5, 1, 5, 3, 5, 3, 5, 3, 5]$
- **Architecture:** single-layer with tied weights
- two *vastly* different local minima found, one with un-informative attention pattern





- figure obtained running code from Cui et al., 2024

[17] Weiss, Goldberg, Yohav, *Thinking like transformers*, ICML, 2021

3. Analysis on a simple model

Summary so far

- **Many** different methods providing explanations
- different results even on single-layer transformer:

G-avg:	attention	based	explanations	are	popular	but	questionable
G-l1:	attention	based	explanations	are	popular	but	questionable
G-l2:	attention	based	explanations	are	popular	but	questionable
G$\times$I:	attention	based	explanations	are	popular	but	questionable
lime:	attention	based	explanations	are	popular	but	questionable
α-avg:	attention	based	explanations	are	popular	but	questionable
α-max:	attention	based	explanations	are	popular	but	questionable

- **Figure:** different explainers yield different explanations
- **Our work:** what should we use?
- starting point = attention coefficients = α_t
- **Problem:** no dependency in the linear layer / values (!)

Gradient-based: closed-form expression

- **Recall:** we are looking at $\nabla_{e_t} g$
- straightforward computations yield:

Theorem (Lopardo, Precioso, G., '24): The gradient of our simple model with respect to token e_t is given by

$$\nabla_{e_t} g(x) = \alpha_t W_v^\top W_\ell^\top + \frac{\alpha_t}{\sqrt{d_{\text{att}}}} W_\ell \left(v_t - \sum_{s=1}^{T_{\max}} \alpha_s v_s \right) W_k^\top q \in \mathbb{R}^{d_e}.$$

- **Main insight:** token contributes if $\alpha_t \neq 0$ and v_t deviates from average
- **Remark (i):** easy corollary for K heads by linearity
- **Remark (ii):** straightforward derivations for average, L^1 and L^2 norms, etc.

Additional notation

- much more challenging analysis for LIME
- set h the index for the <UNK> token
- corresponding key / value vectors for <UNK> token at position t are

$$\begin{cases} k_{h,t} & := W_k h + W_k W_p(t) \in \mathbb{R}^{d_{\text{att}}} \\ v_{h,t} & := W_v(h + W_p(t)) \in \mathbb{R}^{d_{\text{out}}} . \end{cases}$$

- define further

$$g_{h,t} := \exp\left(q^\top k_{h,t} / \sqrt{d_{\text{att}}}\right) ,$$

and

$$\alpha_{h,t} := \frac{g_{h,t}}{\sum_u g_{h,u}} .$$

- **Intuition:** attention coefficient for all <UNK> tokens

Perturbation-based, main result

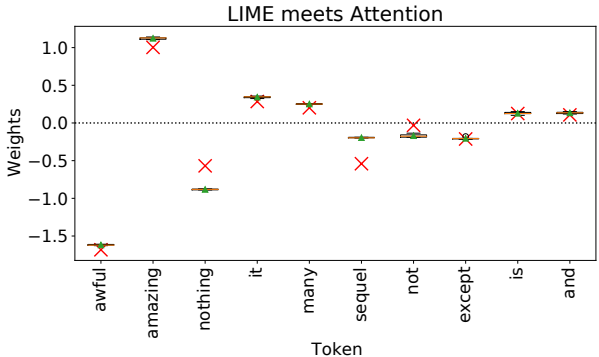
- with these notation:

Theorem (Lopardo, Precioso, G., '24): Assume that $d = T = T^{\varepsilon}_{\max}$, with $\varepsilon \in (0,1)$. Assume further that there exist positive constants $0 < c < C$ such that, as $T \to +\infty$, for all $t \in [T_{\max}]$, $\max(|v_t|, |v_{h,t}|) \leq C$, and $c \leq \min(g_t, g_{h,t}) \leq C$.

$$\forall j \in [d], \qquad \beta_j^{\infty} \approx \frac{3}{2} \sum_{t=1}^{T_{\max}} W_{\ell} \left(\alpha_t v_t - \alpha_{h,t} v_{h,t} \right) \mathbb{1}_{\xi_t = j} \, .$$

- approximate expression of LIME coefficients for a single-layer transformer
- **Main insight:** token contributes if $\alpha_t v_t$ deviates from "average"
- **Remark:** straightforward extension to several heads

Experimental check



- **Figure:** boxplots = 5 runs of LIME, red crosses = approximation. $T = d = 99$ in this example

Sketch of proof (i)

- we use previous work doing the analysis in the asymptotic setting:[18]

Theorem (Mardaoui, G., '21): take $\lambda = 0$, assume f is bounded, then $\hat{\beta} \xrightarrow{\mathbb{P}} \beta^f$, where β^f is defined as

$$\forall j \in [d], \quad \beta_j^f = c_d^{-1} \left\{ \sigma_1 \mathbb{E}\left[\pi f(X)\right] + \sigma_2 \mathbb{E}\left[\pi Z_j f(X)\right] + \sigma_3 \sum_{\substack{k=1 \\ k \neq j}}^{d} \mathbb{E}\left[\pi Z_k f(X)\right] \right\}.$$

Here, X has the distribution of the perturbed document described previously, and c_d, $\sigma_1, \sigma_2, \sigma_3$ have explicit expressions.

- **Intuition:** weighted least squares $\rightarrow$ closed-form

[18] Mardaoui and Garreau, *An analysis of LIME for text data*, AISTATS, 2021

Sketch of proof (ii)

- in the large bandwidth regime, expression simplifies somewhat:

Corollary (Mardaoui and G., '21): same assumptions, $\nu \to +\infty$, then β_j^f converges to

$$\beta_j^\infty = 3\mathbb{E}\left[f(X) \mid j \notin S\right] - \frac{3}{d}\sum_k \mathbb{E}\left[f(X) \mid k \notin S\right],$$

where S is the random set defined in the sampling scheme.

- very challenging to deal with this expectation (f non-linear and complicated distribution)
- we resort to *approximations*

Sketch of proof (iii)

- **Crux of the proof:** approximate

$$\mathbb{E}\left[f(X) \mid j \notin S\right] = \mathbb{E}\left[\sum_{t=1}^{T_{\max}} A_t V_t \mid \ell \notin S\right] = \sum_{t=1}^{T_{\max}} \mathbb{E}\left[\frac{G_t V_t}{\sum_{u=1}^{T_{\max}} G_u} \mid \ell \notin S\right],$$

  where A_t and V_t are the *random* version of attention / values
- **Proof technique:** split expectation according to $|S| = s$, then approximate each piece using the following:

Lemma: Let X and Y be two random variables with finite variance. Assume that there exist two positive constants c and C such that $|X| \leq C$ and $cn \leq Y \leq Cn$ a.s. Then

$$\left|\mathbb{E}\left[\frac{X}{Y}\right] - \frac{\mathbb{E}[X]}{\mathbb{E}[Y]}\right| \leq \frac{C\mathrm{Var}(Y)}{c^3 n^3} + \frac{C^2\sqrt{\mathrm{Var}(Y)}}{c^2 n^2}.$$

Sketch of proof (iv)

- finally, computation of expectation and variance of

$$H_S := \sum_i \{a_i \mathbb{1}_{i \notin S} + b_i \mathbb{1}_{i \in S}\} ,$$

conditionally to $|S| = s$ and $\ell \notin S$

Lemma: Let H_s be as before, then

$$\mathbb{E}_s [H_S | \ell \notin S] = \frac{n-1-s}{n-1} \sum_i a_i + \frac{s}{n-1} \sum_i b_i + \frac{s}{n-1}(a_\ell - b_\ell) ,$$

and

$$\mathrm{Var}_s(H_S \mid \ell \notin S) = \frac{ns(n-s-1)}{(n-1)(n-2)} \left[\widehat{\mathrm{Var}}\,(a-b) - \frac{1}{n-1} \left(a_\ell - b_\ell - (\overline{a} - \overline{b}) \right)^2 \right] .$$

4. Conclusion

Conclusion

- **Summary:**
  - single-layer attention-based model
  - closed-form or exact approximations for token-importance measure
  - methods are very un-alike no clear recommended choice
- **Main reference:**
  - Lopardo, Precioso, Garreau,*Attention Meets Post-hoc Interpretability: A Mathematical Perspective*, ICML, 2024
- **Future directions:**
  - Anchors[19] meets attention (existing theoretical framework[20])
  - more realistic architecture (skip connection, non-linearities, more layers)

[19]Ribeiro, Singh, Guestrin, *Anchors: High-precision model-agnostic explanations*, AAAI, 2018
[20]Lopardo, Precioso, Garreau, *A sea of words: an in-depth analysis of anchors for text data*, AISTATS, 2023

Thank you for your attention!