Anatomia degli LLMs

5. Analisi e Interpretabilità dei Transformers

Alessio Miaschi
ItaliaNLP Lab, Istituto di Linguistica Computazionale (CNR-ILC), Pisa
alessio.miaschi@ilc.cnr.it
https://alemiaschi.github.io/
http://www.italianlp.it/alessio-miaschi/

Interpretability dei LLMs

- Lo sviluppo di Large Language Models (LLMs) allo stato dell'arte comporta un costo in termini della loro **interpretabilità**, poiché modelli basati su reti neurali offrono poca trasparenza sul loro funzionamento interno e sulle loro capacità

Obiettivi:

- Comprendere la natura degli AI systems → comprendere ciò che influenza il processo decisionale di un modello
- Responsabilizzare gli utenti dei sistemi di IA → trarre intuizioni/scelte a partire dalle risposte del sistema

Interpretability in NLP

“In the context of NLP, this question needs to be understood in light of earlier NLP work. [...] In some of these systems, features are more easily understood by humans. [...] In contrast, it is more difficult to understand what happens in an end-to-end neural network model that takes input (say, word embeddings) and generates an output.”

Belinkov and Glass, Analysis Methods in Neural Language Processing: A Survey (2019). In Transactions of ACL, Volume 7, pages 49-72.

Domande di ricerca:

- Cosa succede in una rete neurale addestrata a risolvere il task di language modeling?
- Che tipo di competenze implicite (i.e. features) vengono codificate implicitamente nelle rappresentazioni di tali modelli?
- Esiste una relazione fra tali competenze e l’abilità dei modelli nel risolvere specifici task?

Interpretability in NLP

- L'analisi del funzionamento interno dei NLMs è diventata una delle linee di ricerca più affrontate nel contesto di studi di NLP

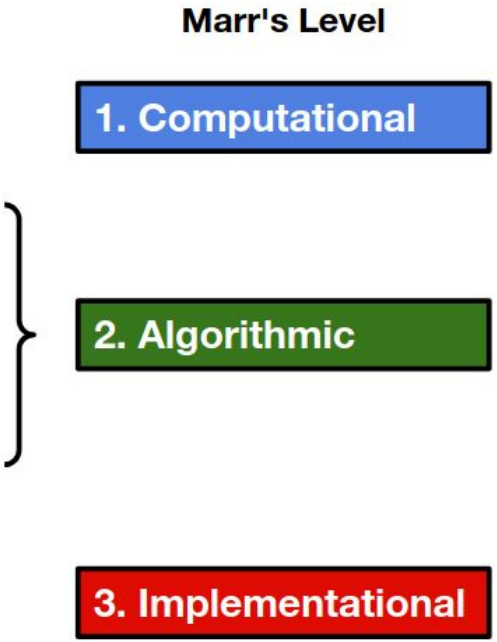
- Diversi metodi sono stati sviluppati al fine di ottenere spiegazioni significative e per capire come questi modelli siano in grado di catturare implicitamente specifici fenomeni linguistici e non

- Diversi approcci:
  - Test comportamentali (es. Goldberg, 2019)
  - Probing tasks (es. Hewitt e Manning, 2019; Pimentel et al., 2020);
  - Analisi dei meccanismi di attenzione (es. Clark et al., 2019);
  - Feature Attribution Methods (es. Ramnath, 2020);
  - Mechanistic Interpretability (es. Elhage et al., 2021).

Diversi livelli di Interpretability

Levels of explanation *granularity*:

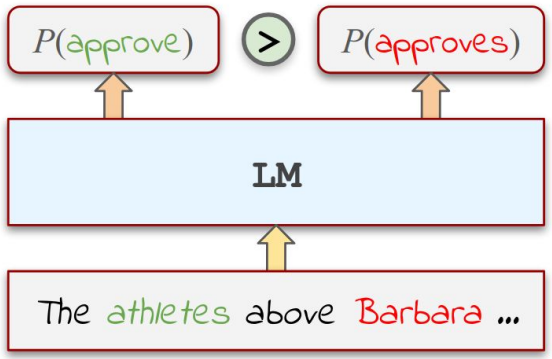
1. **Behavioural**
   - How does the model behave on certain phenomena?
2. **Attributional**
   - Which input features were most *important* for a prediction?
3. **Probing**
   - What *abstract features* are encoded by the model?
4. **Mechanistic**
   - Can we identify specific *circuits* responsible for a particular behaviour?



Da Tutorial EACL 2024 Transformer-specific Interpretability: https://github.com/interpretingdl/eacl2024_transformer_interpretability_tutorial/tree/main

Behavioural Interpretability

- Come si può comprendere meglio il funzionamento di un NLM senza aprire la “*black box*”?
- Approccio:
  - Creare **minimal pairs** per studiare il comportamento del modello di fronte ad uno (o più) fenomeni specifici
  - Tale approccio richiede solamente l’accesso alle **output probabilities** del modelo



Assessing BERT's Syntactic Abilities (Goldberg. 2019)

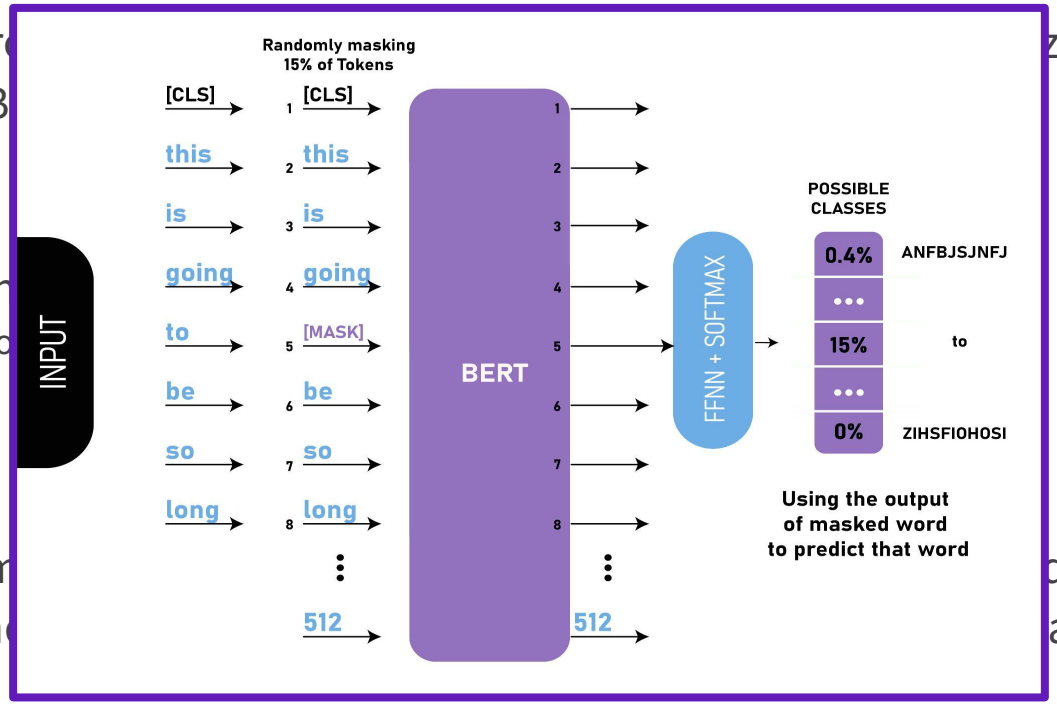
- Approccio proposto da Goldberg (2019) per studiare le competenze linguistiche implicite di BERT

- Due fenomeni linguistici:
  - Accordo soggetto-verbo;
  - Anafora.

- **Approccio**: mascherare delle parole target e chiedere al modello di predirle ("fill in the gap"), andando ad estrarre le parole con la probabilità più alta

Assessing BERT's Syntactic Abilities (Goldberg. 2019)

- Approccio pr… …ze linguistiche implicite di B…
- Due fenomen…
  - Accordo so…
  - Anafora.
- **Approccio**: m… …di predirle ("fill in the gap"), an… …a



Assessing BERT's Syntactic Abilities (Goldberg. 2019)

the game that the guard hates is bad

Assessing BERT's Syntactic Abilities (Goldberg. 2019)

the game that the guard hates **[MASK]** bad

Assessing BERT's Syntactic Abilities (Goldberg. 2019)

the game that the guard hates **[MASK]** bad

- *p(is)* = ?
- *p(are)* = ?

Assessing BERT's Syntactic Abilities (Goldberg. 2019)

	BERT Base	BERT Large	LSTM (M&L)	Humans (M&L)	# Pairs (# M&L Pairs)
SUBJECT-VERB AGREEMENT:					
Simple	1.00	1.00	0.94	0.96	120 (140)
In a sentential complement	0.83	0.86	0.99	0.93	1440 (1680)
Short VP coordination	0.89	0.86	0.90	0.82	720 (840)
Long VP coordination	0.98	0.97	0.61	0.82	400 (400)
Across a prepositional phrase	0.85	0.85	0.57	0.85	19440 (22400)
Across a subject relative clause	0.84	0.85	0.56	0.88	9600 (11200)
Across an object relative clause	0.89	0.85	0.50	0.85	19680 (22400)
Across an object relative (no *that*)	0.86	0.81	0.52	0.82	19680 (22400)
In an object relative clause	0.95	0.99	0.84	0.78	15960 (22400)
In an object relative (no *that*)	0.79	0.82	0.71	0.79	15960 (22400)
REFLEXIVE ANAPHORA:					
Simple	0.94	0.92	0.83	0.96	280 (280)
In a sentential complement	0.89	0.86	0.86	0.91	3360 (3360)
Across a relative clause	0.80	0.76	0.55	0.87	22400 (22400)

Table 3: Results on the Marvin and Linzen (2018) stimuli. M&L results numbers are taken from Marvin and Linzen (2018). The BERT and M&L numbers are *not* directly comparable, as the experimental setup differs in many ways.

Assessing BERT's Syntactic Abilities (Goldberg. 2019)

Colab Notebook:
https://colab.research.google.com/github/gsarti/lcl23-xnlm-lab/blob/main/notebooks/1.1_Transformer_Syntactic_Abilities.ipynb

Behavioural Interpretability

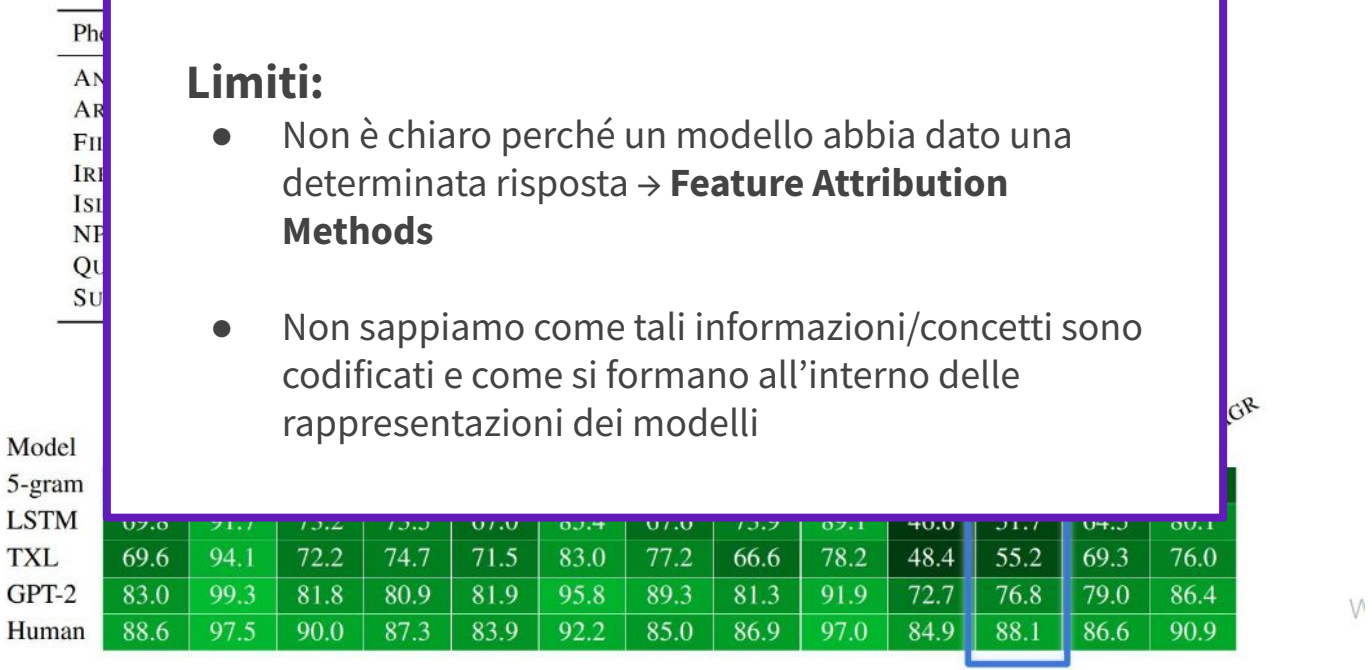
- **BLiMP** & **SyntaxGym**: Benchmark **suites** of different linguistic phenomena:

Phenomenon	N	Acceptable Example	Unacceptable Example
ANAPHOR AGR.	2	*Many girls insulted themselves.*	*Many girls insulted herself.*
ARG. STRUCTURE	9	*Rose wasn't disturbing Mark.*	*Rose wasn't boasting Mark.*
FILLER-GAP	7	*Brett knew what many waiters find.*	*Brett knew that many waiters find.*
IRREGULAR FORMS	2	*Aaron broke the unicycle.*	*Aaron broken the unicycle.*
ISLAND EFFECTS	8	*Which bikes is John fixing?*	*Which is John fixing bikes?*
NPI LICENSING	7	*The truck has clearly tipped over.*	*The truck has ever tipped over.*
QUANTIFIERS	4	*No boy knew fewer than six guys.*	*No boy knew at most six guys.*
SUBJECT-VERB AGR.	6	*These casseroles disgust Kayla.*	*These casseroles disgusts Kayla.*

Model	Overall	ANA. AGR	ARG. STR	BINDING	CTRL. RAIS.	D-N AGR	ELLIPSIS	FILLER. GAP	IRREGULAR	ISLAND	NPI	QUANTIFIERS	S-V AGR
5-gram	61.2	47.9	71.9	64.4	68.5	70.0	36.9	60.2	79.5	57.2	45.5	53.5	60.3
LSTM	69.8	91.7	73.2	73.5	67.0	85.4	67.6	73.9	89.1	46.6	51.7	64.5	80.1
TXL	69.6	94.1	72.2	74.7	71.5	83.0	77.2	66.6	78.2	48.4	55.2	69.3	76.0
GPT-2	83.0	99.3	81.8	80.9	81.9	95.8	89.3	81.3	91.9	72.7	76.8	79.0	86.4
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Behavioural Interpretability

- **BLiMP** & **SyntaxGym**: Benchmark **suites** of different linguistic phenomena:



Model												
5-gram												
LSTM												
TXL	69.6	94.1	72.2	74.7	71.5	83.0	77.2	66.6	78.2	48.4	55.2	69.3	76.0
GPT-2	83.0	99.3	81.8	80.9	81.9	95.8	89.3	81.3	91.9	72.7	76.8	79.0	86.4
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

Limiti:

- Non è chiaro perché un modello abbia dato una determinata risposta → **Feature Attribution Methods**
- Non sappiamo come tali informazioni/concetti sono codificati e come si formano all'interno delle rappresentazioni dei modelli

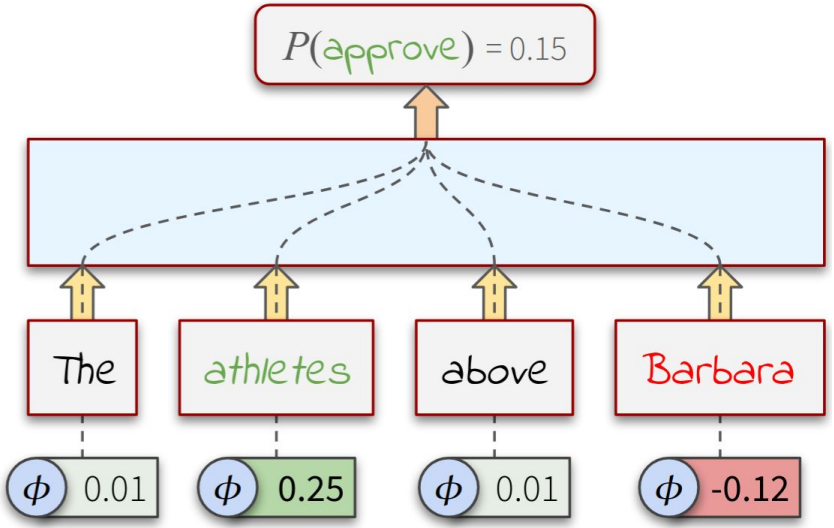
Da Tutorial EACL 2024 Transformer-specific Interpretability: https://github.com/interpretingdl/eacl2024_transformer_interpretability_tutorial/tree/main

Feature Attribution Methods

- I **feature attribution methods** cercano di spiegare le predizioni del modello in termini delle features (e.g. parole, token) che contribuiscono di più nella fase di inferenza
- Approccio che mostra la “logica” di un modello dietro una determinata previsione
- L’approccio più comune è perturbare parti dell’input e misurare il cambiamento nell'output del modello

Feature Attribution Methods

- I **feature attribution methods** cercano di spiegare le predizioni del modello in termini delle features (e.g. parole, token) che contribuiscono di più nella fase di inferenza
- Approccio che mostra la “logica” di un modello dietro una determinata previsione
- L’approccio più comune è perturbare parti dell’input e misurare il cambiamento nell'output del modello



Feature Attribution Methods

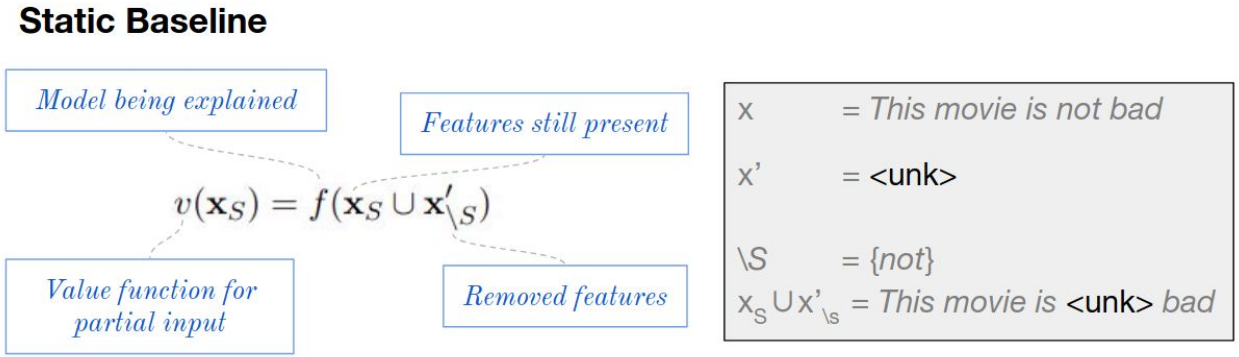
- Solitamente si confronta la sequenza originale rispetto ad una baseline neutrale

- La baseline è il valore che utilizziamo per sostituire una caratteristica (o un insieme di caratteristiche) nella sequenza originale

Feature Attribution Methods

- Solitamente si confronta la sequenza originale rispetto ad una baseline neutrale

- La baseline è il valore che utilizziamo per sostituire una caratteristica (o un insieme di caratteristiche) nella sequenza originale



Feature Attribution Methods

- Solitamente si confronta la sequenza originale rispetto ad una baseline neutrale

- La baseline è il valore che utilizziamo per sostituire una caratteristica (o un insieme di caratteristiche) nella sequenza originale

Interventional Baseline

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S})\right]$$

Expectation over removed features

x = “This movie is not bad”

\S = {*not*}

$x_S \cup x'_{\setminus s}$ = “This movie is *the* *is* *walk* ... bad”

Feature Attribution Methods

- Solitamente si confronta la sequenza originale rispetto ad una baseline neutrale

- La baseline è il valore che utilizziamo per sostituire una caratteristica (o un insieme di caratteristiche) nella sequenza originale

Observational Baseline

Conditioned on present features

$$v(\mathbf{x}_S) = \mathbb{E}_{\mathbf{x}'_{\setminus S}}\left[f(\mathbf{x}_S \cup \mathbf{x}'_{\setminus S}) \mid \mathbf{x}_S\right]$$

Expectation over removed features

x = “This movie is not bad”

\S = {*not*}

$x_S \cup x'_{\setminus s}$ = “This movie is *very* *that* *quite* ... bad”

Feature Attribution Methods

Input: *Can you stop the dog from*
Output: barking

1. Why did the model predict "barking"?
Can you stop the dog from

2. Why did the model predict "barking" *instead of* "crying"?
Can you stop the dog from

3. Why did the model predict "barking" *instead of* "walking"?
Can you stop the dog from

Importance of feature: difference of output when removed

$$S_E(x_i) = q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})$$

$$S^*_E(x_i) = (q(y_t|\boldsymbol{x}) - q(y_t|\boldsymbol{x}_{\neg i})) - (q(y_f|\boldsymbol{x}) - q(y_f|\boldsymbol{x}_{\neg i}))$$

*Explanation with respect to **foil***

Integrated Gradients

- L’**Integrated Gradients** è uno degli approcci più popolari di feature attribution (Sundararajan et al., 2017)

- Approccio:
  - Stimare quanto ogni input contribuisce all'output confrontandolo con un input “neutro” (baseline)
  - Utilizzare il gradiente del modello rispetto all'input, integrato lungo un percorso tra input e baseline

Integrated Gradients

- L'**Integrated Gradients** è uno degli approcci più popolari di feature attribution (Sundararajan et al., 2017)

- Approccio:
  - Stimare quanto ogni input contribuisce all'output confrontandolo con un input "neutro" (baseline)
  - Utilizzare il gradiente del modello rispetto all'input, integrato lungo un percorso tra input e baseline

$$\mathrm{IG}_i(x) = (x_i - x_i') \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

Integrated Gradients

- L'**Integrated Gradients** è uno degli approcci più popolari di feature attribution (Sundararajan et al., 2017)

- Approccio:
  - Stimare quanto ogni input contribuisce all'output confrontandolo con un input "neutro" (baseline)
  - Utilizzare il gradiente del modello rispetto all'input, integrato lungo un percorso tra input e baseline

$$\mathrm{IG}_i(x) = (x_i - x'_i) \cdot \int_{\alpha=0}^{1} \frac{\partial F(x' + \alpha(x - x'))}{\partial x_i} d\alpha$$

Dove:

- *x*: input del modello
- *x'*: baseline (es. token [PAD] o tutti zero)
- *F*: funzione del modello (es. logit della classe)
- *α*: scala tra baseline e input
  i: dimensione/feature

Integrated Gradients

- Definizione della baseline x' (es. input vuoto o neutro)
- Calcolo dei passaggi intermedi tra baseline e input: $x'+\alpha(x-x')x'$ per diversi $\alpha \in [0,1]$
- Per ogni punto, calcolo del gradiente dell'output rispetto all'input
- Calcolo dell’integrale dei gradienti
- Moltiplicazione per $(x-x')$ → valore di attribuzione per ogni input token

Integrated Gradients

- Definizione della baseline x' (es. input vuoto o neutro)
- Calcolo dei passaggi intermedi tra baseline e input: $x'+\alpha(x-x')x'$ per diversi $\alpha \in [0,1]$
- Per ogni punto, calcolo del gradiente dell'output rispetto all'input
- Calcolo dell'i

- Moltiplicazione per $(x-x')$ → valore di attribuzione per ogni input token

Integrated Gradients

Legend: ■ Negative □ Neutral ■ Positive

True Label	Predicted Label	Attribution Label	Attribution Score	Word Importance
pos	pos (0.96)	pos	1.29	it was a fantastic performance ! #pad
pos	pos (0.87)	pos	1.56	best film ever #pad #pad #pad #pad
pos	pos (0.92)	pos	1.14	such a great show ! #pad #pad
neg	neg (0.29)	pos	-1.11	it was a horrible movie #pad #pad
neg	neg (0.22)	pos	-1.03	i 've never watched something as bad
neg	neg (0.07)	pos	-0.84	that is a terrible movie . #pad

Captum Library: https://captum.ai/tutorials/IMDB_TorchText_Interpret

Probing Tasks

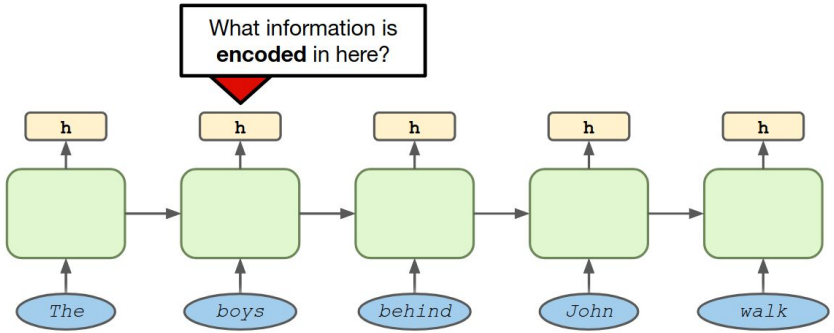
- Opening the black box: focus sulle rappresentazioni interne dei modelli

- **Approccio**: addestrare un classificatore/regressore usando le rappresentazioni interne del modello (i.e. embeddings) come input features

Probing Tasks

- Opening the black box: focus sulle rappresentazioni interne dei modelli

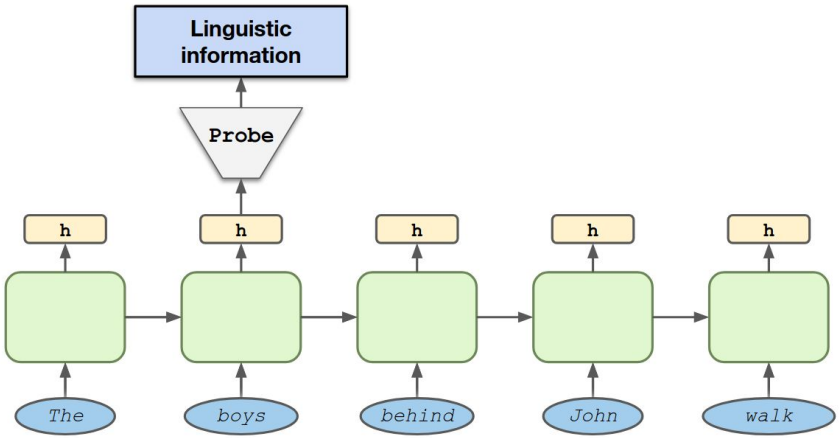
- **Approccio**: addestrare un classificatore/regressore usando le rappresentazioni interne del modello (i.e. embeddings) come input features



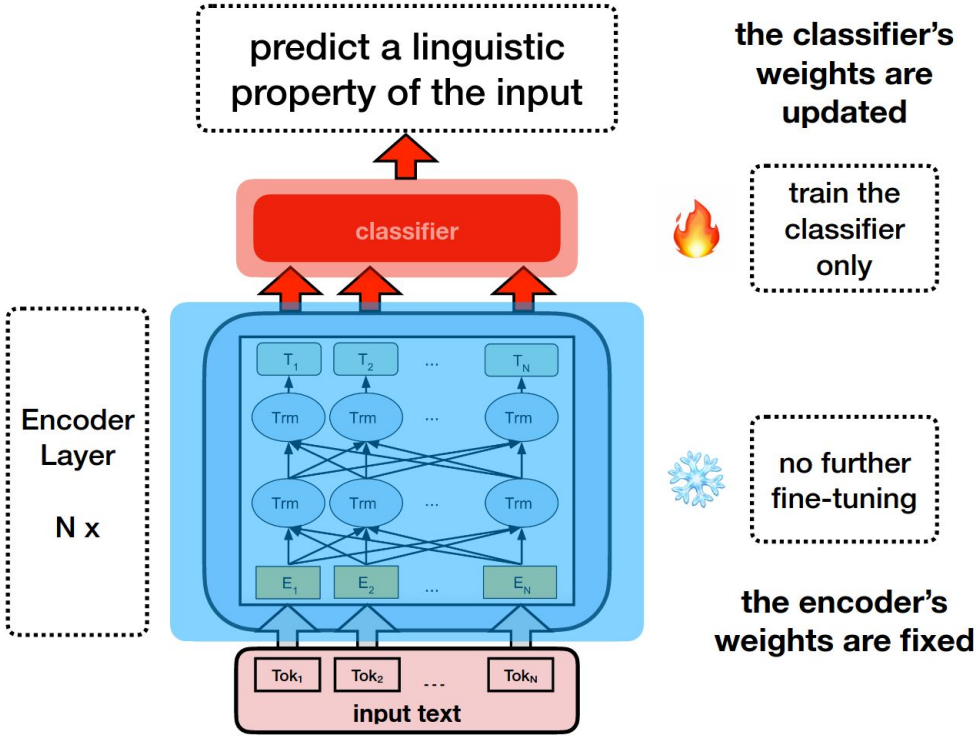
Probing Tasks

- Opening the black box: focus sulle rappresentazioni interne dei modelli

- **Approccio**: addestrare un classificatore/regressore usando le rappresentazioni interne del modello (i.e. embeddings) come input features

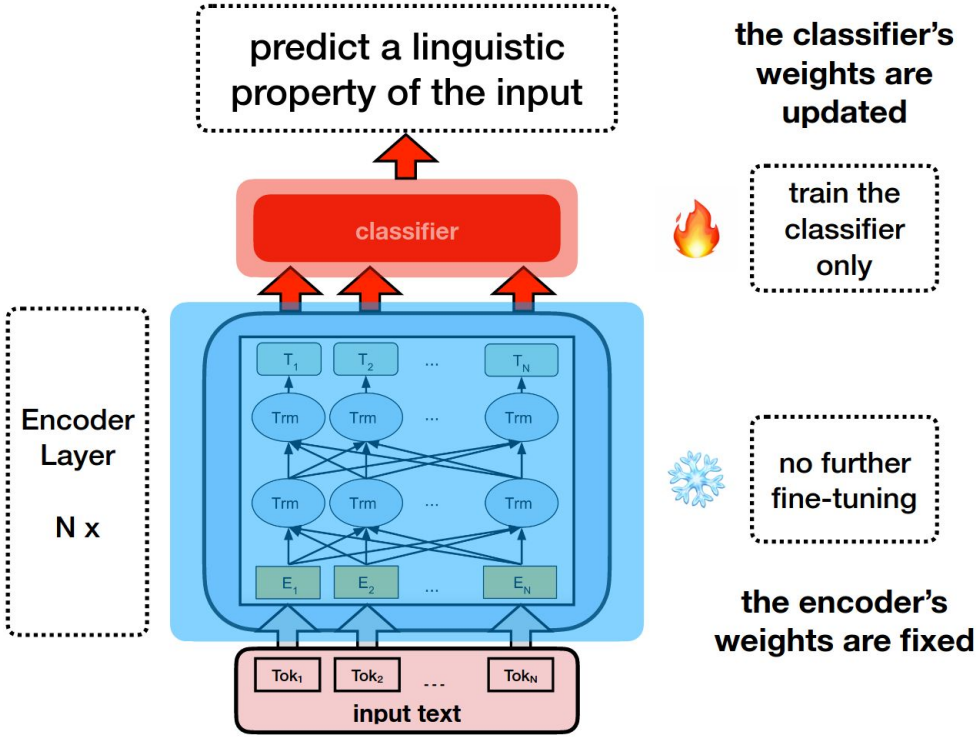


Probing Tasks



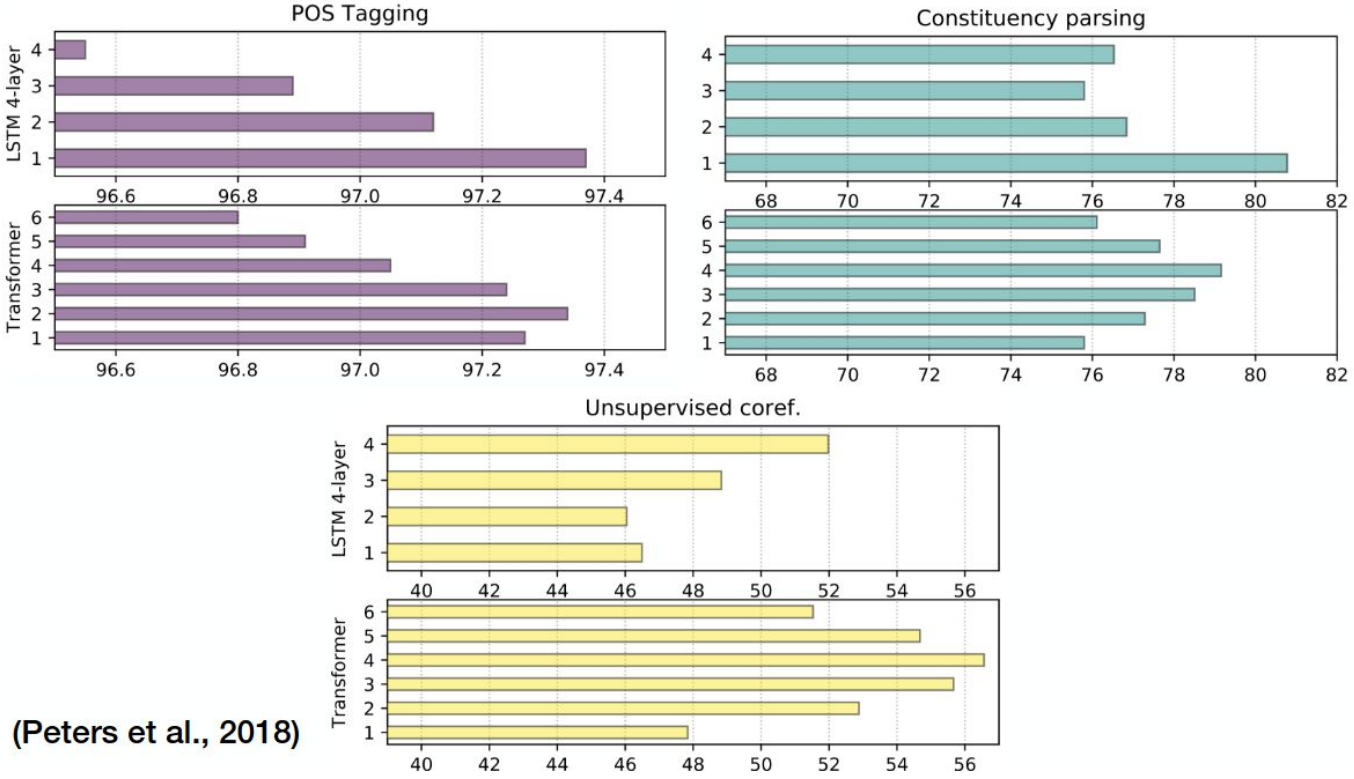
Fonte: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/19-probes.pdf

Probing Tasks



Fonte: https://people.cs.umass.edu/~miyyer/cs685_f20/slides/19-probes.pdf

Probing Tasks



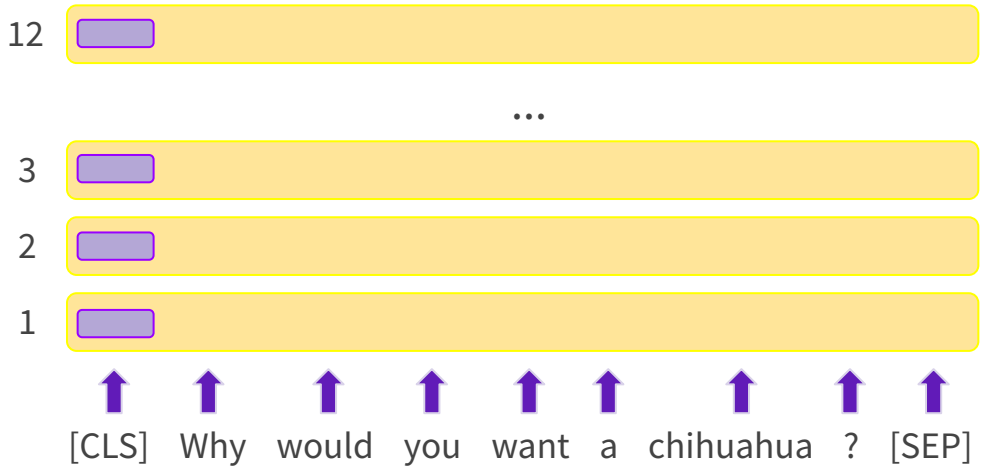
(Peters et al., 2018)

Profiling di un NLM

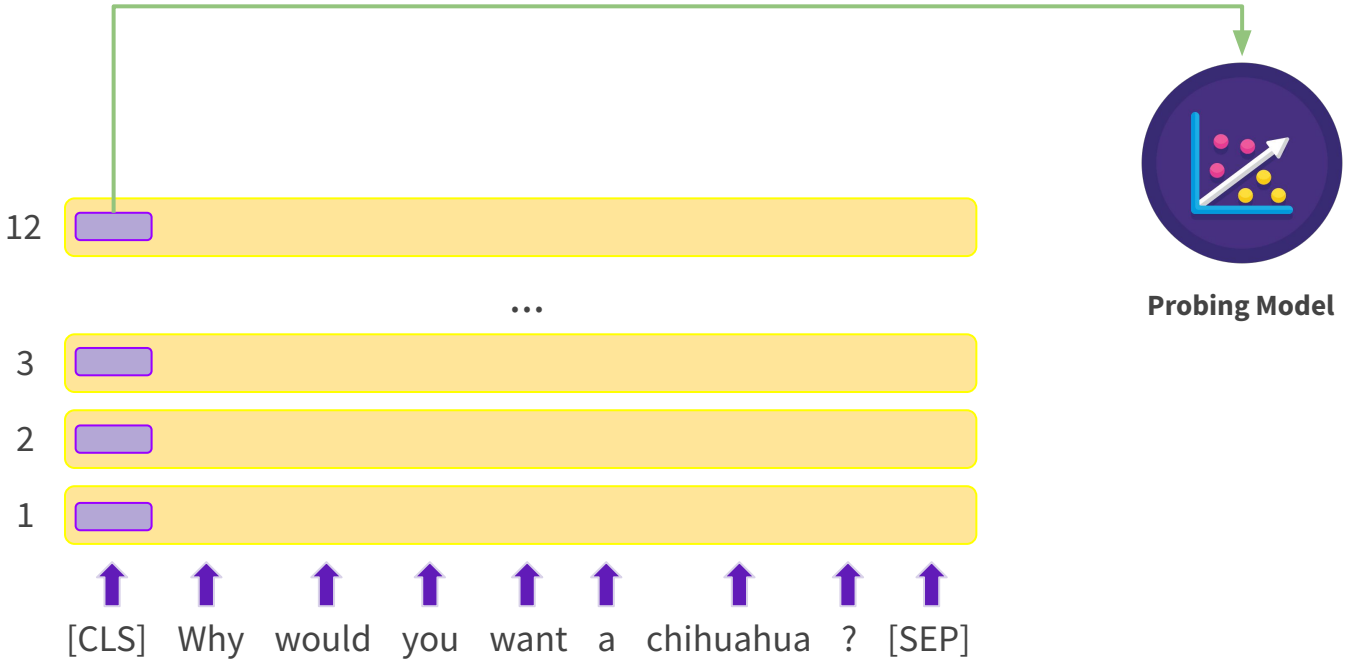
- La metodologia del "*linguistic profiling*" (van Halteren, 2004) parte dal presupposto che un ampio numero di caratteristiche linguistiche sia particolarmente utile per la risoluzione di diversi compiti di NLP, ad esempio:
  - Text Profiling (e.g. text readability, textual genres)
  - Author Profiling (e.g. author's age and native language)

Research Question:
Il potere predittivo di queste caratteristiche potrebbe essere utile anche per comprendere il comportamento di un NLM?

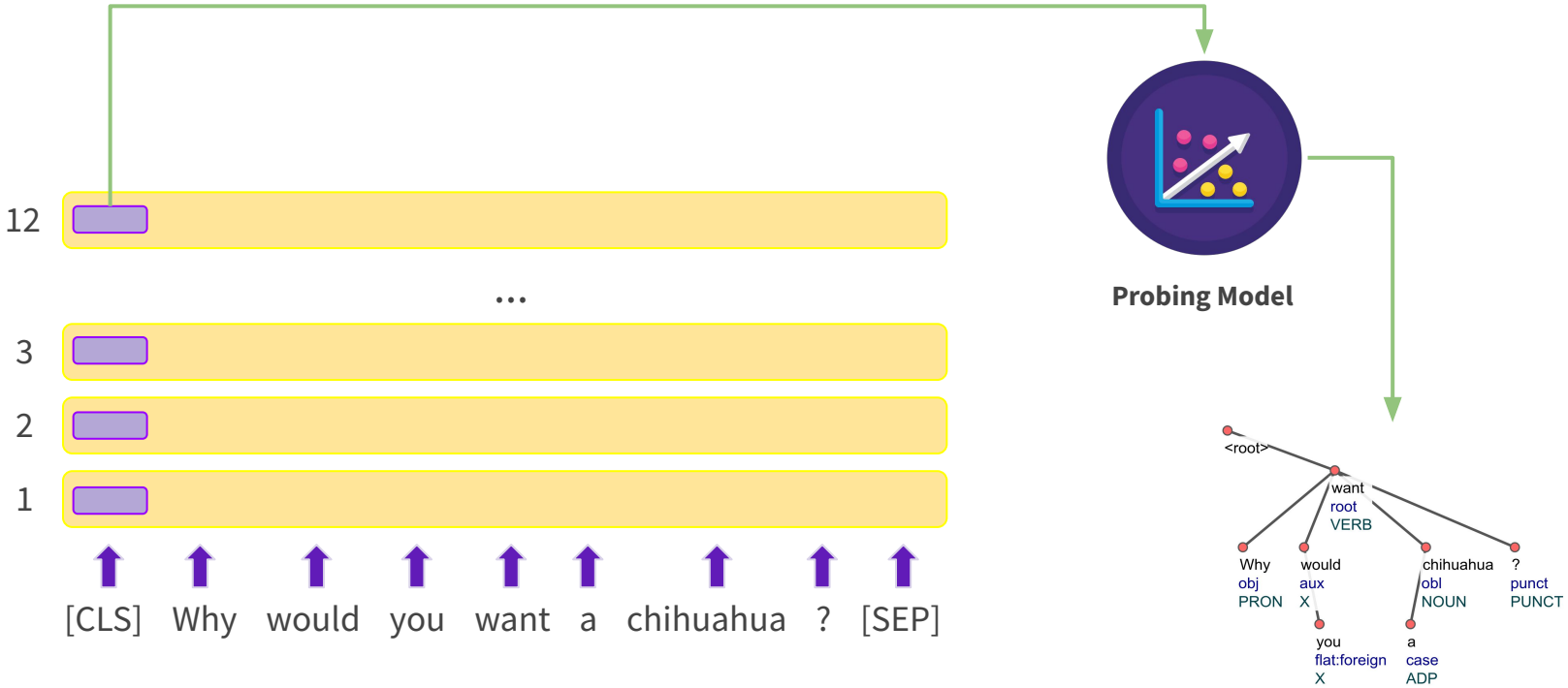
Profiling di un NLM



Profiling di un NLM



Profiling di un NLM



Profiling-UD: un tool per il Linguistic Profiling

- ProfilingUD (Brunato et al., 2020) è un'applicazione web che permette di eseguire il profiling linguistico di un testo, o di una vasta collezione di testi, per più lingue

- Consente l'estrazione di oltre 130 caratteristiche, che coprono diversi livelli di descrizione linguistica

- Link: http://linguistic-profiling.italianlp.it/

Linguistic Feature
Raw Text Properties
Sentence Length
Word Length
Vocabulary Richness
Type/Token Ratio for words and lemmas
Morphosyntactic information
Distibution of UD and language–specific POS
Lexical density
Inflectional morphology
Inflectional morphology of lexical verbs and auxiliaries
Verbal Predicate Structure
Distribution of verbal heads and verbal roots
Verb arity and distribution of verbs by arity
Global and Local Parsed Tree Structures
Depth of the whole syntactic tree
Average length of dependency links and of the longest link
Average length of prepositional chains and distribution by depth
Clause length
Relative order of elements
Order of subject and object
Syntactic Relations
Distribution of dependency relations
Use of Subordination
Distribution of subordinate and principal clauses
Average length of subordination chains and distribution by depth
Relative order of subordinate clauses

Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

- Studio delle competenze linguistiche codificate implicitamente nelle rappresentazioni di BERT

Domande di Ricerca:

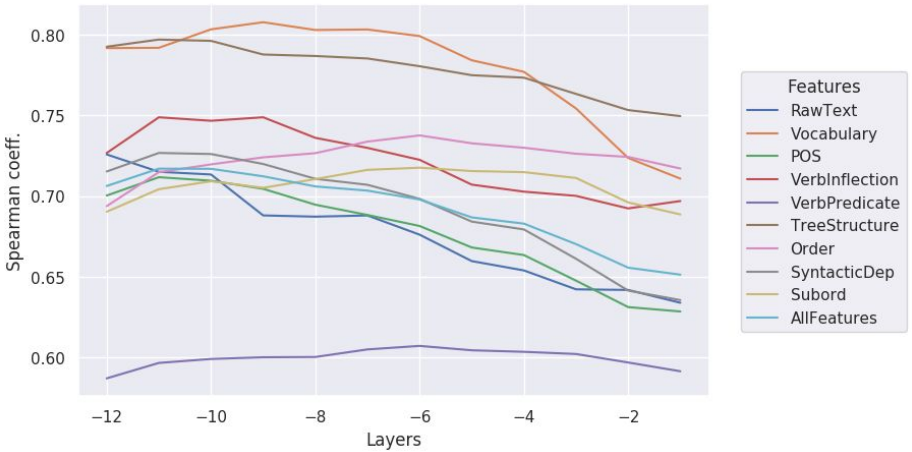
1. Quali proprietà linguistiche sono codificate in una versione pre-addestrata di BERT?

2. Come questa conoscenza si modifica al seguito di una fase di fine-tuning?

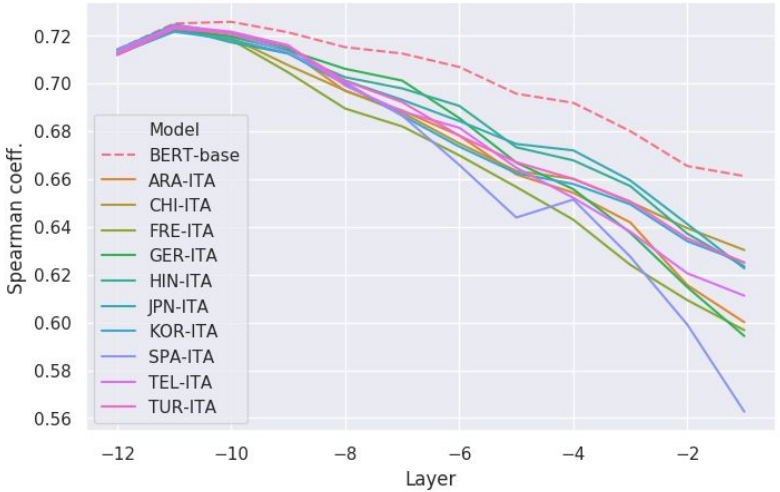
3. Esiste una relazione fra questa competenza e l'abilità del modello nel risolvere un downstream task?

Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

Pre-trained Model:

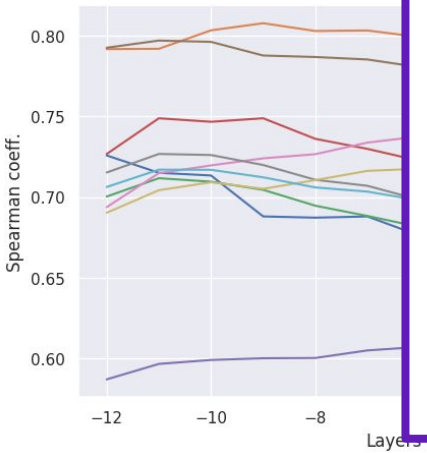


Fine-tuned Model:

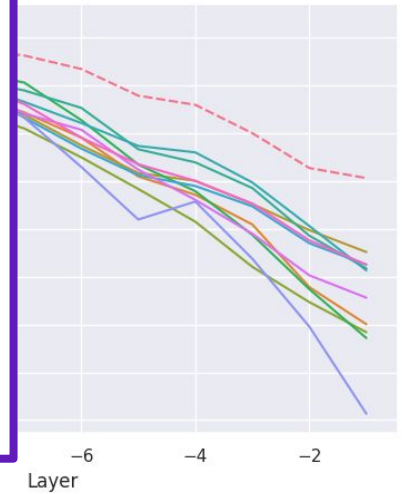


Linguistic Profiling of a Neural Language Model (Miaschi et al., 2020)

Pre-trained Model:



Fine-tuned Model:



Limitazioni:

- Quante delle informazioni apprese dal “probing model” sono effettivamente codificate nelle rappresentazioni del modello e non un effetto del learning del modello stesso?
- I probing tasks mostrano se determinati concetti /proprietà sono codificate nelle rappresentazioni di modelli → Non ci mostra se e come il modello utilizza effettivamente queste informazioni per le sue predizioni

Mechanistic Interpretability

Da Tutorial EACL 2024 Transformer-specific Interpretability: https://github.com/interpretingdl/eacl2024_transformer_interpretability_tutorial/tree/main

Mechanistic Interpretability

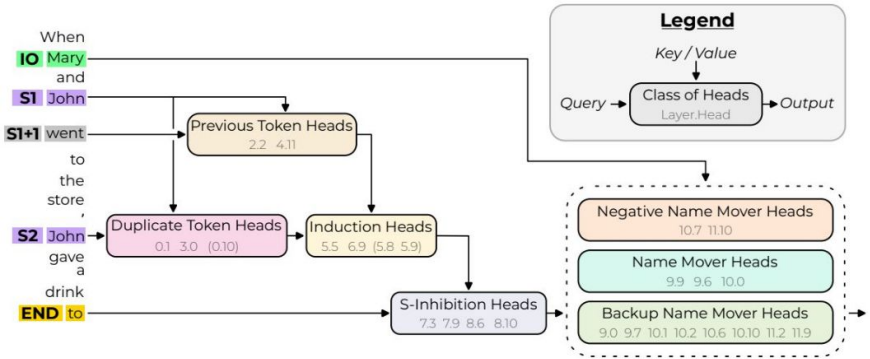
- La Mechanistic Interpretability è un sotto-dominio degli studi sull'interpretabilità dei NLMs con lo scopo di fare reverse-engineering della struttura della rete

- Si basa sullo studio di tecniche di interpretabilità causale e delle sotto parti della rete (e.g. attention heads, MLPs, neuroni)

Mechanistic Interpretability

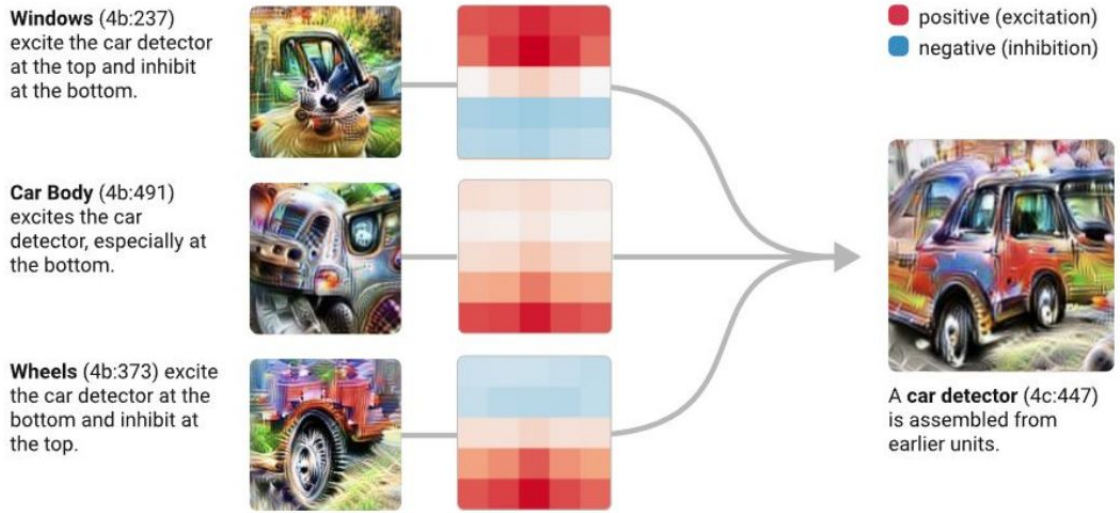
- La Mechanistic Interpretability è un sotto-dominio degli studi sull'interpretabilità dei NLMs con lo scopo di fare reverse-engineering della struttura della rete

- Si basa sullo studio di tecniche di interpretabilità causale e delle sotto parti della rete (e.g. attention heads, MLPs, neuroni)



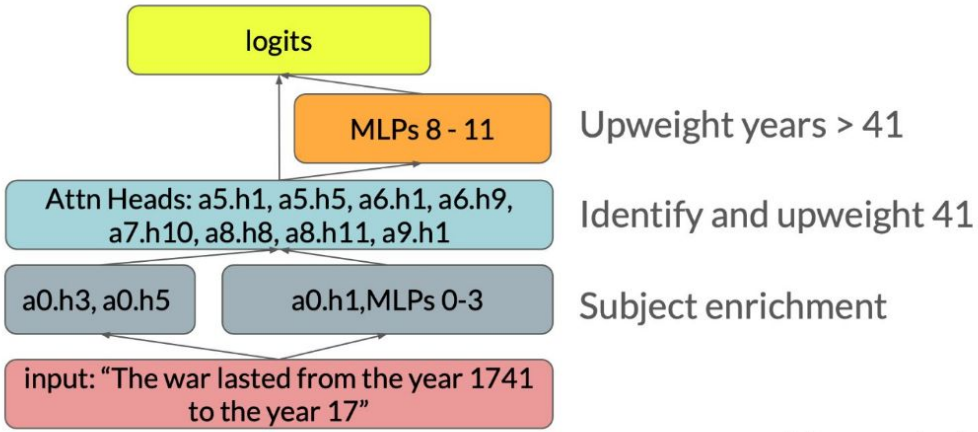
Transformer Circuits

- Cosa sono i circuiti?
- Olah et al. (2020) hanno definito i circuiti come “sub-graphs of the network, consisting a set of tightly linked features and the weights between them”



Transformer Circuits

- Cosa sono i circuiti?
- Olah et al. (2020) hanno definito i circuiti come “sub-graphs of the network, consisting a set of tightly linked features and the weights between them”



Hanna et al., 2023

Transformer Circuits

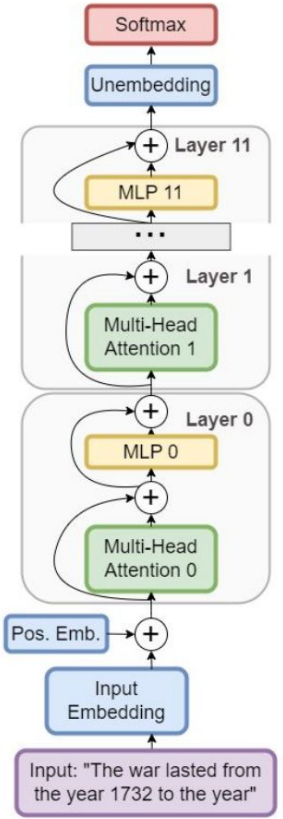
- **Circuit**: computational subgrapgh minimo di un dato modello responsabile (faithful) della risoluzione di un determinato compito

- **Minimal Computational Subgraph**: set minimo di “nodes” and “edges” di un modello
- **Task**: set di input e outputs, misurabile tramite una loss function
- **Faithful**: la loss del modello rimane invariata nel momento in cui tutti gli edges esterni al circuito vengono rimossi

Transformer Circuits

- **Circuit**: computational subgrapgh minimo di un dato modello responsabile (faithful) della risoluzione di un determinato compito

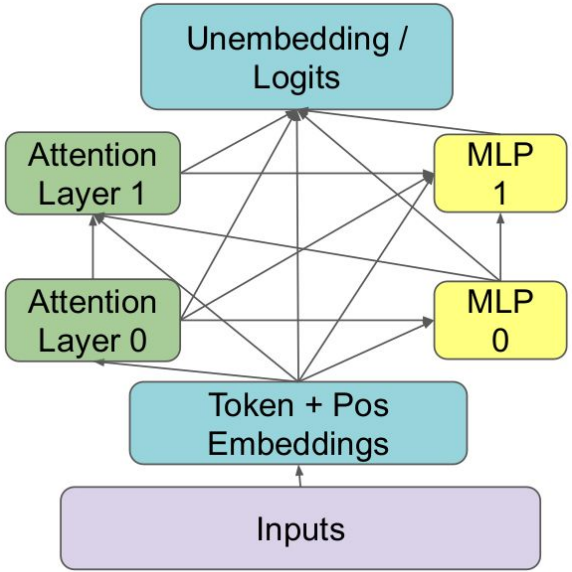
- **Minimal Computational Subgraph**: set minimo di “nodes” and “edges” di un modello
- **Task**: set di input e outputs, misurabile tramite una loss function
- **Faithful**: la loss del modello rimane invariata nel momento in cui tutti gli edges esterni al circuito vengono rimossi



Transformer Circuits

- **Circuit**: computational subgrapgh minimo di un dato modello responsabile (faithful) della risoluzione di un determinato compito

- **Minimal Computational Subgraph**: set minimo di “nodes” and “edges” di un modello
- **Task**: set di input e outputs, misurabile tramite una loss function
- **Faithful**: la loss del modello rimane invariata nel momento in cui tutti gli edges esterni al circuito vengono rimossi



Transformer Circuits

Task: Greater-Than

Inputs: “The war lasted from 1741 to 17”
Expected outputs: a 2-digit number greater than 41
Metric: $\sum_{y>41} p(y) - \sum_{y<=41} p(y)$

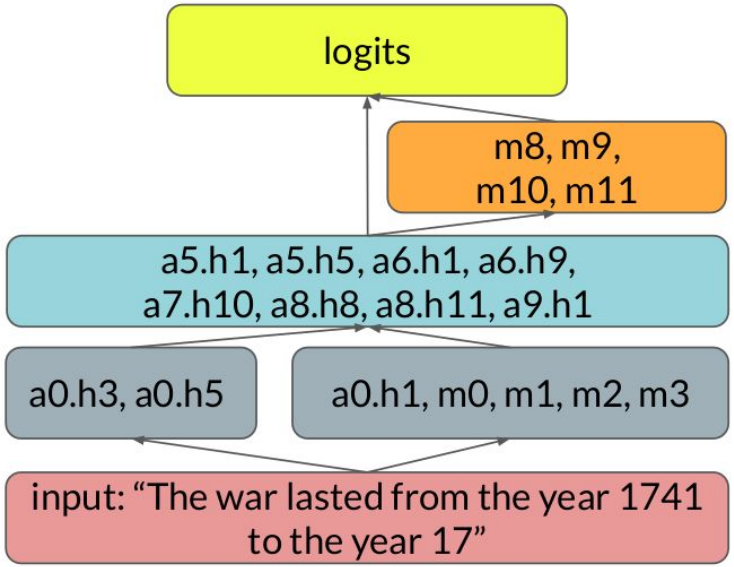
Tasks should be solvable by your model, and evaluable in one forward pass.

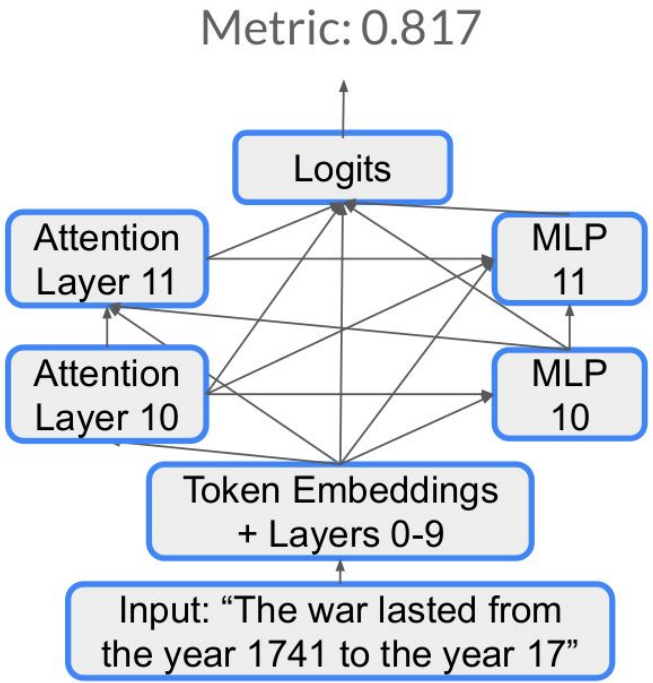
Average Metric Value: 0.817

For circuit-finding, we also need corrupted inputs.

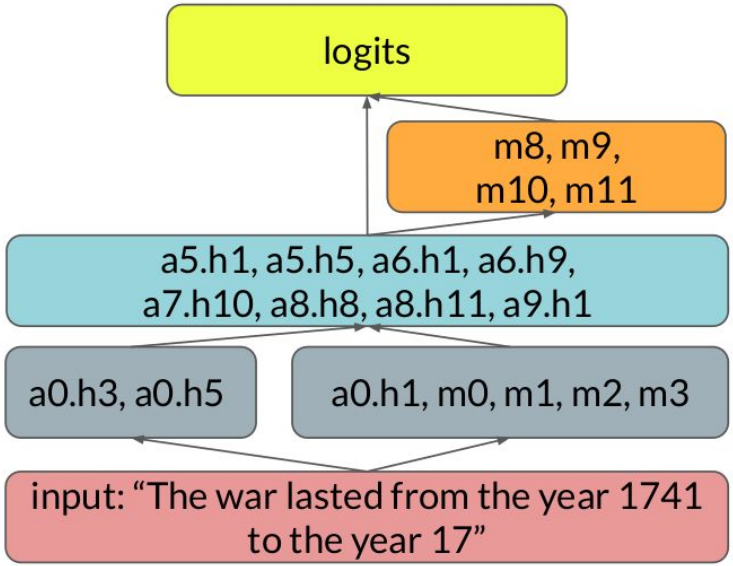
Corrupted inputs: “The war lasted from 1701 to 17”

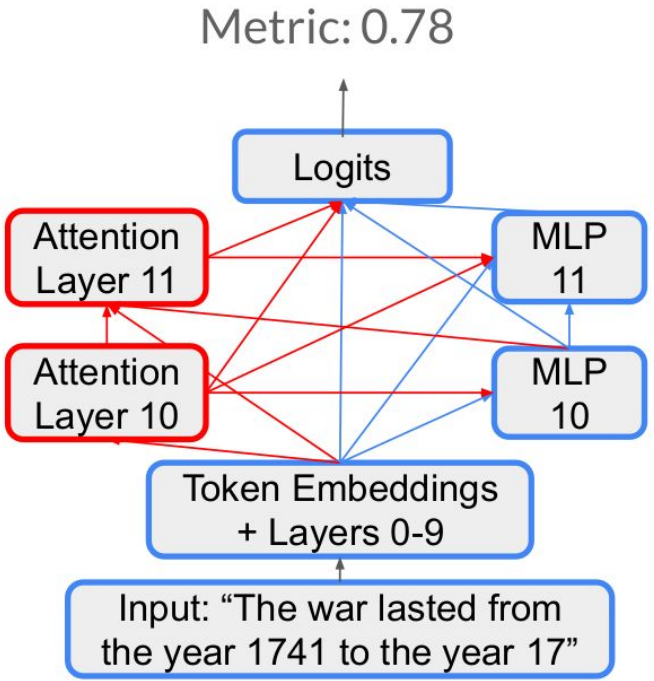
Transformer Circuits





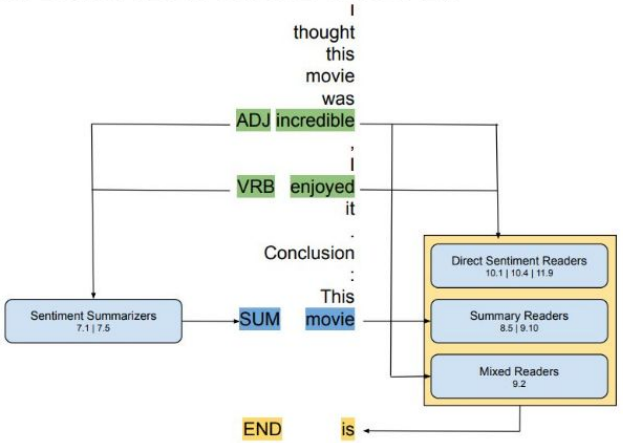
Transformer Circuits





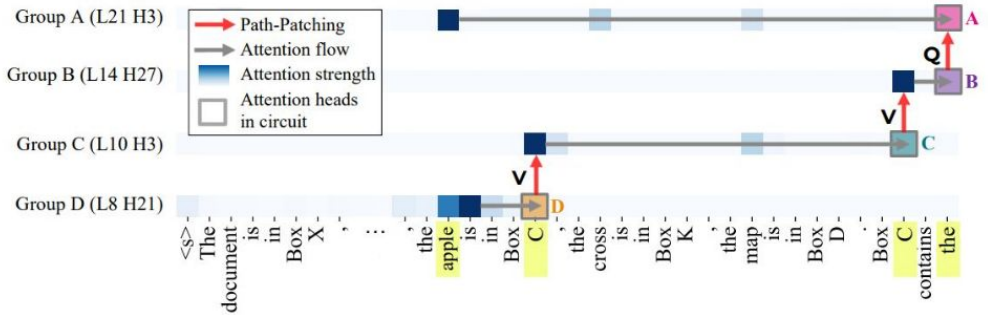
Transformer Circuits

LINEAR REPRESENTATIONS OF SENTIMENT IN LARGE LANGUAGE MODELS



Tigges et al., 2023

FINE-TUNING ENHANCES EXISTING MECHANISMS: A CASE STUDY ON ENTITY TRACKING



Prakash et al., 2024

Lezione Pratica