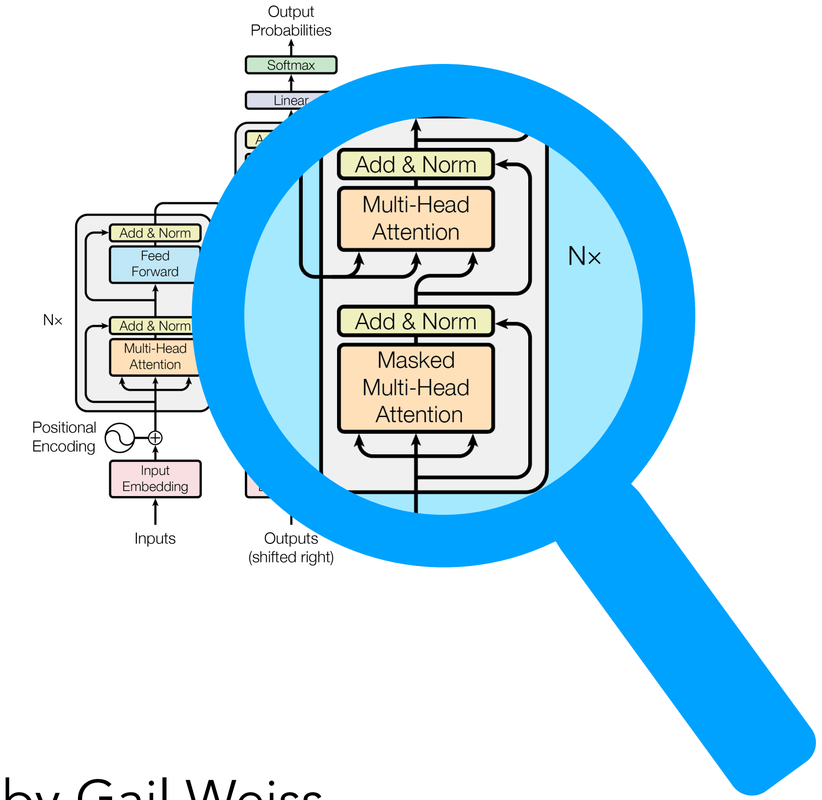
Interpretability



Given by Gail Weiss

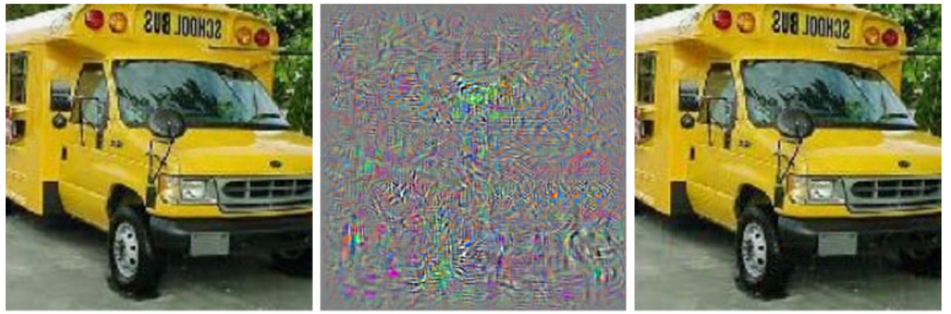
Outline

- ✦ Introduction
- ✦ Methods and Concepts
  - Several black and white box methods, intuition vs hypotheses, observation vs intervention
  - Background interrupt: Classical NLP tasks
- ✦ Friends
  - Formal Analysis; Extraction
- ✦ Conclusion

Can we trust neural networks?

Introduction

Why did my model do that?



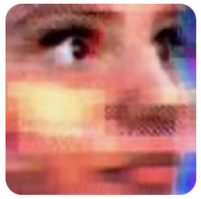
bus ostrich

CBS News

Microsoft shuts down AI chatbot, Tay, after it turned into a Nazi

Microsoft got a swift lesson this week on the dark side of social media. Yesterdav the company launched "Tay," an artifi

25 Mar 2016



Introduction

"Interpretability is the degree to which a human can understand the cause of a decision"

Tim Miller, 2017

"Interpretability is the degree to which a human can consistently predict the model's result"

Been Kim et al, 2016

Introduction

"Interpretability is the degree to which a human can understand the cause of a decision"

Tim Miller, 2017

"Interpretability is the degree to which a human can consistently predict the model's result"

Been Kim et al, 2016

A: Intrinsic Interpretability:

Rule Based Models

Decision Trees; Linear Models; Finite State Machines…

Introduction

"Interpretability is the degree to which a human can understand the cause of a decision"

Tim Miller, 2017

"Interpretability is the degree to which a human can consistently predict the model's result"

Been Kim et al, 2016

A: Intrinsic Interpretability:

Rule Based Models

Decision Trees; Linear Models; Finite State Machines…

Introduction

"Interpretability is the degree to which a human can understand the cause of a decision"

Tim Miller, 2017

"Interpretability is the degree to which a human can consistently predict the model's result"

Been Kim et al, 2016

A: Intrinsic Interpretability:

Rule Based Models

Decision Trees; Linear Models; Finite State Machines…

B: Post Hoc Interpretability

Extraction

Converting to decision trees; FSMs; other rules…

Introduction

"Interpretability is the degree to which a human can understand the cause of a decision"

Tim Miller, 2017

"Interpretability is the degree to which a human can consistently predict the model's result"

Been Kim et al, 2016

A: Intrinsic Interpretability:

Rule Based Models

Decision Trees; Linear Models; Finite State Machines…

B: Post Hoc Interpretability

Investigating trained models

Feature Visualisation; LIME; SHAP; Leave-one-out; Saliency maps; Probing; Causal Tracing; De-embedding space; Key-Value Pairs; …

Extraction

Converting to decision trees; FSMs; other rules…

Introduction

"Interpretability is the degree to which a human can understand the cause of a decision"

Tim Miller, 2017

"Interpretability is the degree to which a human can consistently predict the model's result"

Been Kim et al, 2016

A: Intrinsic Interpretability:

Rule Based Models

Decision Trees; Linear Models; Finite State Machines…

B: Post Hoc Interpretability

Investigating trained models

Feature Visualisation; LIME; SHAP; Leave-one-out; Saliency maps; Probing; Causal Tracing; De-embedding space; Key-Value Pairs; …

Extraction

Converting to decision trees; FSMs; other rules…

Intuitions and Visualisations

Cell sensitive to position in line:

```
The sole importance of the crossing of the Berezina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutuzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energy was directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
made for crossing as by what took place at the bridges. When the bridges
broke down, unarmed soldiers, people from Moscow and women with children
who were with the French transport, all--carried on by vis inertiae--
pressed forward into boats and into the ice-covered water and did not,
surrender.
```

Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."
```

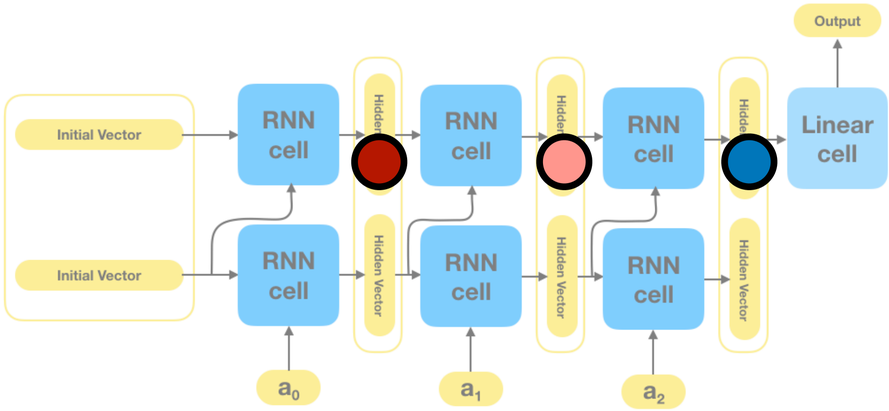
Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
        siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```

A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
```

Individual cells in an RNN hidden state



The Unreasonable Effectiveness of Recurrent Neural Networks, Karpathy 2015

Architecture	White box embedder
Required data	None
Explains	Neuron, Sample
Focus	Neuron
Use for...	Intuition

Intuitions and Visualisations

Cell sensitive to position in line:

```
The sole importance of the crossing of the Berezina lies in the fact
that it plainly and indubitably proved the fallacy of all the plans for
cutting off the enemy's retreat and the soundness of the only possible
line of action--the one Kutuzov and the general mass of the army
demanded--namely, simply to follow the enemy up. The French crowd fled
at a continually increasing speed and all its energy was directed to
reaching its goal. It fled like a wounded animal and it was impossible
to block its path. This was shown not so much by the arrangements it
made for crossing as by what took place at the bridges. When the bridges
broke down, unarmed soldiers, people from Moscow and women with children
who were with the French transport, all--carried on by vis inertiae--
pressed forward into boats and into the ice-covered water and did not,
surrender.
```

Cell that turns on inside quotes:

```
"You mean to imply that I have nothing to eat out of.... On the
contrary, I can supply you with everything even if you want to give
dinner parties," warmly replied Chichagov, who tried by every word he
spoke to prove his own rectitude and therefore imagined Kutuzov to be
animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating
smile: "I meant merely to say what I said."
```

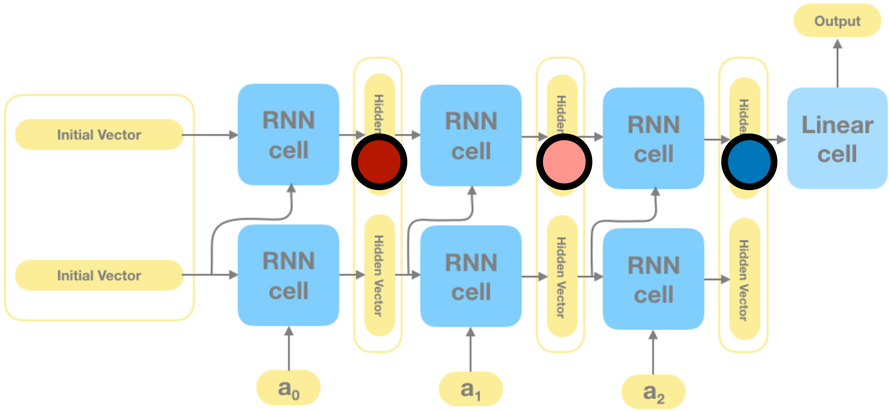
Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
        siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```

A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
```

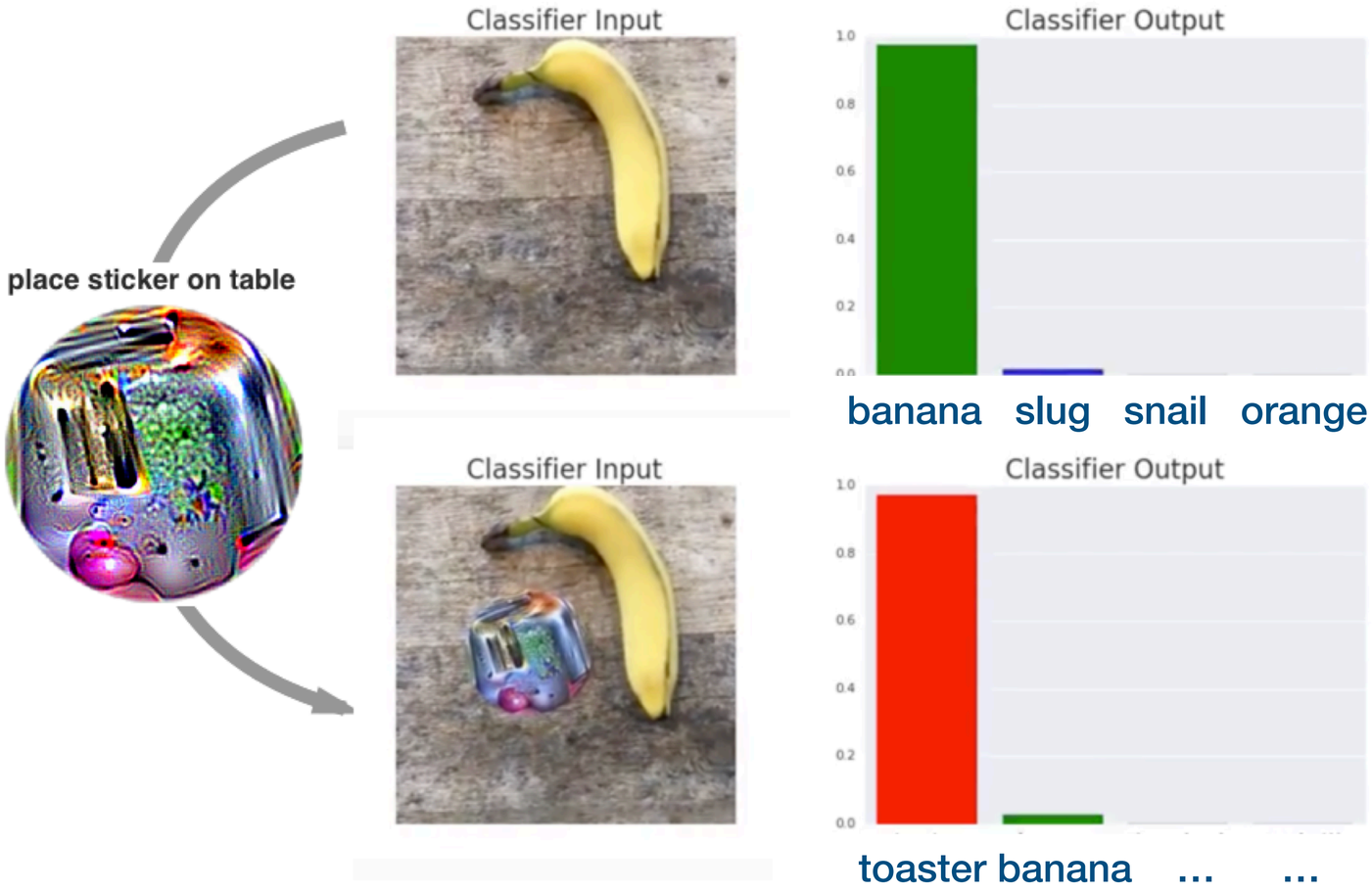
Individual cells in an RNN hidden state



The Unreasonable Effectiveness of Recurrent Neural Networks, Karpathy 2015

Architecture	White box classifier
Required data	None
Explains	Model
Focus	Class
Use for…	Modifying output across samples!

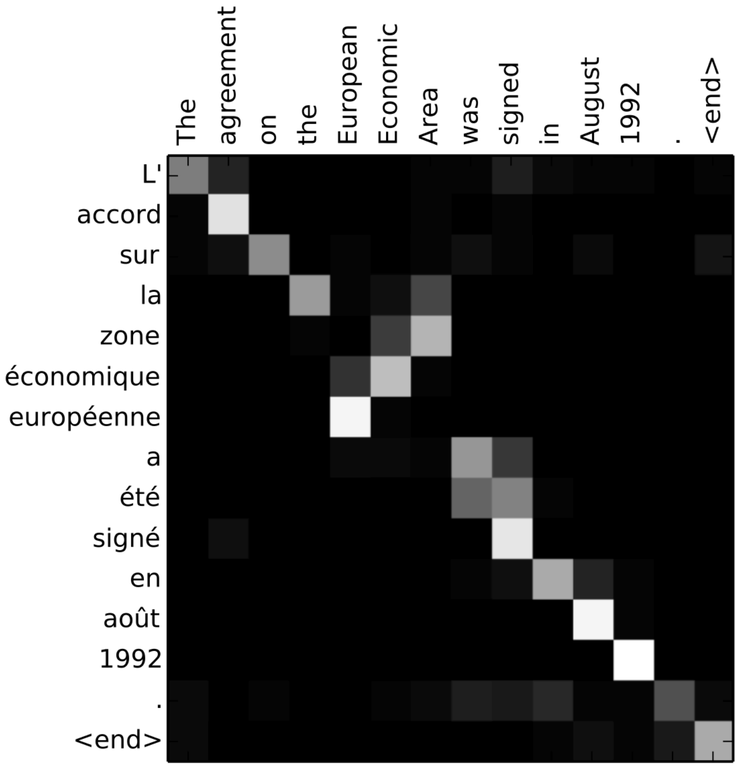
Intuitions and Visualisations



Adversarial patch, Brown et al, 2018

Architecture	White box with attention
Required data	None
Explains	Model
Focus	Partial Computation
Use for...	Intuition

Intuitions and Visualisations

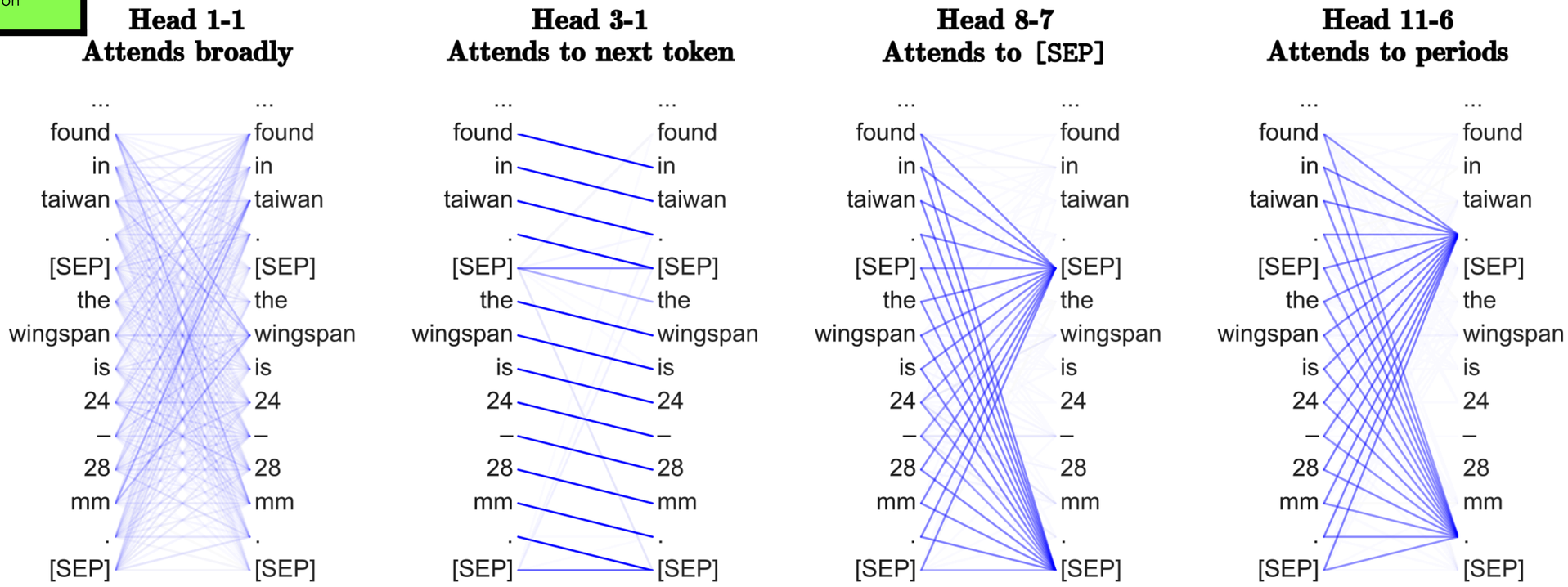


Attention!

Neural Machine Translation by Jointly Learning to Align and Translate, Bahdanau et al, 2014

Architecture	White box with attention
Required data	None
Explains	Model
Focus	Partial Computation
Use for...	Intuition

Intuitions and Visualisations



What does BERT look at? An analysis of BERT's attention, Clark et al, 2019

Attention is not explanation, Jain and Wallace, 2019

Attention is not not explanation, Wiegreffe and Pinter, 2019

Outline

- ✦ Introduction
- ✦ Methods and Concepts
  - Several black and white box methods, intuition vs hypotheses, observation vs intervention
  - Background interrupt: Classical NLP tasks
- ✦ Friends
  - Formal Analysis; Extraction
- ✦ Conclusion

We have a model and a classification - what can we check?

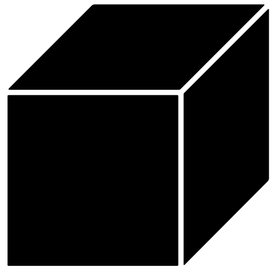
Basic approaches

complete black box

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

Leave One Out

I couldn’t like this more!



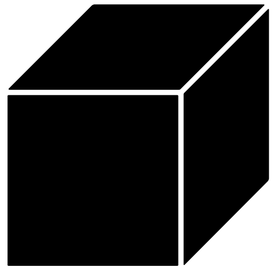
👍 **0.9**

👎 **0.1**

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

Leave One Out

I couldn’t like this more!



👍 **0.9**

👎 **0.1**

Manipulation		Result
I couldn’t like this more!	👍	**0.9**
I couldn’t like this more!	👍	**0.7**

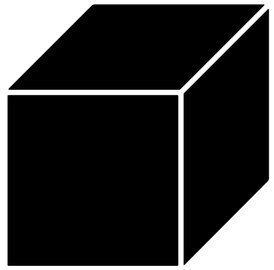
Conclusion?

“I” and “couldn’t” were not very important in this sample

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

Leave One Out

I couldn’t like this more!



👍 **0.9**

👎 **0.1**

Manipulation		Result
I couldn’t like this more!	👍	**0.9**
I couldn’t like this more!	👍	**0.7**
I couldn’t like this more!	👍	**0.5**

Conclusion?

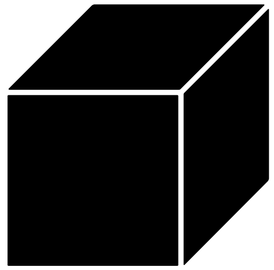
“I” and “couldn’t” were not very important in this sample

“like” quite important

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

Leave One Out

I couldn’t like this more!



👍 **0.9**

👎 **0.1**

Manipulation	Result
I couldn’t like this more!	👍 **0.9**
I couldn’t like this more!	👍 **0.7**
I couldn’t like this more!	👍 **0.5**
I couldn’t like this more!	👍 **0.3**
I couldn’t like this more!	👍 **0.1**

Conclusion?

“I” and “couldn’t” were not very important in this sample

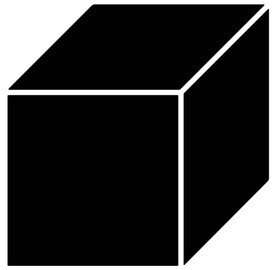
“like” quite important

“this”, “more” critical???

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

Leave One Out

I couldn’t like this more!



👍 **0.9**

👎 **0.1**

Manipulation	Result
I couldn’t like this more!	👍 **0.9**
I couldn’t like this more!	👍 **0.7**
I couldn’t like this more!	👍 **0.5**
I couldn’t like this more!	👍 **0.3**
I couldn’t like this more!	👍 **0.1**

Conclusion?

“I” and “couldn’t” were not very important in this sample

“like” quite important

“this”, “more” critical???

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

Weaknesses

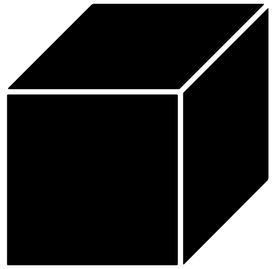
- Doesn’t catch interactions between features
- Explanation can be based on illegal (i.e. OOD) inputs

Can we consider multiple features?

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

Local Surrogates

I couldn’t like this more!



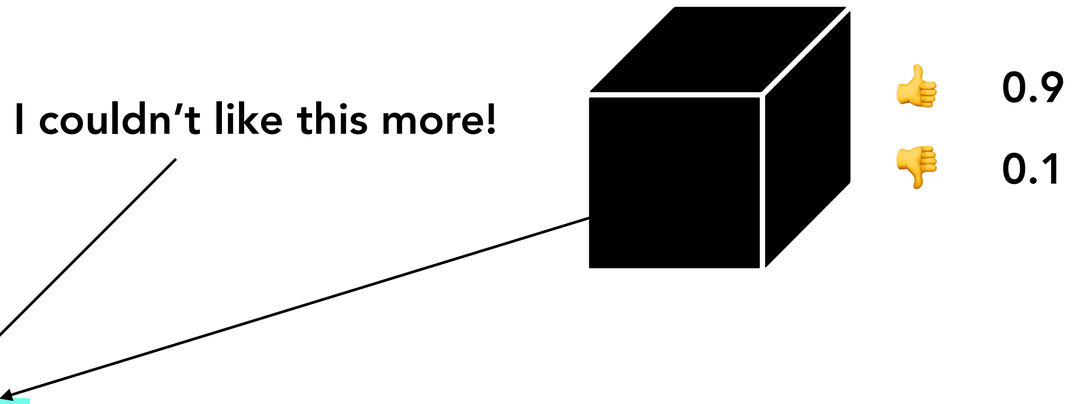
👍 **0.9**

👎 **0.1**

“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et al, 2016

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for...	Intuition

Local Surrogates



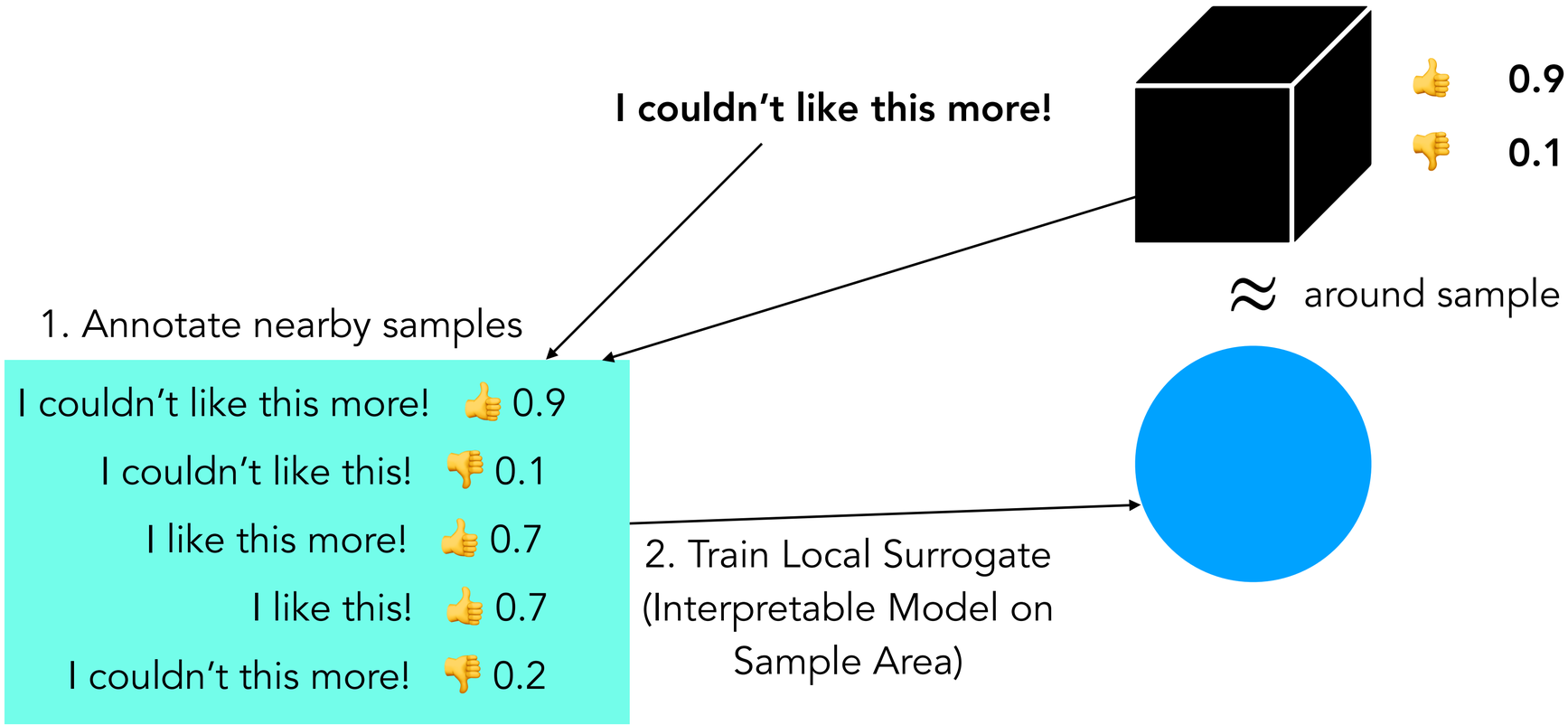
1. Annotate nearby samples

I couldn’t like this more!	👍 0.9
I couldn’t like this!	👎 0.1
I like this more!	👍 0.7
I like this!	👍 0.7
I couldn’t this more!	👎 0.2

“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et al, 2016

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for...	Intuition

Local Surrogates

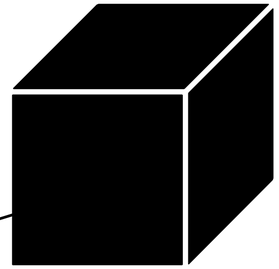


“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et al, 2016

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for...	Intuition

Local Surrogates

I couldn't like this more!



👍 **0.9**

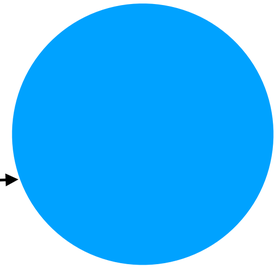
👎 **0.1**

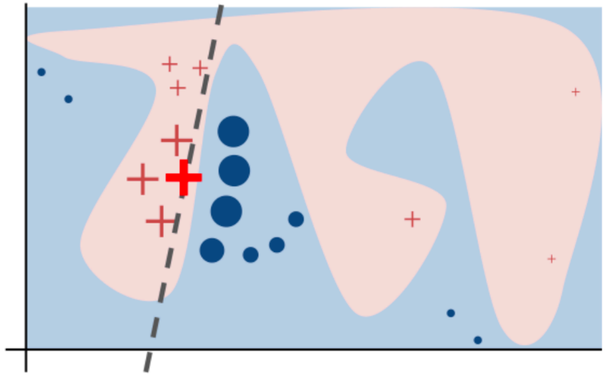
≈ around sample

1. Annotate nearby samples

I couldn't like this more!	👍 0.9
I couldn't like this!	👎 0.1
I like this more!	👍 0.7
I like this!	👍 0.7
I couldn't this more!	👎 0.2

2. Train Local Surrogate (Interpretable Model on Sample Area)





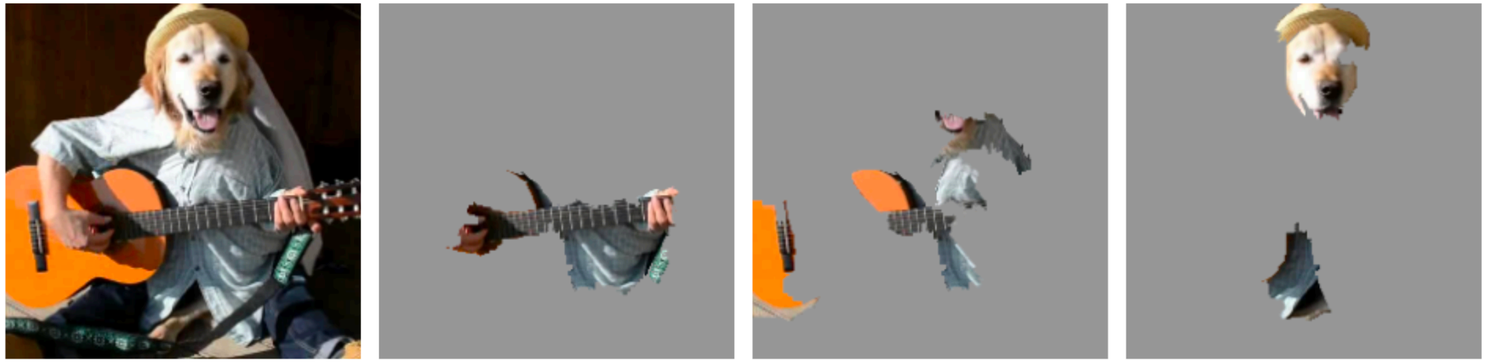
"Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et al, 2016

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for...	Intuition

Local Surrogates

LIME: Local Interpretable Model-agnostic Explanations

On images: use “superpixels”
Include and exclude similarly to tokens in sequences



Original | Electric Guitar | Acoustic Guitar | Labrador

“Why Should I Trust You?” Explaining the Predictions of Any Classifier, Ribeiro et al, 2016

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for...	Intuition

Local Surrogates

LIME: Local Interpretable Model-agnostic Explanations

Weaknesses

Unstable - kernel (sample "area") can change explanation!

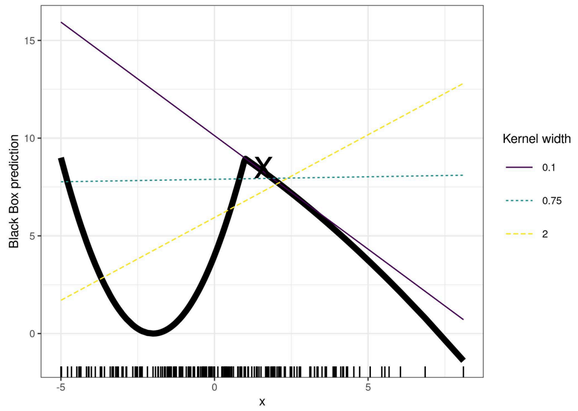


Image: Interpretable Machine Learning: A guide for making black boxes explainable, Molner, 2023

Explanation may be based on illogical (OOD) inputs

Image: Introduction to Model Interpretability, Jonathan Mak

"Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et al, 2016

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

Local Surrogates

LIME: Local Interpretable Model-agnostic Explanations

Weaknesses

Unstable - kernel (sample "area") can change explanation!

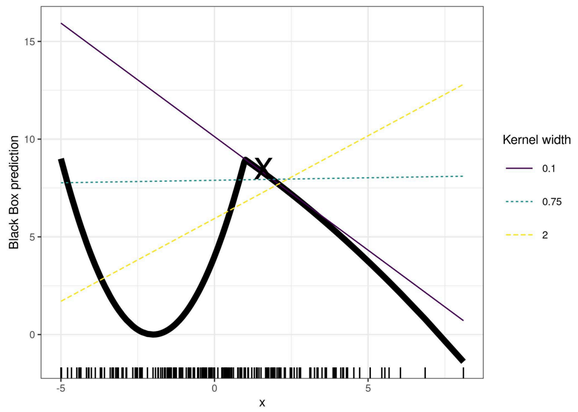


Image: Interpretable Machine Learning: A guide for making black boxes explainable, Molner, 2023

Explanation may be based on illogical (OOD) inputs

Image: Introduction to Model Interpretability, Jonathan Mak

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

"Why Should I Trust You?" Explaining the Predictions of Any Classifier, Ribeiro et al, 2016

SHAP

LIME

Architecture	Black box
Required data	None
Explains	Sample area
Focus	Features
Use for…	Intuition

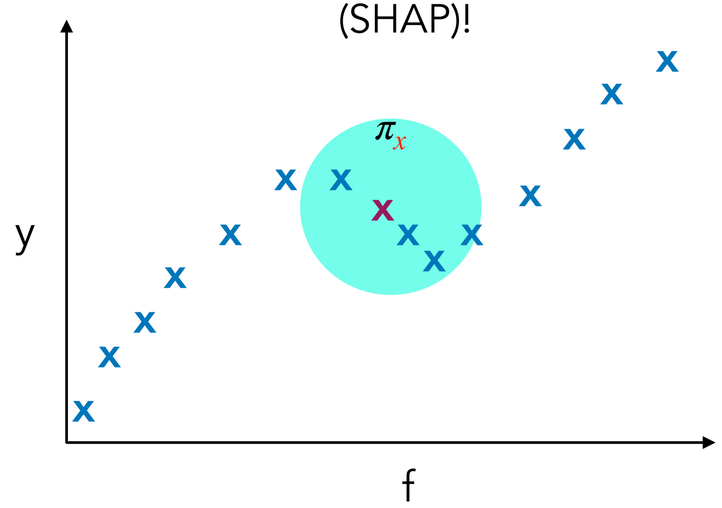
SHAP

Architecture	Black box
Required data	Train samples
Explains	Model
Focus	Features
Use for…	Intuition

A Unified Approach to Interpreting Model Predictions, Lundberg and Lee, 2017

Warning

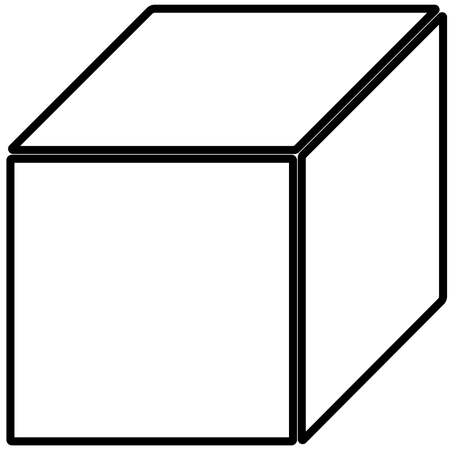
A feature can have negative correlation with output locally (LIME), and positive globally (SHAP)!



Interpreter Beware!

"An adversarial entity [can] craft an arbitrary desired explanation" -

Fooling LIME and SHAP: Adversarial Attacks on Post hoc Explanation Methods, Slack et al, 2020



Looking inside

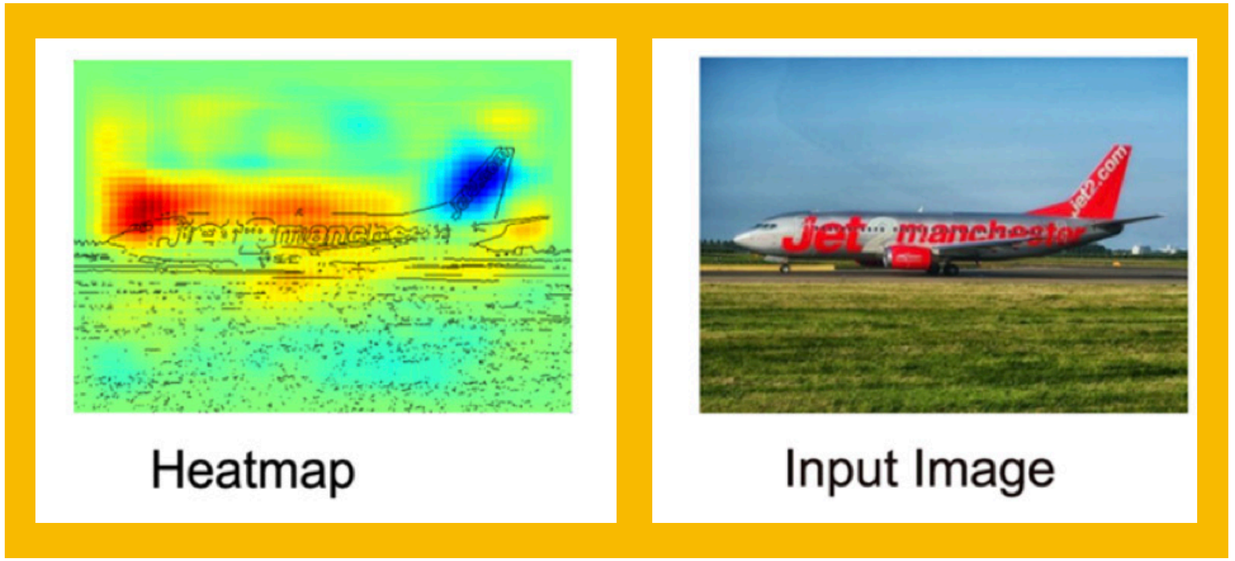
White Box

Architecture	White box with layers
Required data	None
Explains	Sample
Focus	Class in sample
Use for…	Intuition

Layer-Wise Relevance Propagation

Iteratively propagate output class score through layers of network, assigning relevance scores to the neurons at each layer

Adapted for deeper models, which have deep non linear “path” from input to output

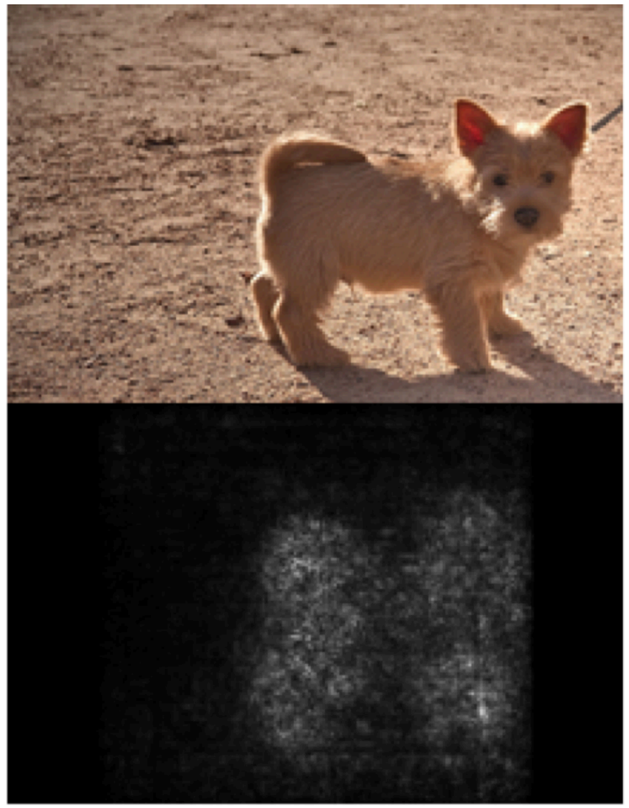


On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation, Bach et al, 2015

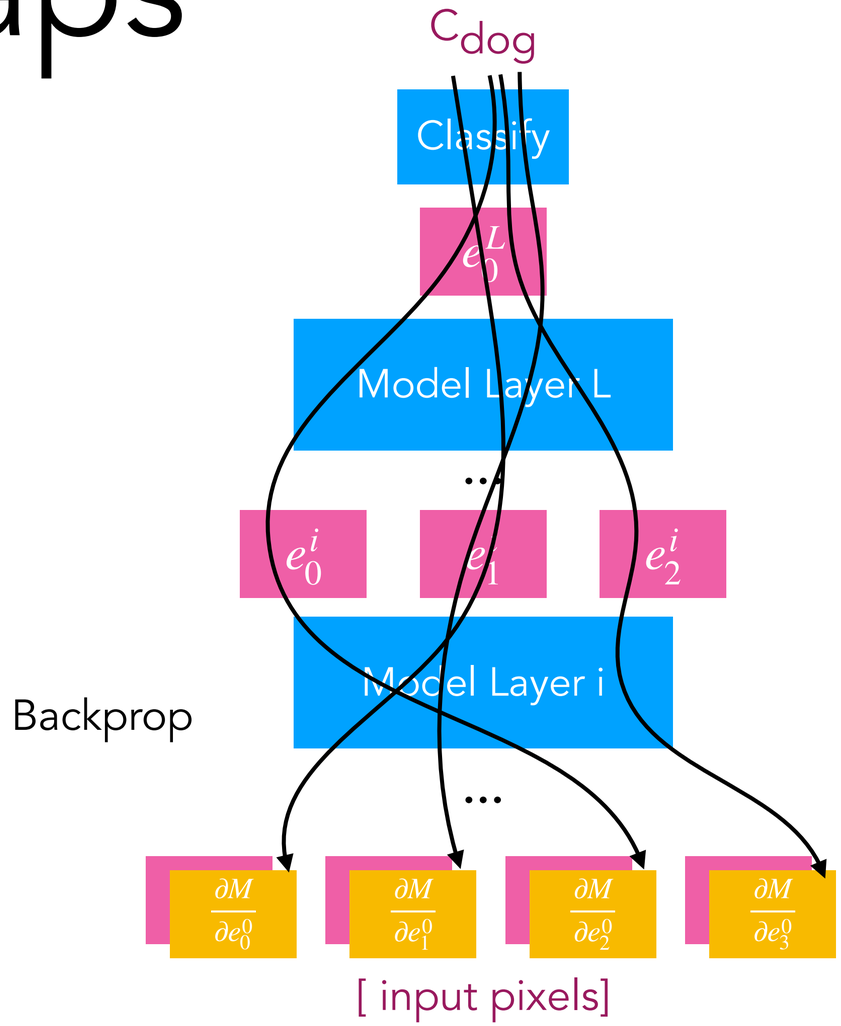
Architecture	Differentiable White box
Required data	None
Explains	Sample
Focus	Class in sample
Use for…	Intuition

Saliency Maps

AKA Pixel Attribution



$$\frac{\partial M}{\partial p}(C_{\text{dog}})$$



Backprop

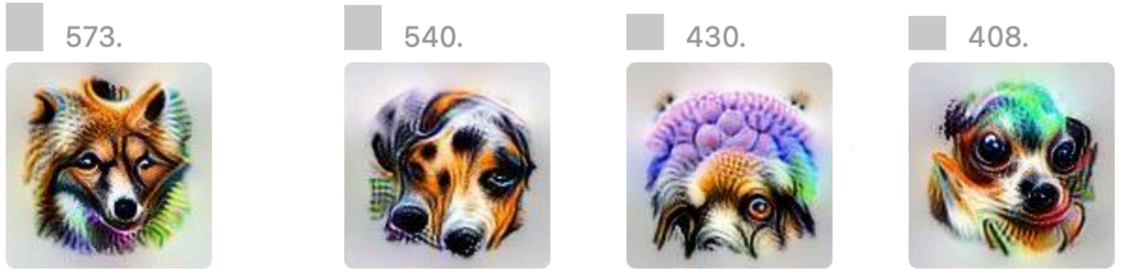
Deep inside convolutional networks: Visualising image classification models and saliency maps, Simonyan et al, 2013

Can we do better?

Gradients

Architecture	Differentiable white box embedder
Required data	None
Explains	Model
Focus	Neuron
Use for…	Intuition

Activation Maximisation

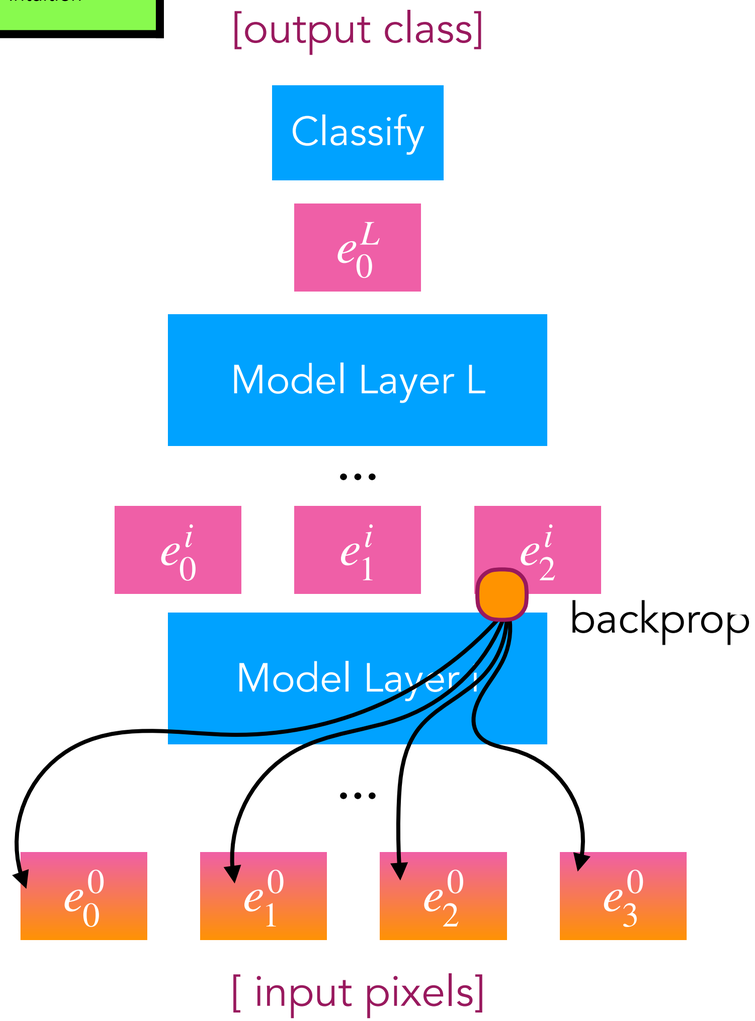


Visualizing Higher-Layer Features of a Deep Network, Erhan et al, 2009

The building blocks of interpretability, Olah et al, 2018

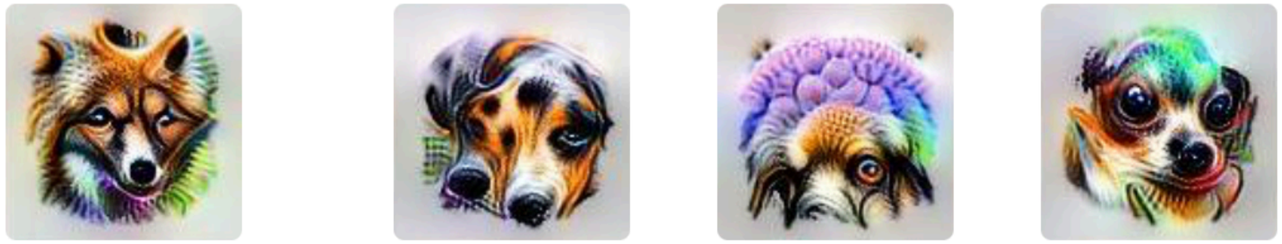
Architecture	Differentiable white box embedder
Required data	None
Explains	Model
Focus	Neuron
Use for…	Intuition

Activation Maximisation



Single Activation Maximisation

$[e_2^i]_{326}$ $[e_2^i]_{293}$ $[e_2^i]_{282}$ $[e_2^i]_{256}$

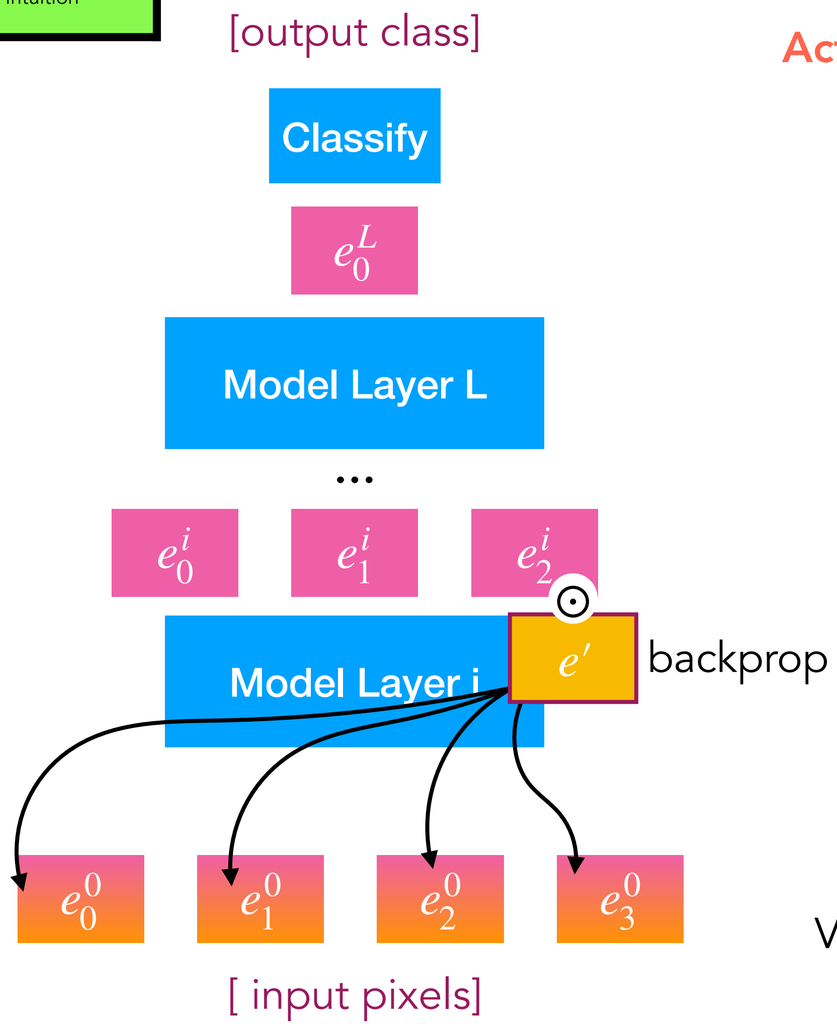


Visualizing Higher-Layer Features of a Deep Network, Erhan et al, 2009

The building blocks of interpretability, Olah et al, 2018

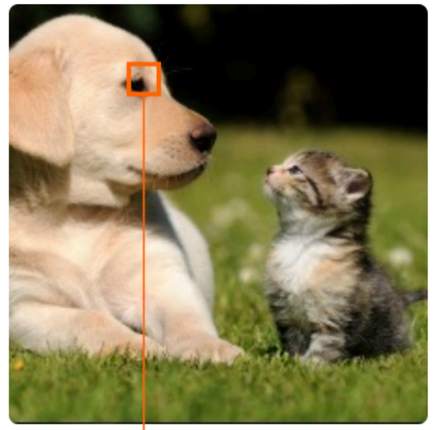
Architecture	Differentiable white box embedder
Required data	None
Explains	Sample
Focus	Layer
Use for…	Intuition

Activation Maximisation



Activation Vector Maximisation

image optimised for $[e_2^i]$
(input position "2", layer i)

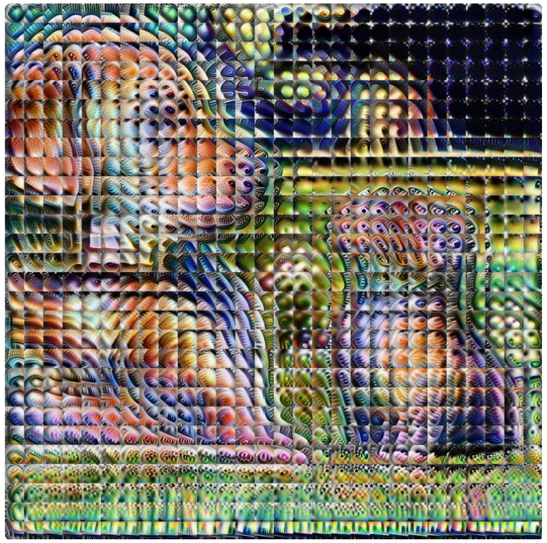
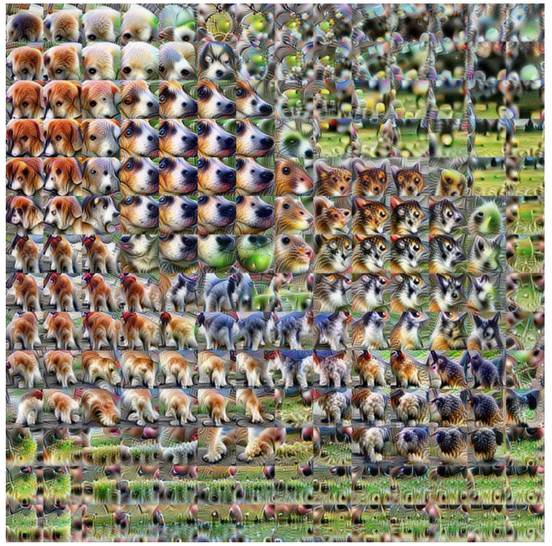


Visualizing Higher-Layer Features of a Deep Network, Erhan et al, 2009

The building blocks of interpretability, Olah et al, 2018

Architecture	Differentiable white box embedder
Required data	None
Explains	Sample
Focus	Layer
Use for…	Intuition

Activation Maximisation

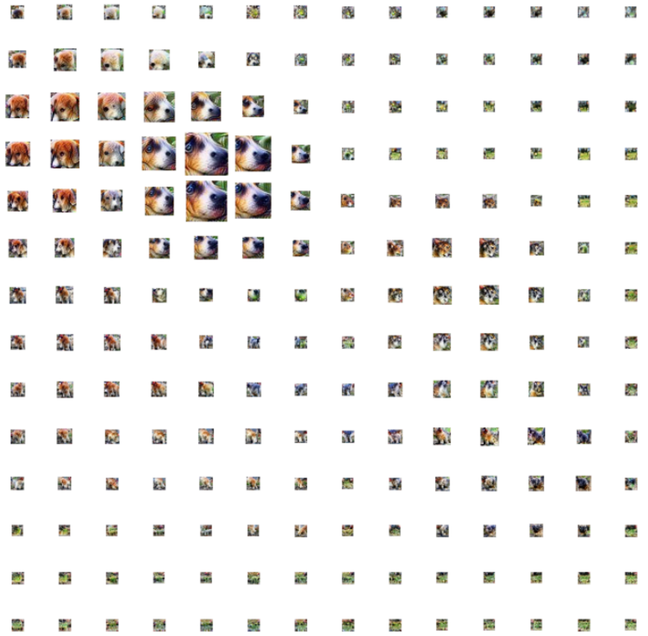

(kind of like local caricatures…)

Visualizing Higher-Layer Features of a Deep Network, Erhan et al, 2009

The building blocks of interpretability, Olah et al, 2018

Architecture	Differentiable white box embedder
Required data	None
Explains	Sample
Focus	Layer
Use for…	Intuition

Activation Maximisation + scaling



The “activation atlas”:
https://distill.pub/2019/activation-atlas/

Visualizing Higher-Layer Features of a Deep Network, Erhan et al, 2009

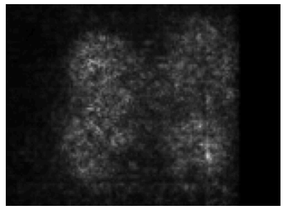
The building blocks of interpretability, Olah et al, 2018

Visualisations

Saliency Maps

Architecture	Differentiable white box
Required data	None
Explains	Sample
Focus	Class in sample
Use for…	Intuition

Backprop to input, no train



Activation (Neuron) Maximisation

Architecture	Differentiable white box embedder
Required data	None
Explains	Model
Focus	Neuron
Use for…	Intuition

Neuron-specific optimisation of input

Activation Vector Maximisation

Architecture	Differentiable white box embedder
Required data	None
Explains	Sample
Focus	Layer
Use for…	Intuition

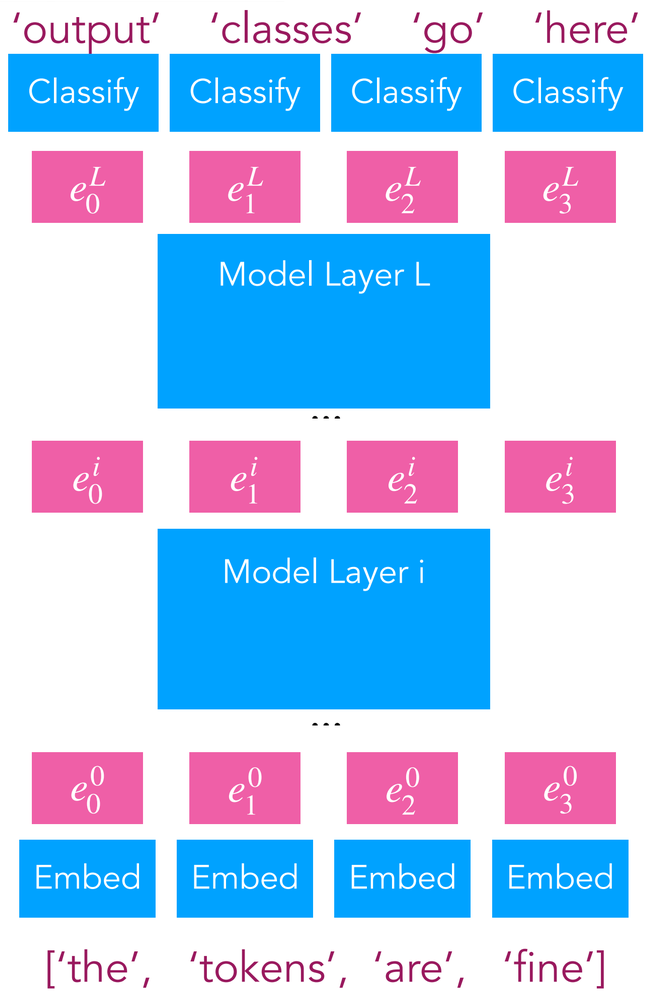
Sample-specific optimisation of input

Nice images, I have convinced myself of all my pre existing beliefs 🙏

Evaluating partial computations
Attention

Architecture	White box with attention
Required data	Annotated task
Explains	Model
Focus	Partial Computation
Use for…	Evaluating hypotheses

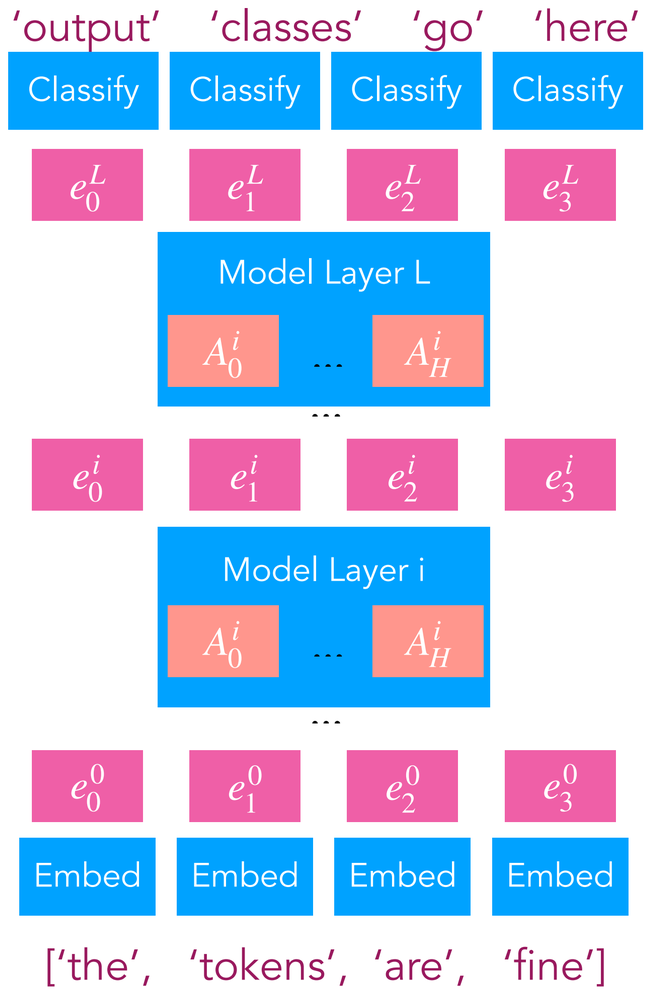
Specialised Attention Heads



Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned, Voita et al, 2019

Architecture	White box with attention
Required data	Annotated task
Explains	Model
Focus	Partial Computation
Use for…	Evaluating hypotheses

Specialised Attention Heads



In trained NMT encoder-decoder models, encoder contains:

Positional Heads: 90% of time, max focus is +1 or -1

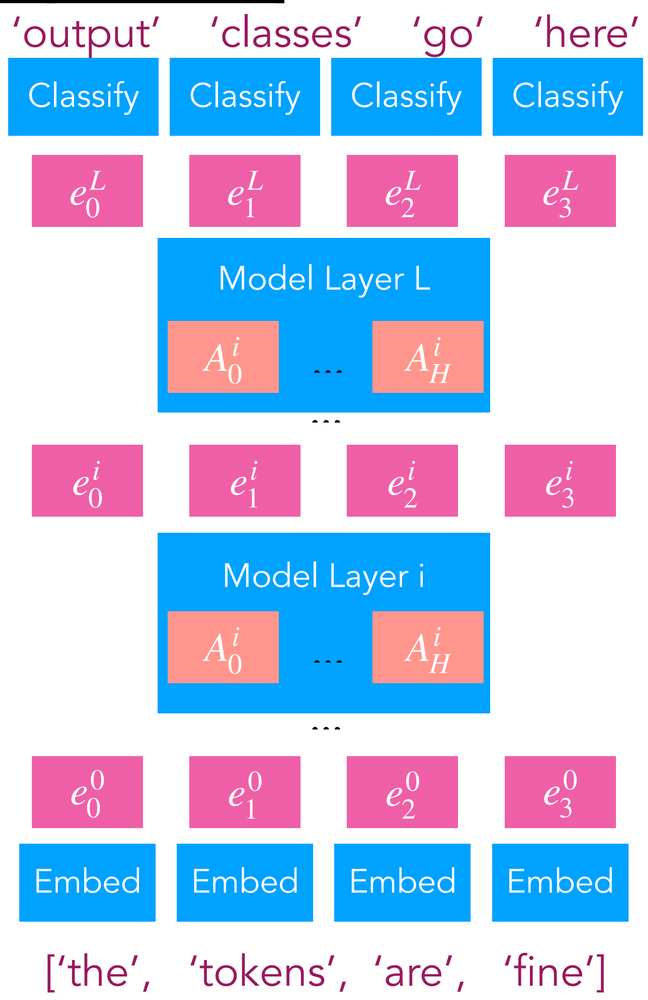
Syntactic Heads: 10% higher than majority baseline for some dependency relation (e.g. advmod: adverb modifier)

Rare Word Head: in sequences with tokens not from top 500, frequently attends to rarest token(s) (post-hoc description..)

Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned, Voita et al, 2019

Architecture	White box with attention
Required data	Annotated task
Explains	Model
Focus	Partial Computation
Use for…	Evaluating hypotheses

Specialised Attention Heads



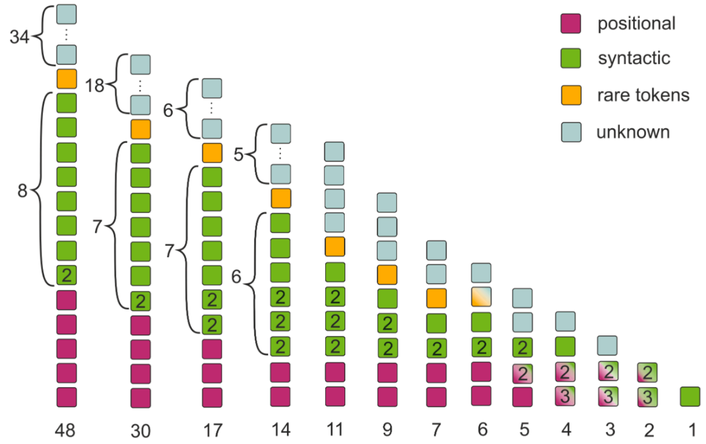
Finding: In trained NMT encoder-decoder models, encoder contains:

Positional Heads: 90% of time, max focus is +1 or -1

Syntactic Heads: 10% higher than majority baseline for some dependency relation (e.g. advmod: adverb modifier)

Rare Word Head: in sequences with tokens not from top 500, frequently attends to rarest token(s)

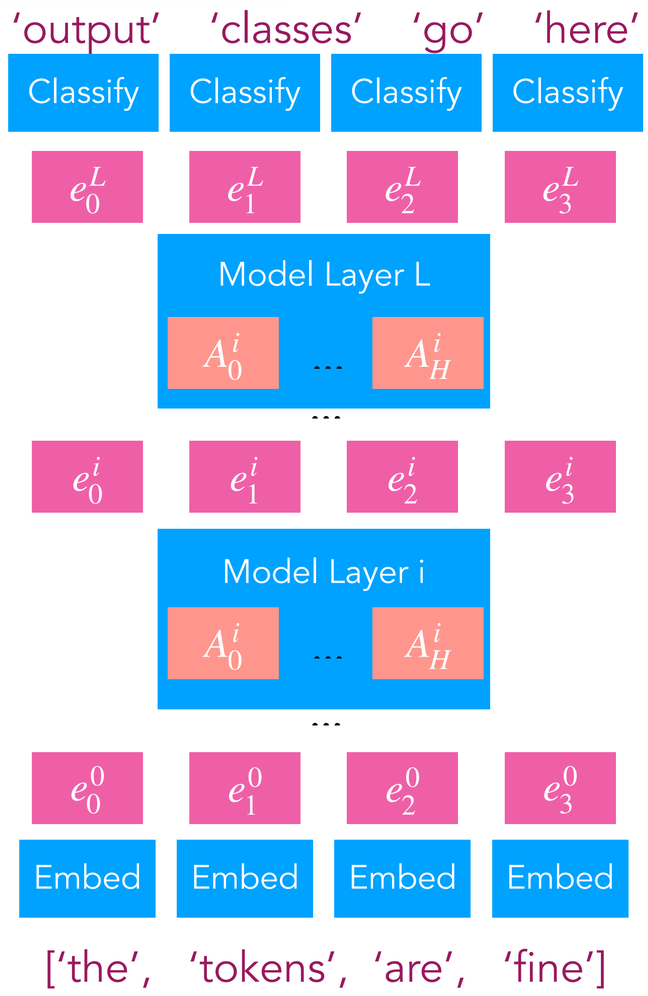
When pruning the network,
these heads are removed last!



Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned, Voita et al, 2019

Architecture	White box with attention
Required data	Annotated task
Explains	Model
Focus	Partial Computation
Use for…	Evaluating hypotheses

Specialised Attention Heads



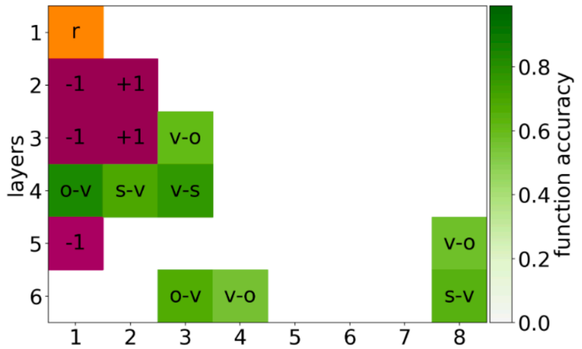
Finding: In trained NMT encoder-decoder models, encoder contains:

Positional Heads: 90% of time, max focus is +1 or -1

Syntactic Heads: 10% higher than majority baseline for some dependency relation (e.g. advmod: adverb modifier)

Rare Word Head: in sequences with tokens not from top 500, frequently attends to rarest token(s)

We see correlation between LRP score and having a clear task



Visualising and Understanding Neural Machine Translation, Ding et al, 2017

Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned, Voita et al, 2019

Evaluating partial computations

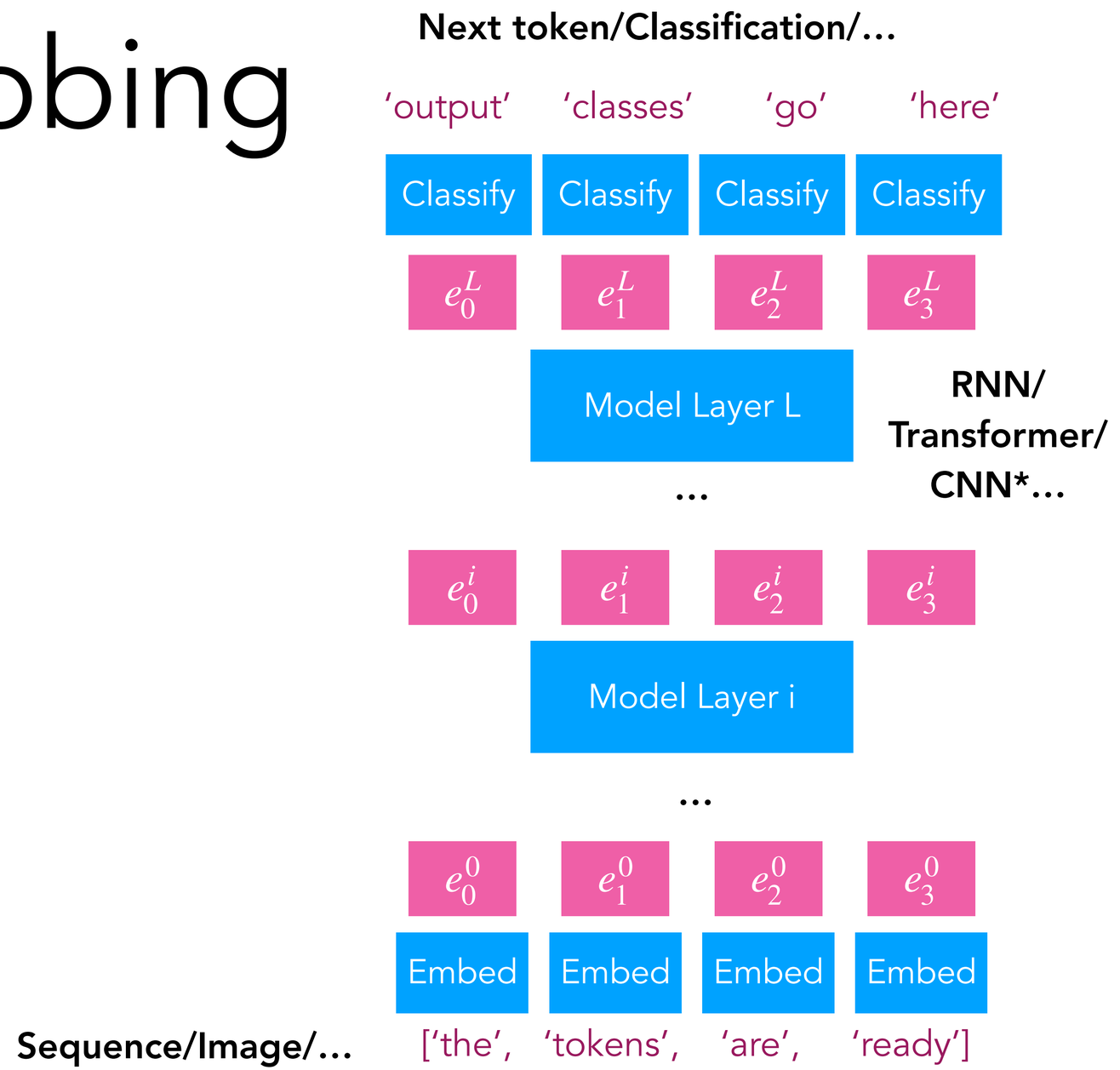
Embeddings

Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for…	Evaluating hypothesis, Single samples if good

Probing

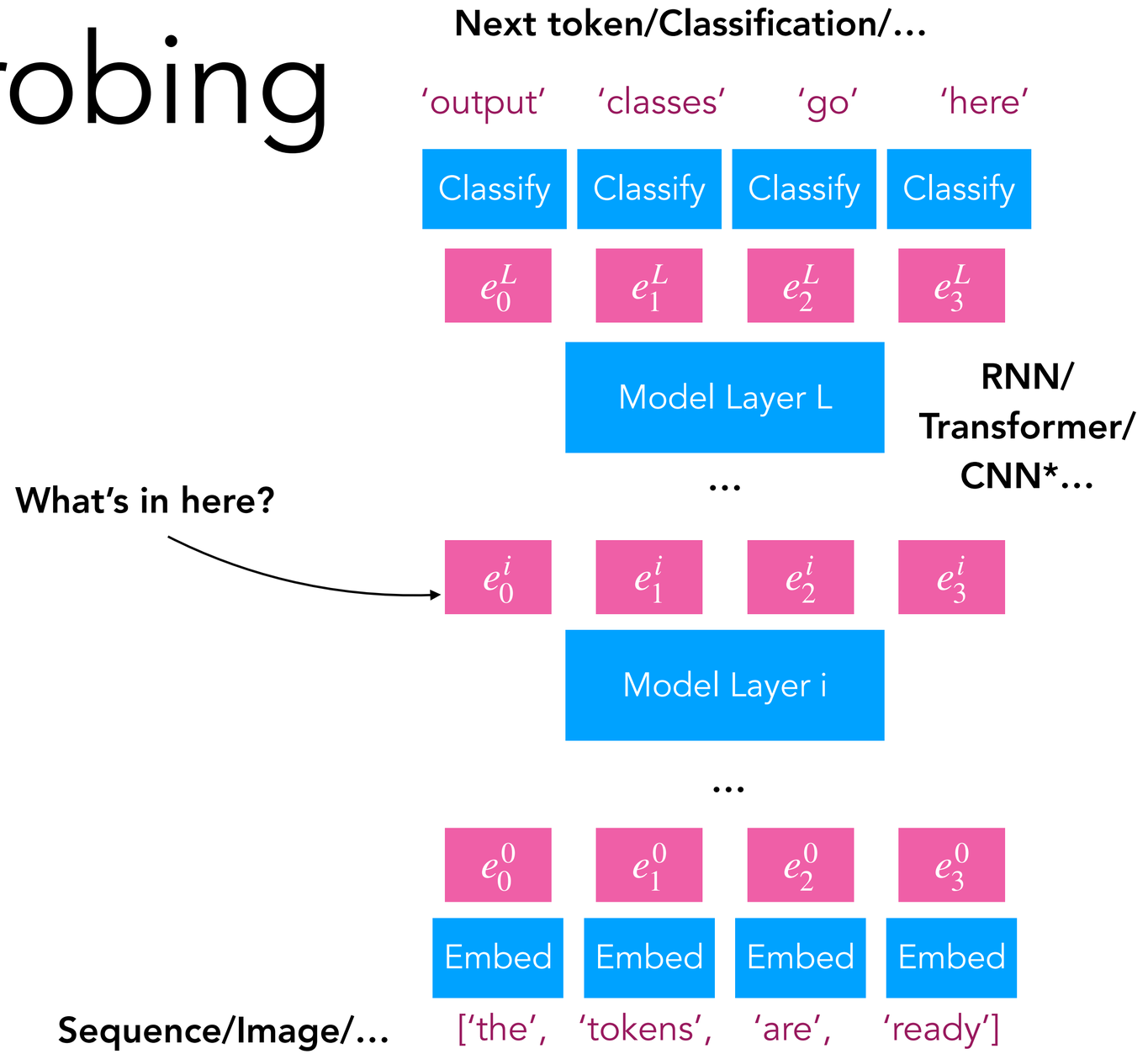
Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for…	Evaluating hypothesis, Single samples if good

Probing



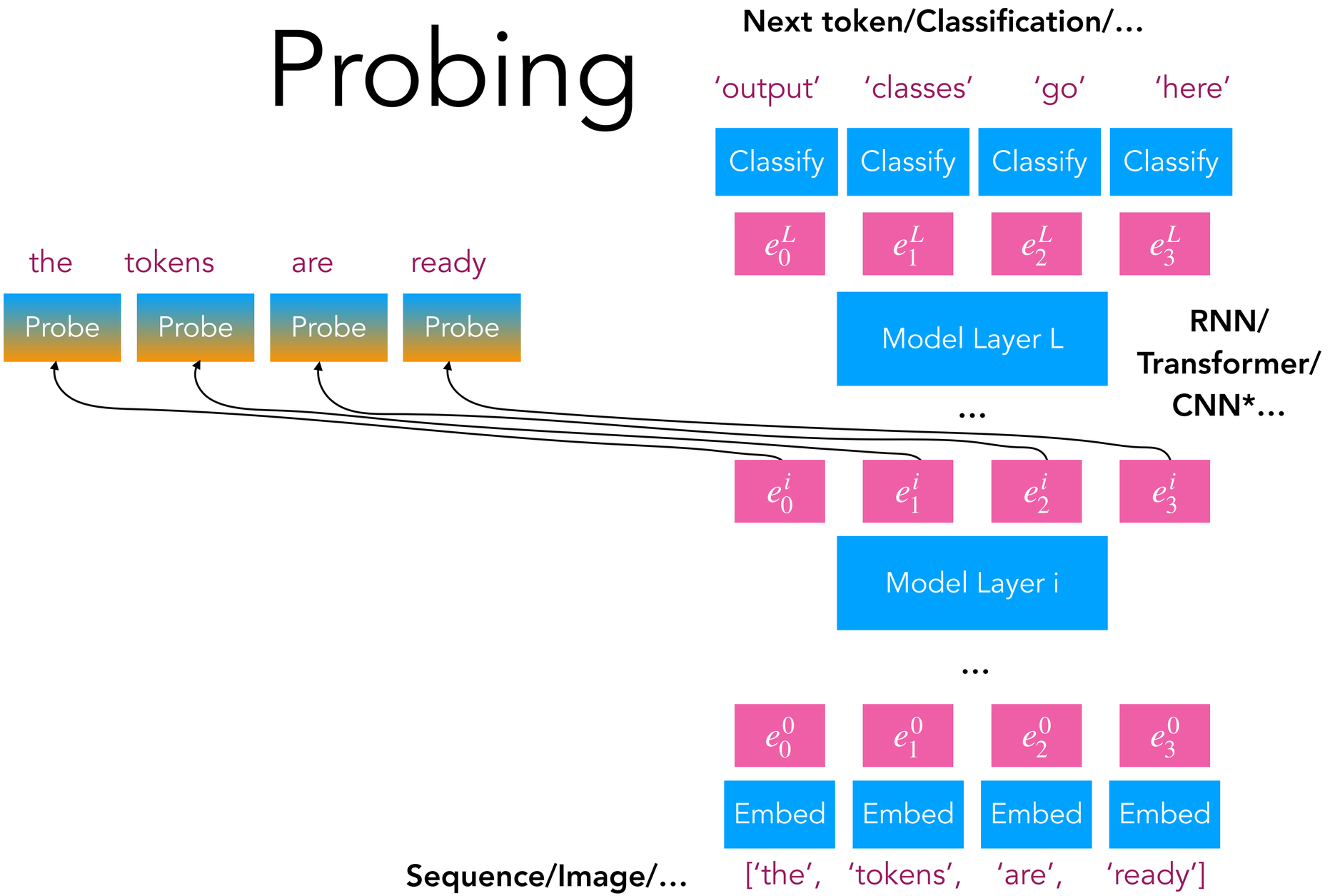
Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for…	Evaluating hypothesis, Single samples if good

Probing



Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for…	Evaluating hypothesis, Single samples if good

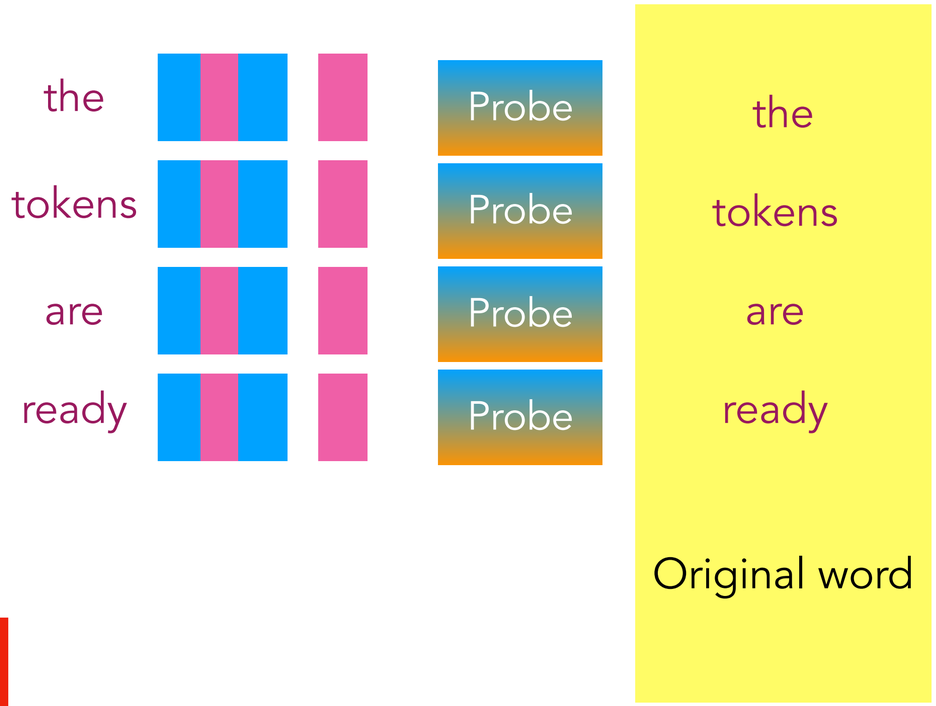
Probing



Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for…	Evaluating hypothesis, Single samples if good

Probing

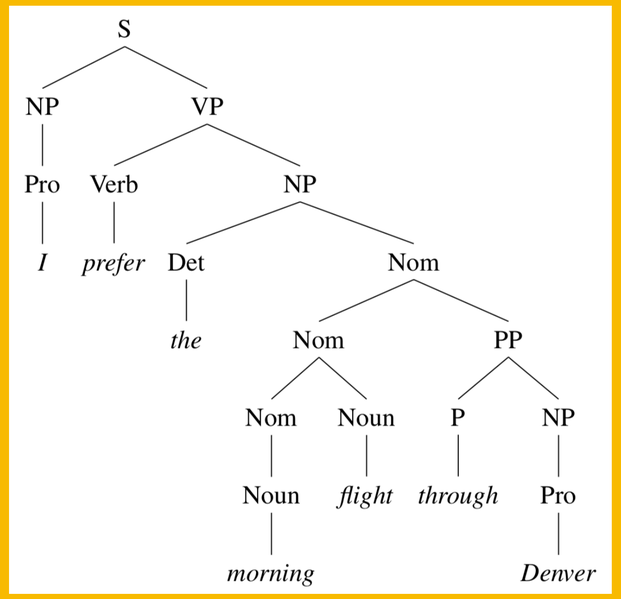
Example: Word-level Tasks



Classical NLP Tasks

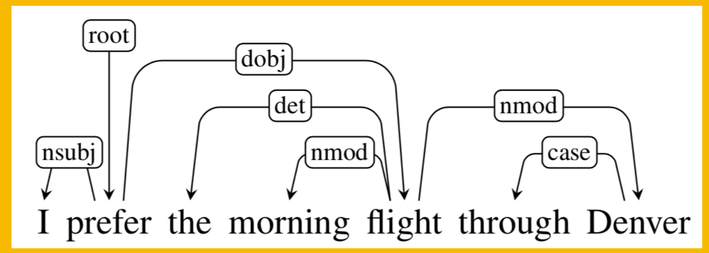
Explicitly assigning structure to sentences

Constituency Parsing



e.g.:
Penn Treebank
Marcus et al, 1999

Dependency Parsing



e.g.:
Universal Dependencies
Nivre et al, 2017

Named Entity Recognition

[Jim]$_{person}$ prefers the [6am]$_{time}$ flight from [Denver]$_{place}$

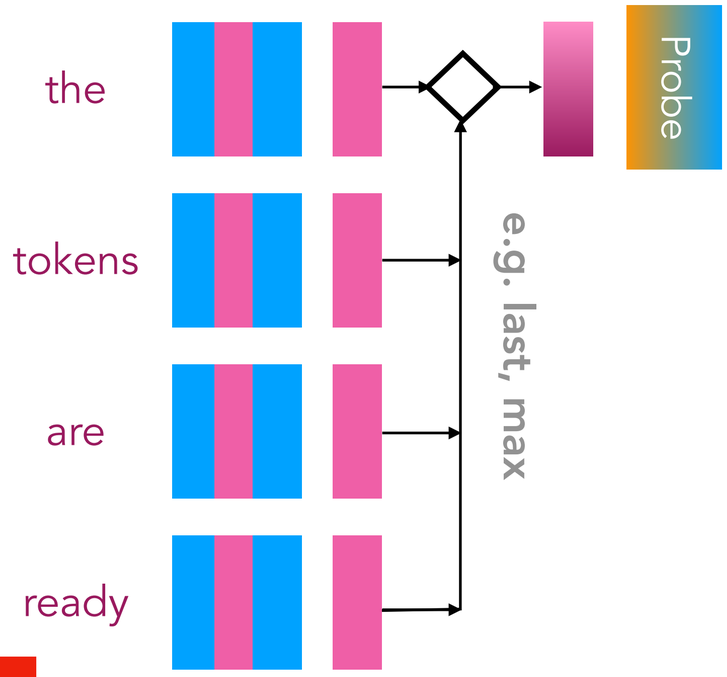
e.g.:
CoNLL-2003 shared task: Language Independent NER
Tjong et al, 2003

NLP researchers invested a lot of effort in solving these tasks explicitly, in pursuit of successful sentence parsing

Do modern models solve them too, or are they doing something else?

Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for...	Evaluating hypothesis, Single samples if good

Probing



Examples: Sentence-level Tasks

Surface level

[<10] length [input] special word

Syntactic

[OK] bigram shift [1] tree depth [NP, VP] topConst …

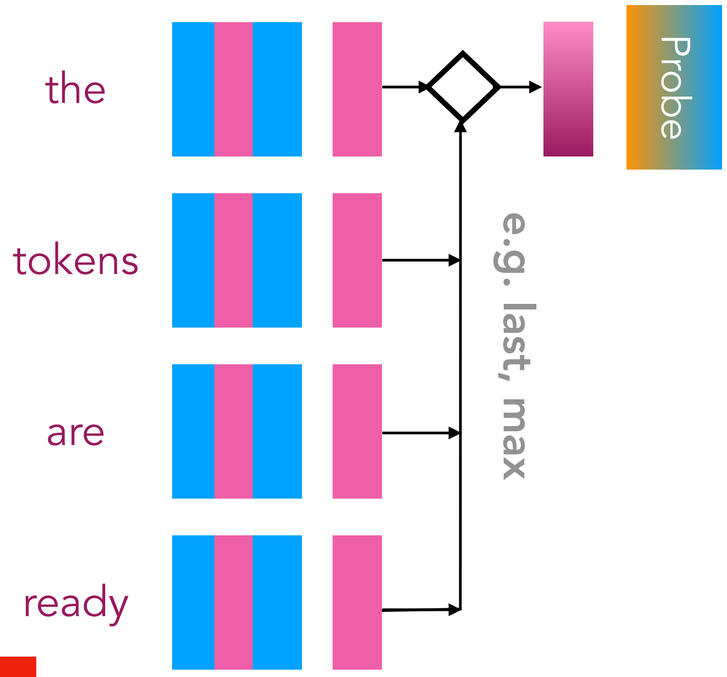
Semantic

plural subjNum [OK] SOMO

What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties, Conneau et al, 2018

Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for…	Evaluating hypothesis, Single samples if good

Probing



Examples: Sentence-level Tasks

Surface level

[<10] length

[input] special word

Syntactic

[OK] bigram shift

[1] tree depth

[NP, VP] topConst

…

Semantic

plural subjNum

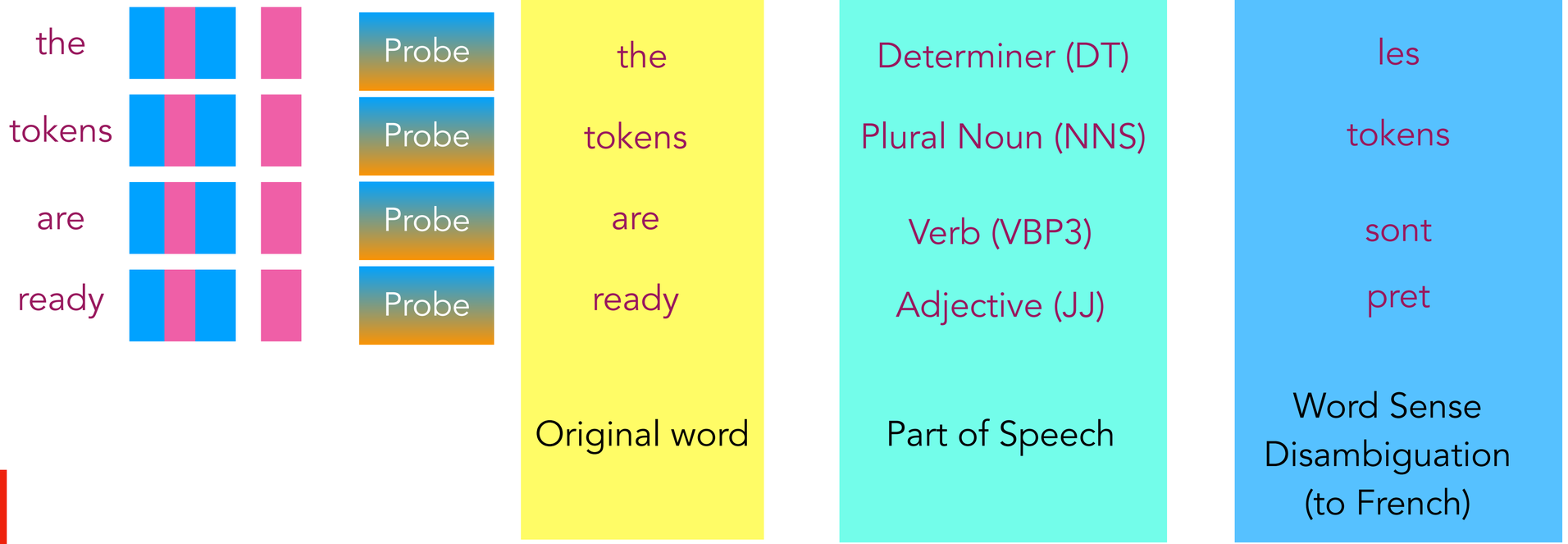
[OK] SOMO

Note: Can also probe over subspans (“Edge Probing”):
What do you learn from context? Probing for sentence structure in contextualized word representations, Tenney et al, 2019

Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for…	Evaluating hypothesis, Single samples if good

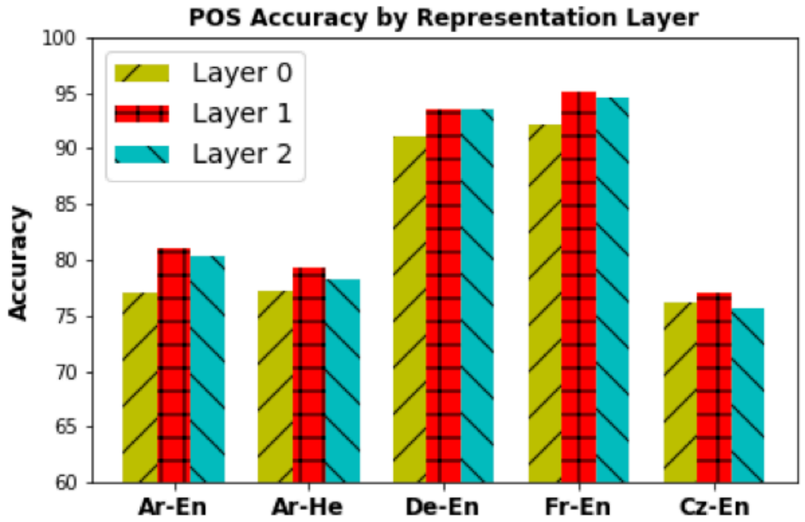
Probing

Example: Word-level Tasks



Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for...	Evaluating hypothesis, Single samples if good

Probing



The embeddings of trained Neural Networks contain ‘solutions’ to a variety of classical NLP tasks

Lower layers appear better at syntax (structure), higher layers at semantics (meaning)

What do Neural Machine Translation Models Learn about Morphology?, Belinkov et al, 2017

Evaluating Layers of Representation in Neural Machine Translation on Part-of-Speech and Semantic Tagging Tasks, Belinkov et al, 2018

Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for...	Evaluating hypothesis, Single samples if good

Probing

Is BERT solving classical NLP tasks as part of its solution? Is it solving them in the order we expect?

Use for...	Localise behaviour to layer

Scalar Mixing Weights

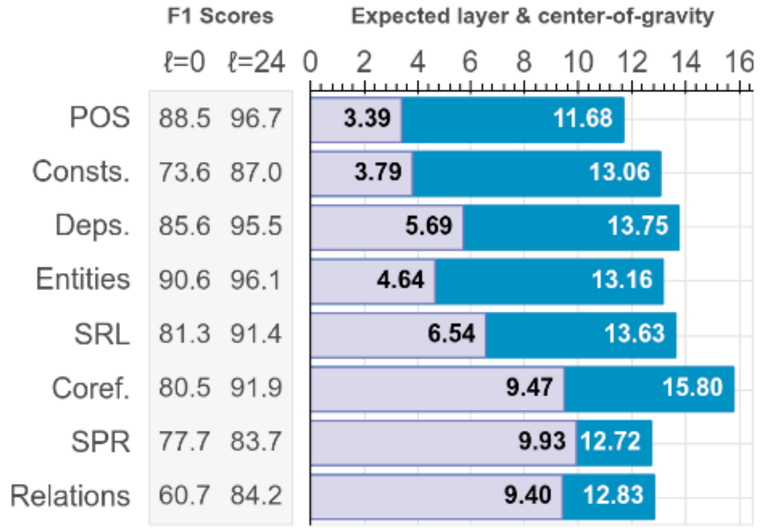
Apply probe to linear combination of the embeddings from **all layers**. Learn the coefficients in addition to the probe

high coefficient = high information for task?

Cumulative Scoring

Train num-layers probes. Each probe L does scalar mixing on **all layers up to L**.

high performance at layer L = task has been solved by time of layer L



Light purple: time (layer) by which task is solved
Semantic tasks solved later in model

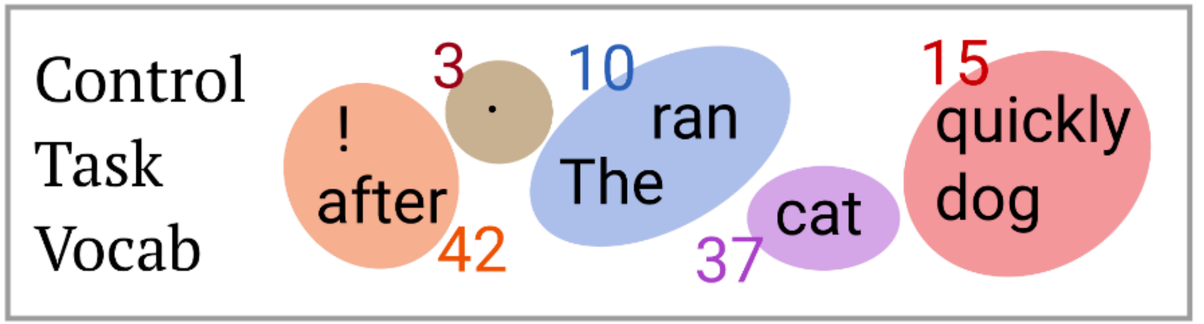
BERT Rediscovers the Classical NLP Pipeline, Tenney et al, 2019

Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for...	Evaluating hypothesis, Single samples if good

Probing

Selectivity

- If our probe is too strong, and our model embeddings very rich, the probe might solve the task regardless of whether the model is doing so itself
- Evaluate the probe on control tasks to stay grounded
  - Design random tasks that can only be learned by memorisation - no generalisation



Sentence 1	The	cat	ran	quickly	.
Part-of-speech	DT	NN	VBD	RB	.
Control task	10	37	10	15	3

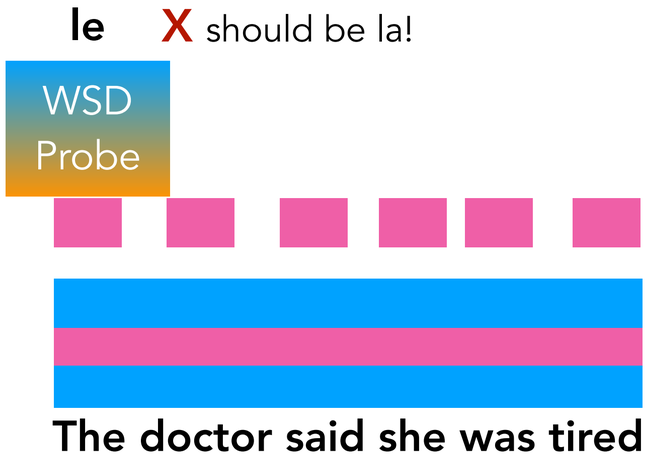
Sentence 2	The	dog	ran	after	!
Part-of-speech	DT	NN	VBD	IN	.
Control task	10	15	10	42	42

Designing and Interpreting Probes with Control Tasks, Hewitt and Liang, 2019

Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for…	Evaluating hypothesis, Single samples if good

Probing

Note: won't give new classes, but could reveal unexpected associations!



Architecture	White box with internal embeddings
Required data	Annotated task
Explains	General Behaviour, Single samples if good
Localises	Partial Computation
Use for…	Evaluating hypothesis, Single samples if good

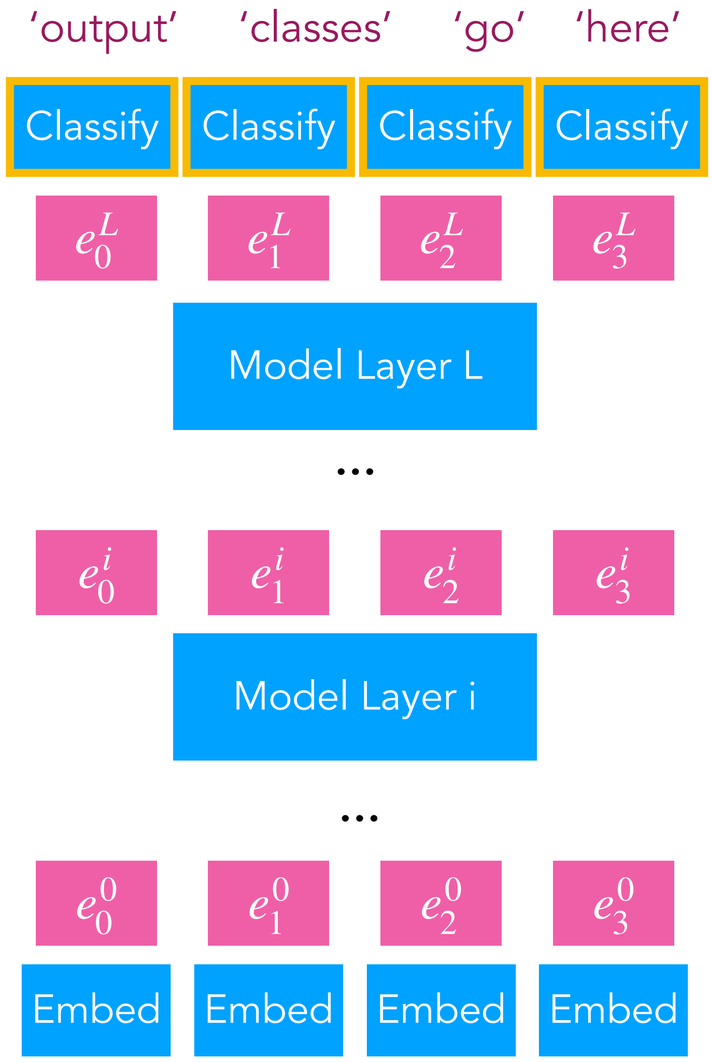
Any ready made probes?

Embedding space

Architecture	White box with internal embeddings
Required data	None
Explains	Global Behaviour, samples when good
Localises	Partial Computation
Use for…	Evaluating hypotheses, samples when good

Embedding Space

aka logit lens

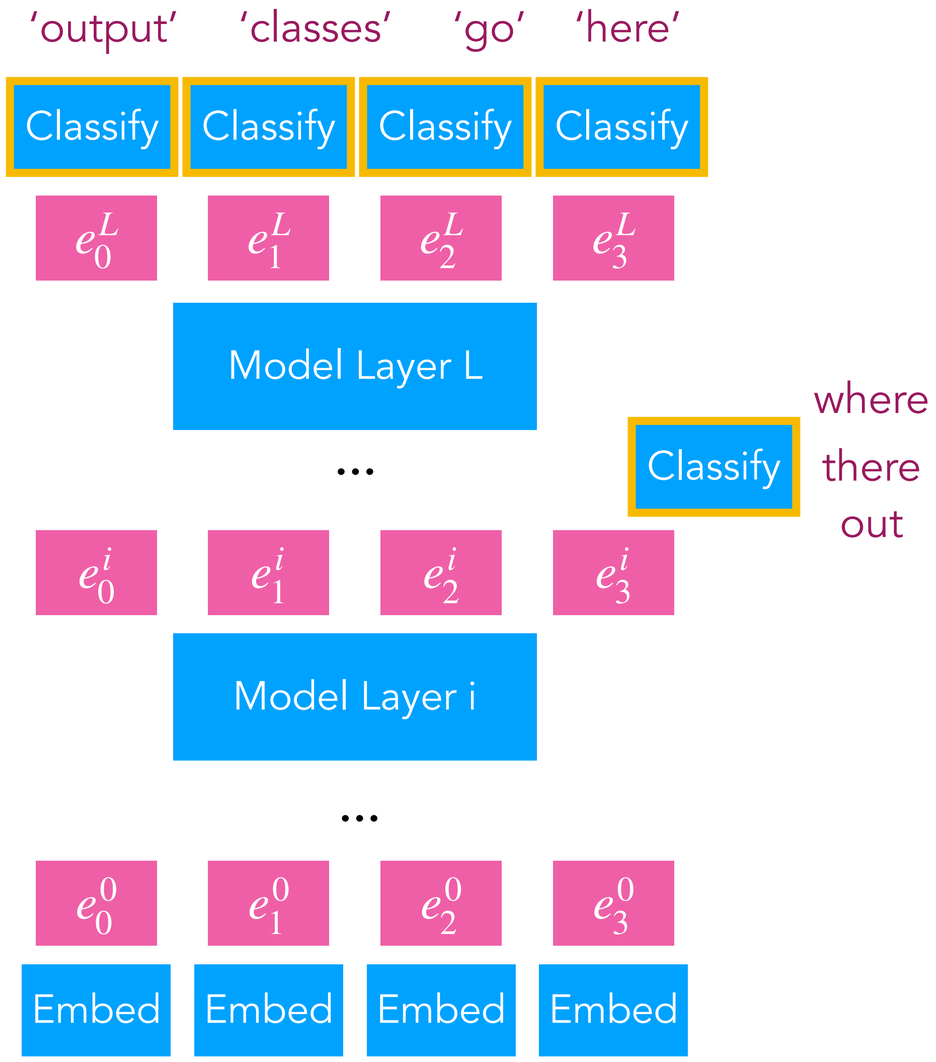


Idea: Treat classifier as probe

Architecture	White box with internal embeddings
Required data	None
Explains	Global Behaviour, samples when good
Localises	Partial Computation
Use for…	Evaluating hypotheses, samples when good

Embedding Space

aka logit lens

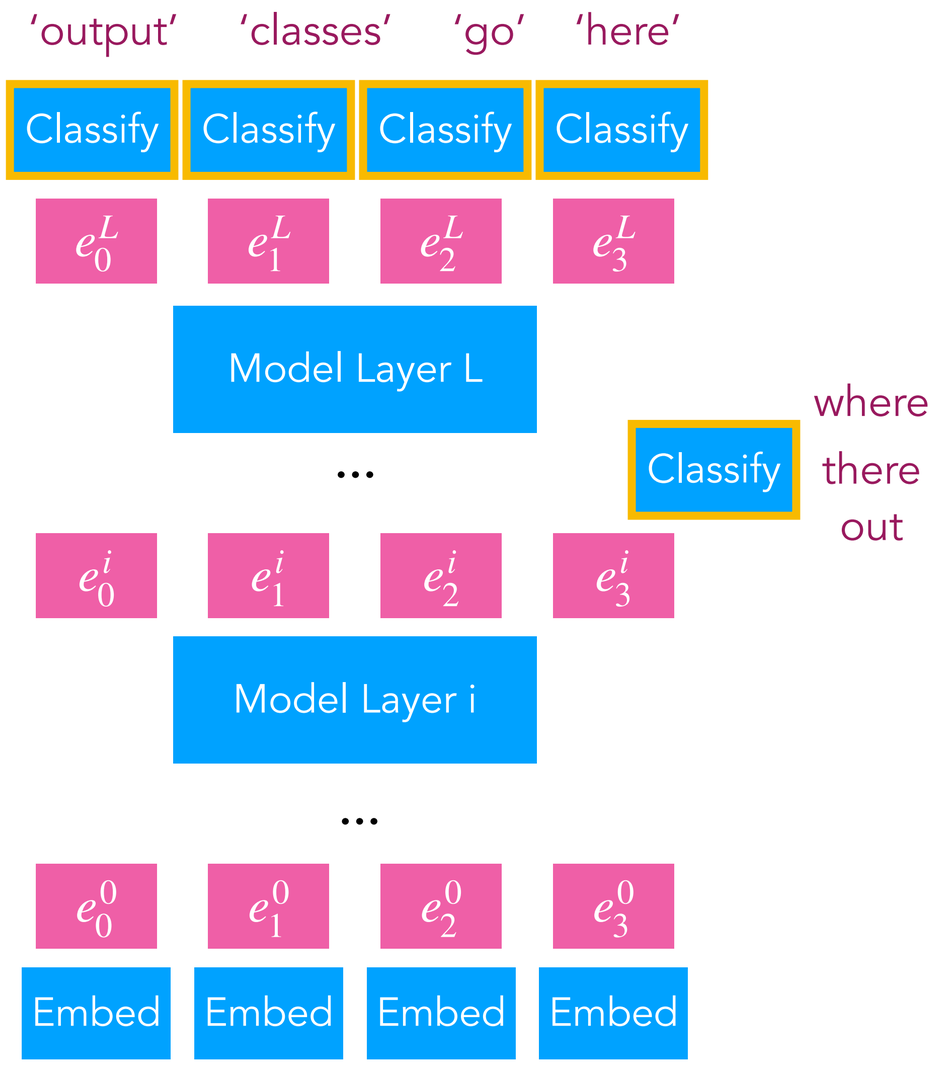


Idea: Treat classifier as probe

Architecture	White box with internal embeddings
Required data	None
Explains	Global Behaviour, samples when good
Localises	Partial Computation
Use for...	Evaluating hypotheses, samples when good

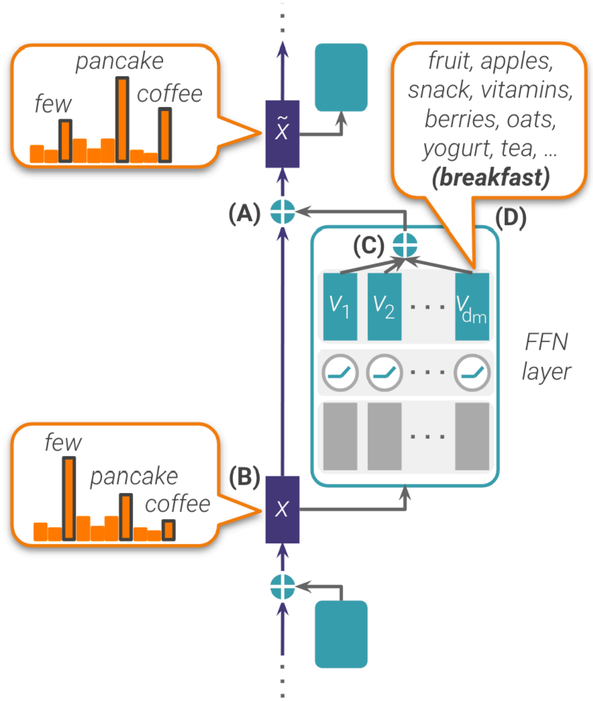
Embedding Space

aka logit lens



Idea: Treat classifier as probe

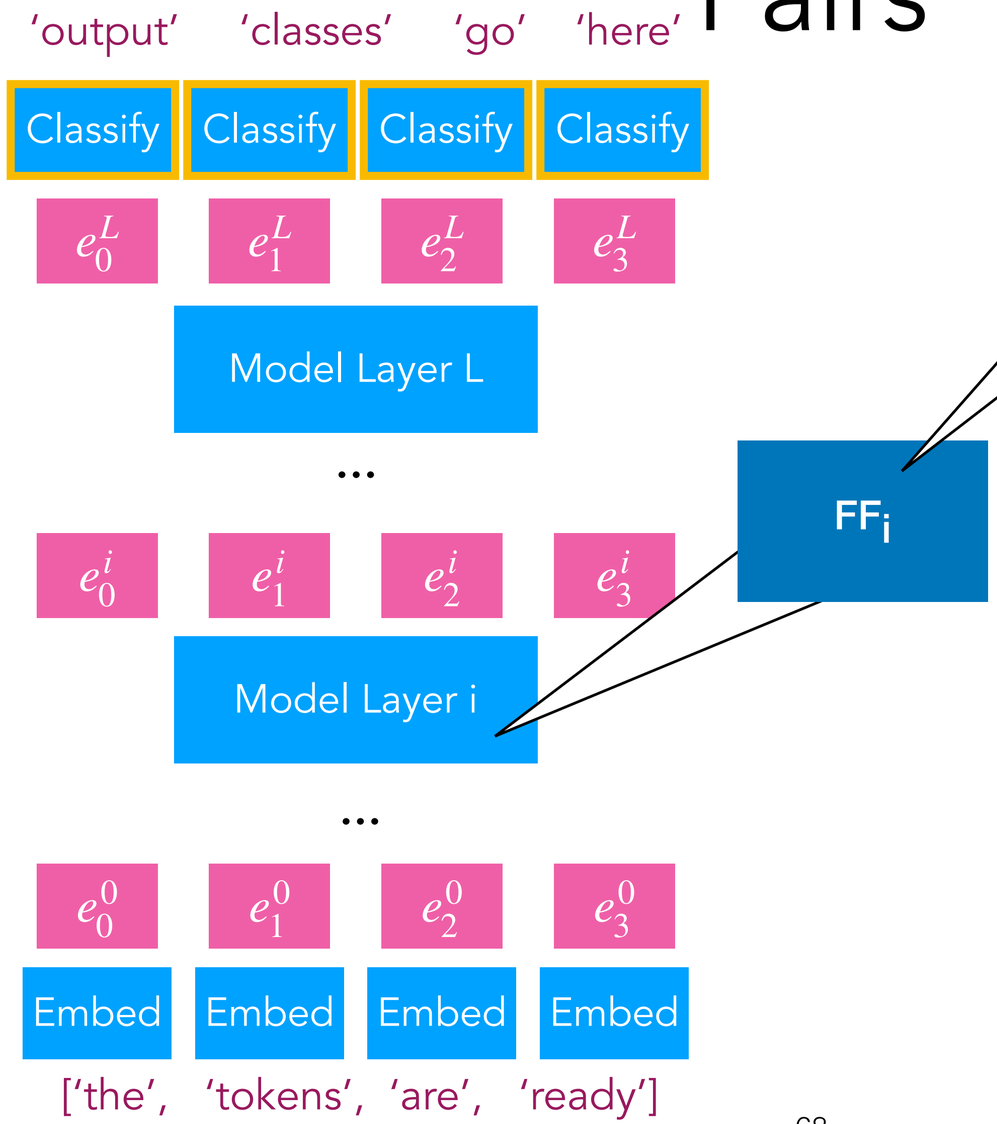
Results seem plausible/interpretable:



Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space, Geva et al, 2022

anything interesting with this?

Transformer FFs as Key-Value Pairs

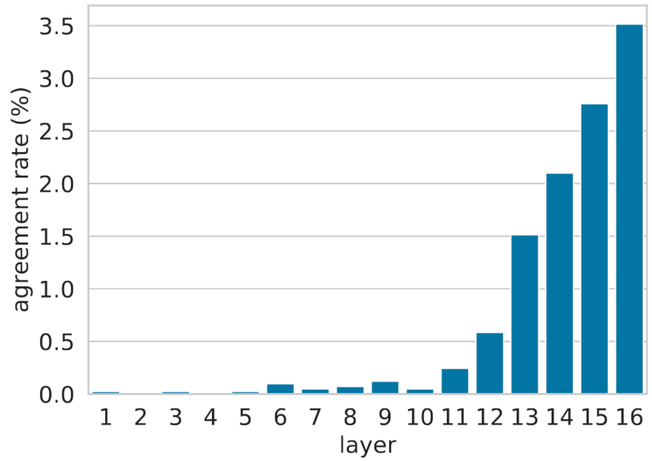


$$FF_i(e) = W_i^B R(W_i^A e)$$

"keys"
(rows respond to "trigger" inputs)

"values"
(columns meaningful in embedding space)

%:
key's "top trigger" next token
==
value's top token



Transformer feed-forward networks are key-value memories, Geva et al, 2021

Was that a hint of associations coming up

Architecture	Black box LM
Required data	Knowledge Triplets
Explains	Model
Localises	No
Use for...	Evaluation of Knowledge

Knowledge

The **LAMA** probe*:
LAnguage Model Analysis

Take knowledge triplet, and present as sentence with object as last token:
(Dante,Born in,Florence)
→ **Dante was born in Florence**

Mask object, feed to LM, and asses ability to recover the object

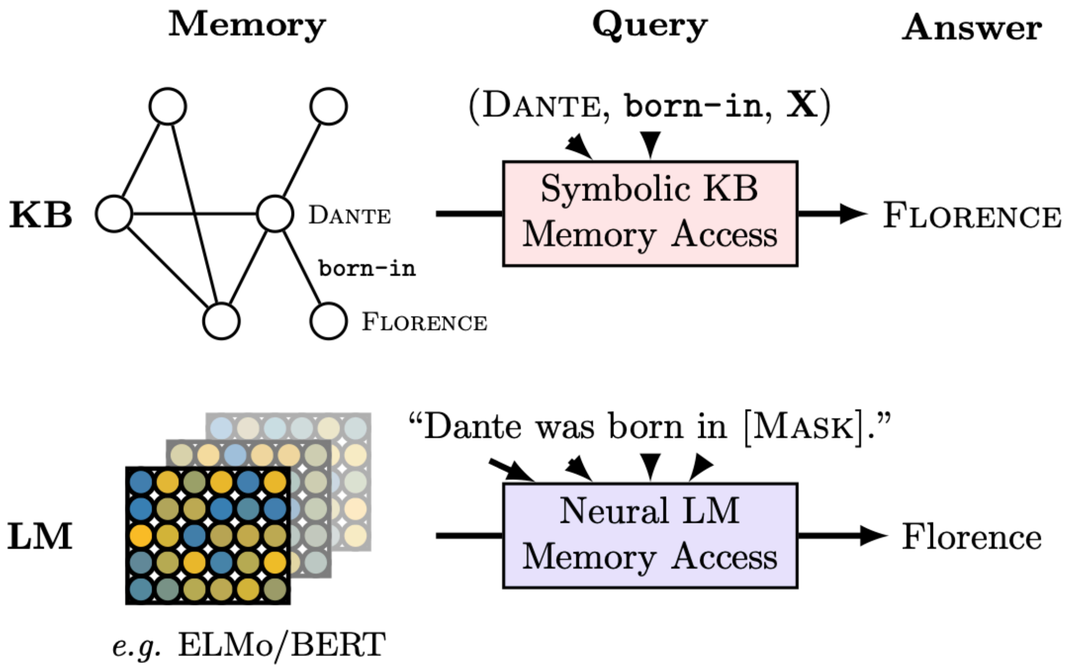


Figure 1: Querying knowledge bases (KB) and language models (LM) for factual knowledge.

*task, prompt

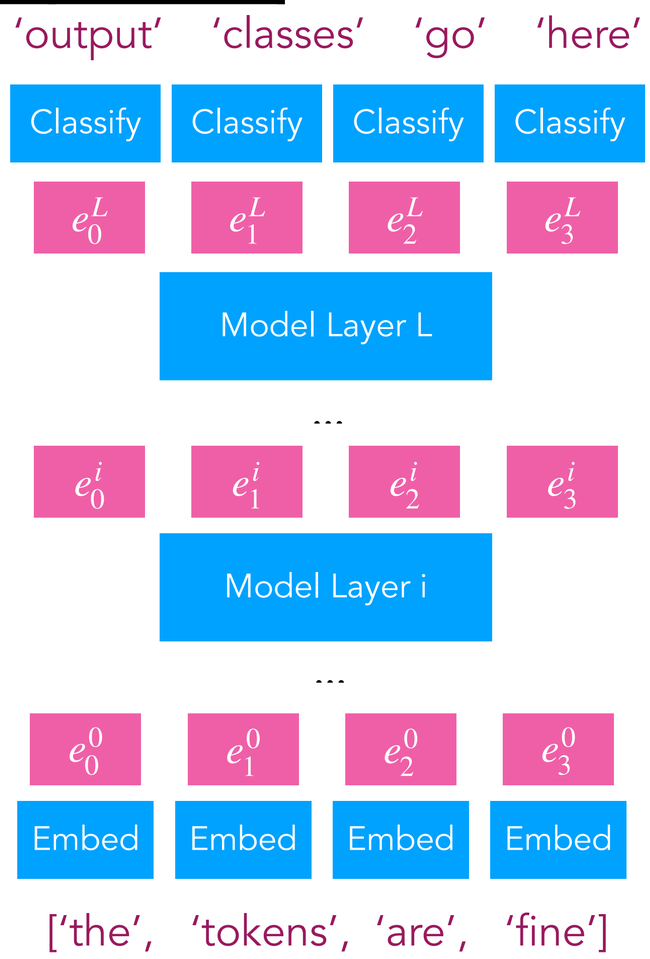
Language Models as Knowledge Bases?, Petroni et al, 2019

Locating knowledge
Interventions

Architecture	White box
Required data	Samples of interest
Explains	General behaviour
Localises	Input x Layer
Use for...	Locating computation

Causal Tracing

AKA Activation Patching



Inspiration: Causal Mediation Analysis, Pearl, 2001

Locating and Editing Factual Associations in GPT, Meng et al, 2022

Architecture	White box
Required data	Samples of interest
Explains	General behaviour
Localises	Input x Layer
Use for...	Locating computation

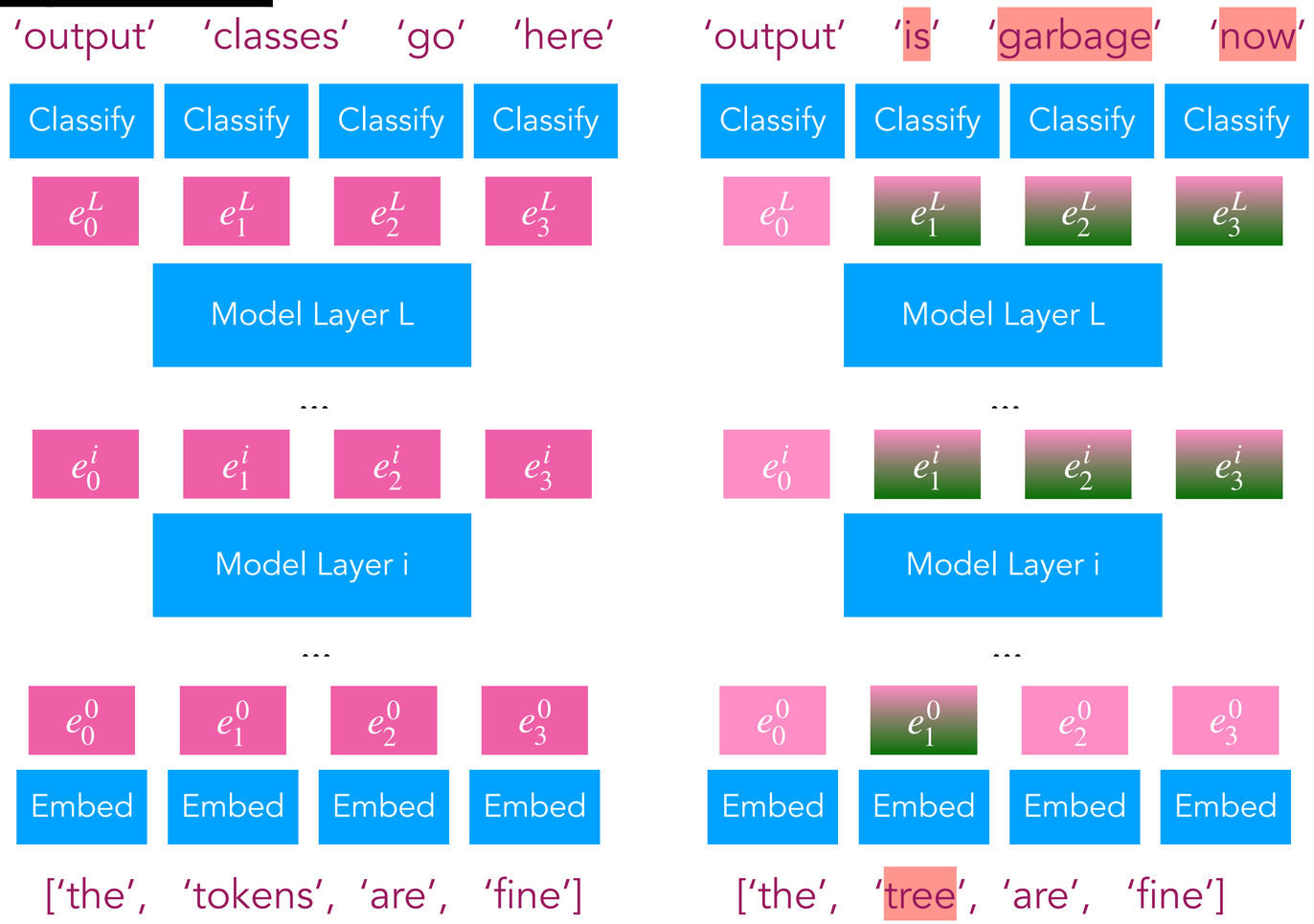
Causal Tracing

AKA Activation Patching

1. corrupt input

corrupt **tokens** or embedding
or FF hidden or attention

$\Longrightarrow$ output gets broken



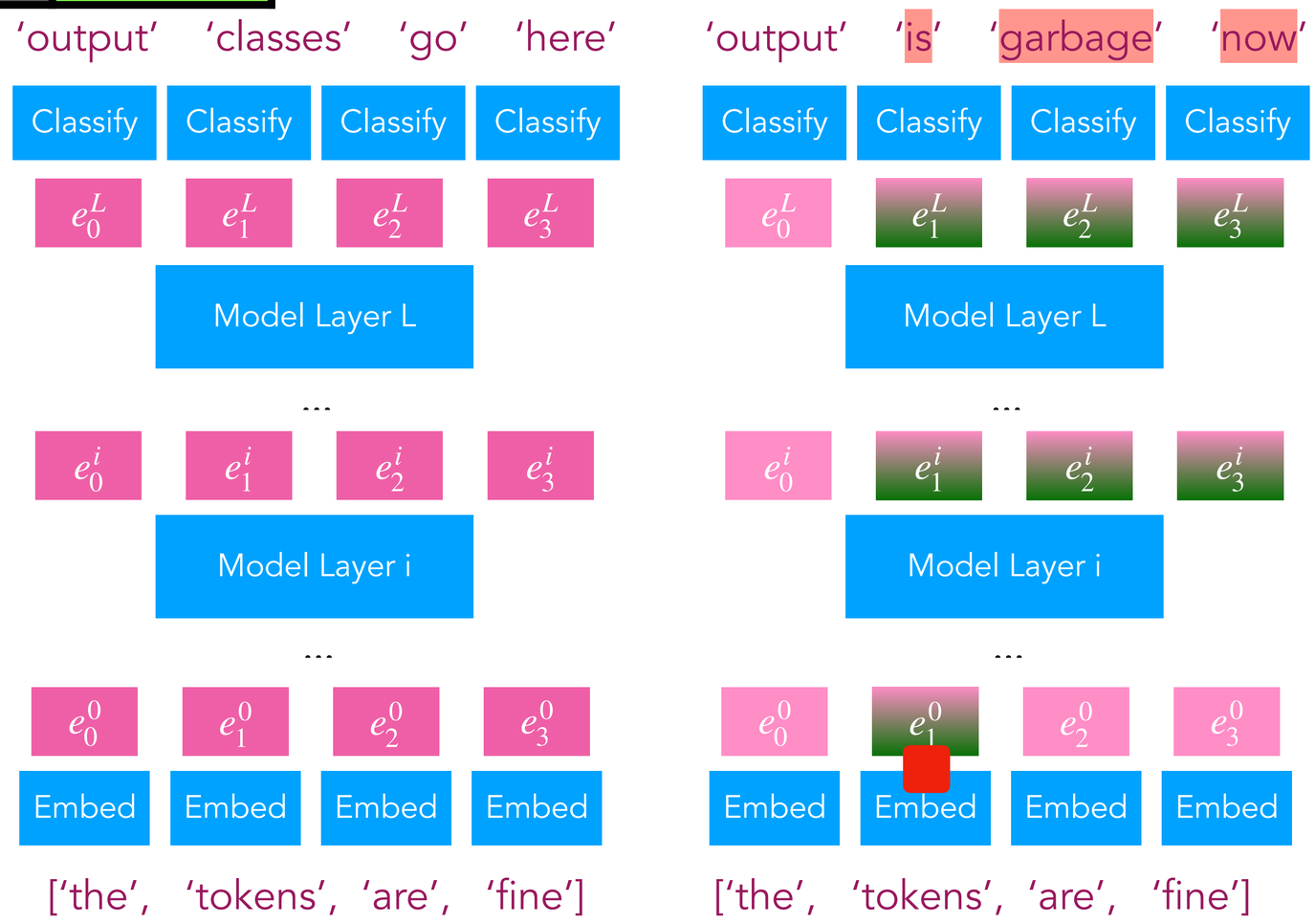
Inspiration: Causal Mediation Analysis, Pearl, 2001

Locating and Editing Factual Associations in GPT, Meng et al, 2022

Architecture	White box
Required data	Samples of interest
Explains	General behaviour
Localises	Input x Layer
Use for...	Locating computation

Causal Tracing

AKA Activation Patching



1. corrupt input

corrupt tokens or **embedding**
or FF hidden or attention

$\Longrightarrow$ output gets broken

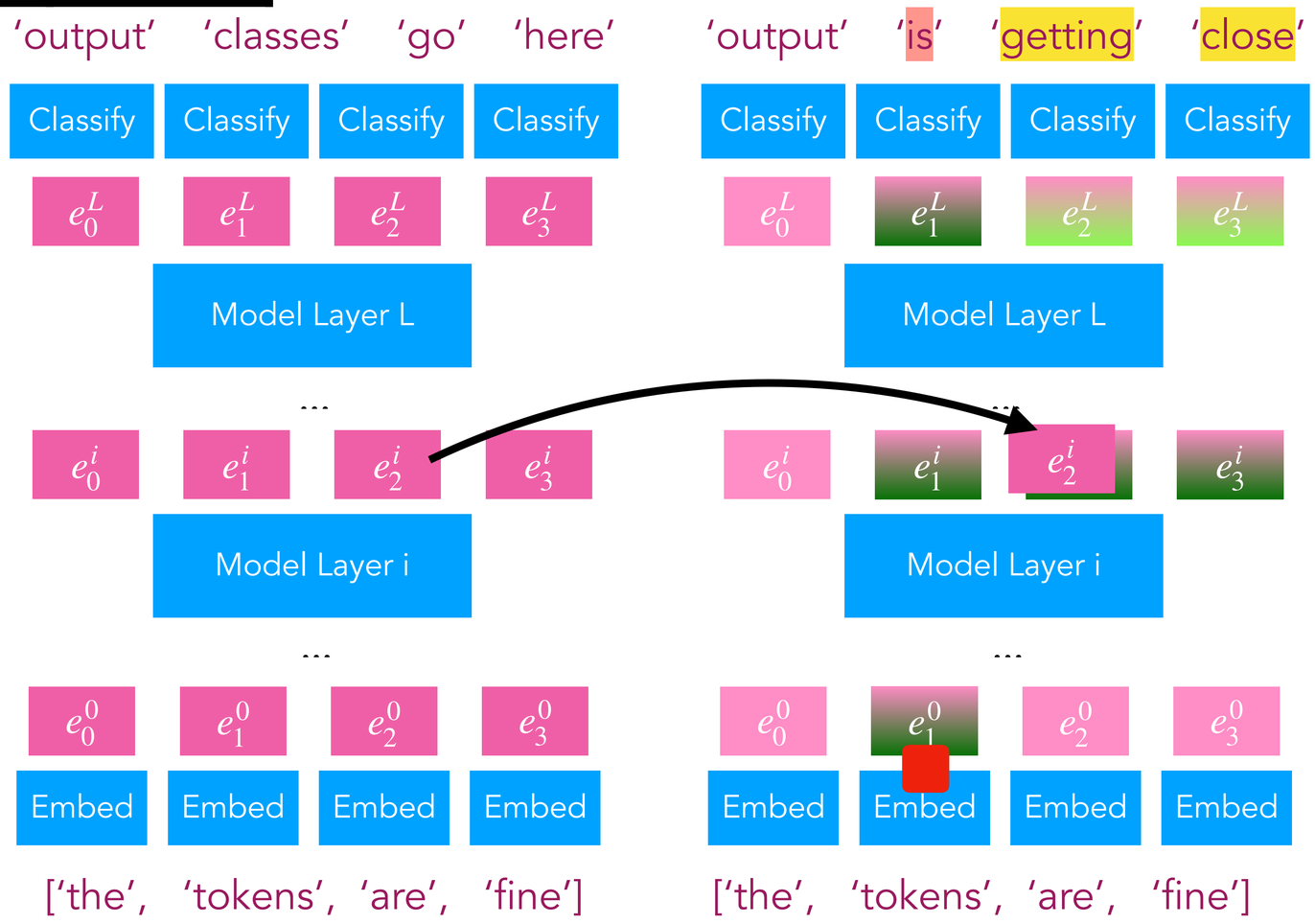
Inspiration: Causal Mediation Analysis, Pearl, 2001

Locating and Editing Factual Associations in GPT, Meng et al, 2022

Architecture	White box
Required data	Samples of interest
Explains	General behaviour
Localises	Input x Layer
Use for...	Locating computation

Causal Tracing

AKA Activation Patching



1. corrupt input

corrupt tokens or embedding
or FF hidden or attention

$\Longrightarrow$ output gets broken

2. patch in embeddings
from correct input

different layers and positions

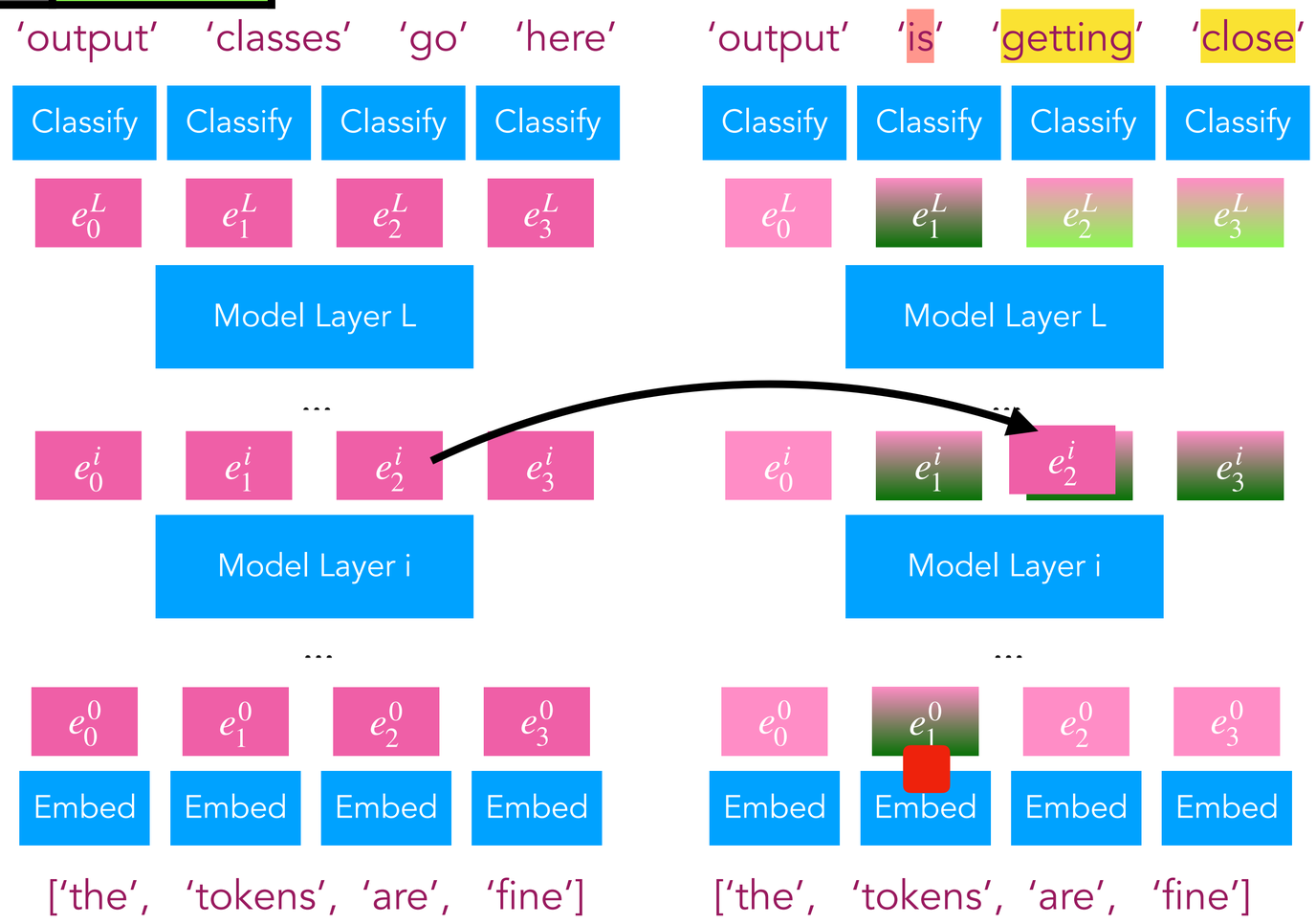
Inspiration: Causal Mediation Analysis, Pearl, 2001

Locating and Editing Factual Associations in GPT, Meng et al, 2022

Architecture	White box
Required data	Samples of interest
Explains	General behaviour
Localises	Input x Layer
Use for...	Locating computation

Causal Tracing

AKA Activation Patching



1. corrupt input

corrupt tokens or embedding
or FF hidden or attention

$\Longrightarrow$ output gets broken

2. patch in embeddings
from correct input

different layers and positions

is output fixed?

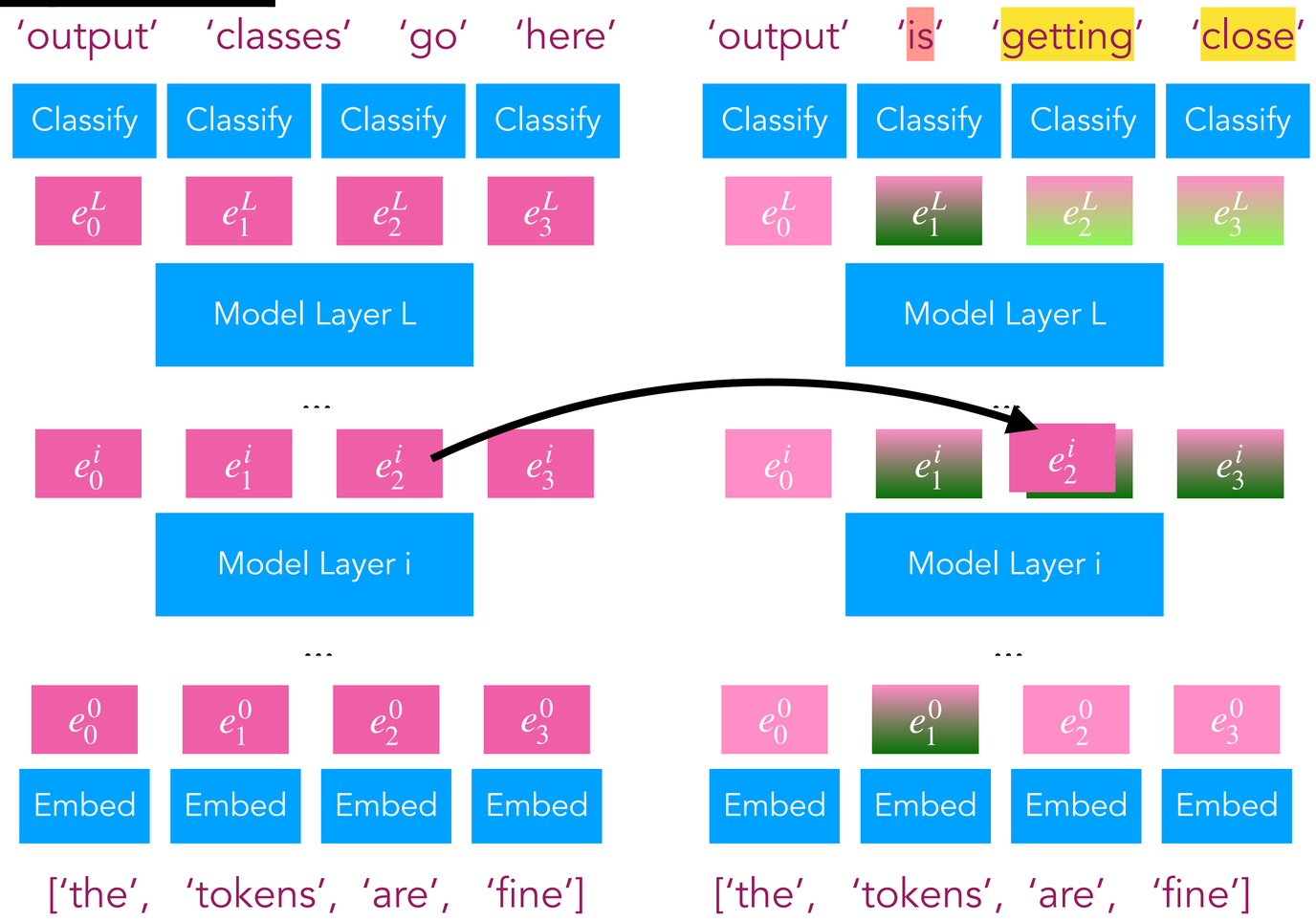
Inspiration: Causal Mediation Analysis, Pearl, 2001

Locating and Editing Factual Associations in GPT, Meng et al, 2022

Architecture	White box
Required data	Samples of interest
Explains	General behaviour
Localises	Input x Layer
Use for...	Locating computation

Causal Tracing

AKA Activation Patching



1. corrupt input

corrupt tokens or embedding
or FF hidden or attention

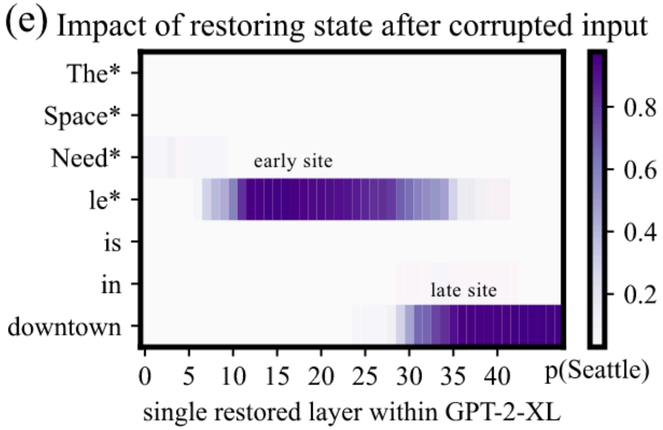
$\Longrightarrow$ output gets broken

2. patch in embeddings
from correct input

different layers and positions

is output fixed?

Result: localise position of associations in network



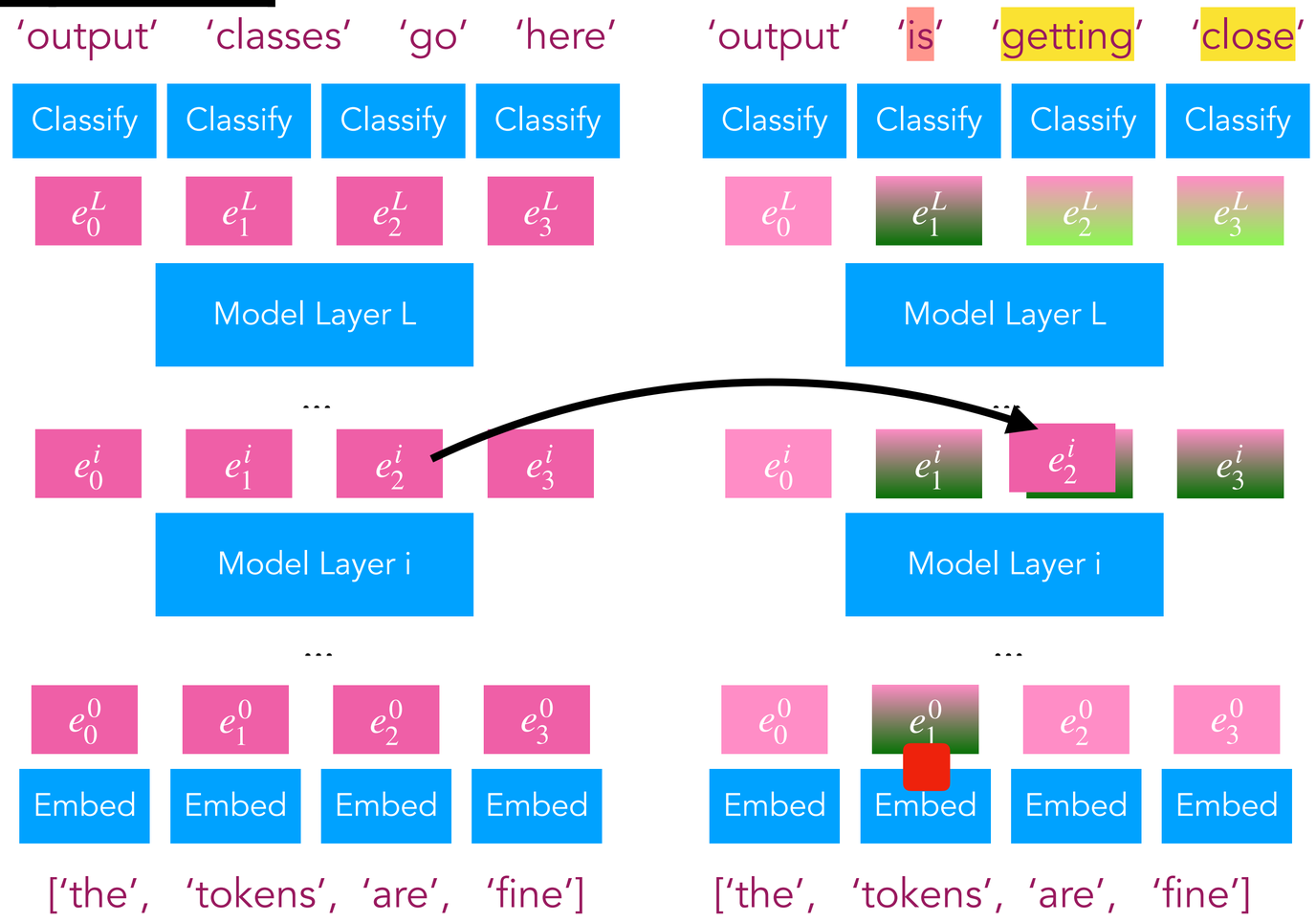
Inspiration: Causal Mediation Analysis, Pearl, 2001

Locating and Editing Factual Associations in GPT, Meng et al, 2022

Architecture	White box
Required data	Samples of interest
Explains	General behaviour
Localises	Input x Layer
Use for...	Locating computation

Causal Tracing

AKA Activation Patching



1. corrupt input

corrupt tokens or embedding
or FF hidden or attention

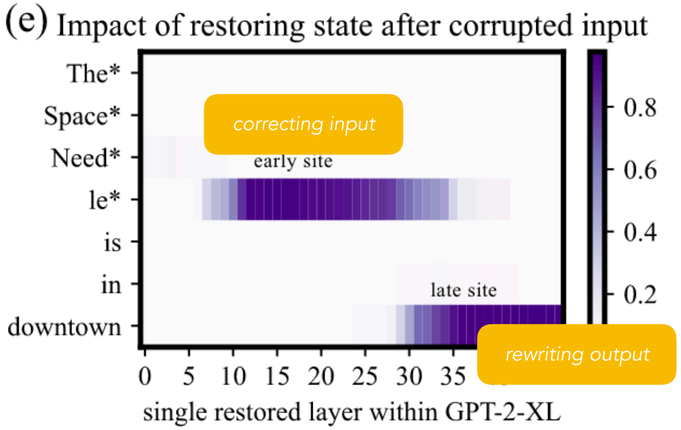
$\Longrightarrow$ output gets broken

2. patch in embeddings
from correct input

different layers and positions

is output fixed?

Result: localise position of associations in network



Inspiration: Causal Mediation Analysis, Pearl, 2001

Locating and Editing Factual Associations in GPT, Meng et al, 2022

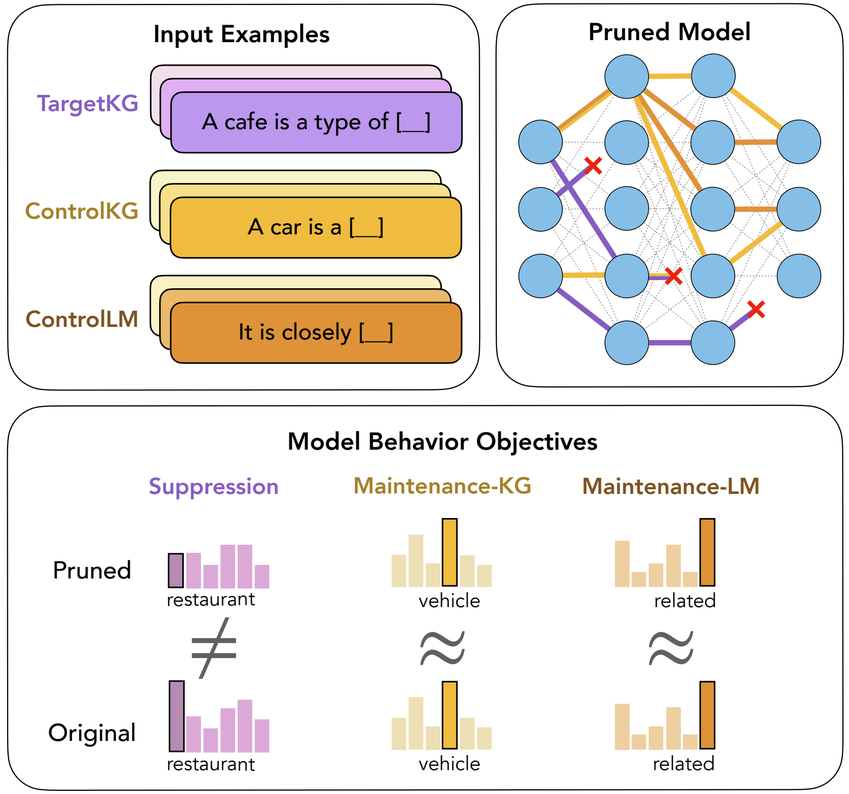
Locating knowledge
Modifications

Architecture	White box LM
Required data	Knowledge Phrases
Explains	Model
Focus	Knowledge
Use for…	Blocking knowledge

Knowledge-Critical Subnets

Learn a mask over model parameters to remove knowledge

Optimise for:
Keep mask minimal
+
Remove target knowledge
+
Maintain other behaviours



*task, prompt

Discovering Knowledge Critical Subnetworks in Pretrained Language Models, Bayazit et al, 2023

Breathe

Recap

	Leave one out	LIME	SHAP	Layerwise Relevance Propagation	Saliency Maps	Activation Neuron Maximisation	Activation Vector Maximisation	Attention head eval	Probing	Embedding Space	LAMA	Causal Tracing	Knowledge Critical Subnets
Architecture	Black box	Black box	Black box	White box (with layers)	Differentiable white box	Differentiable white box embedder	Differentiable white box embedder	White box with attention	White box with internal embeddings	White box with internal embeddings	Black box LM	White box	White box
Required data	None	None	Train Samples	None	None	None	None	Annotated task	Annotated task	None	Knowledge Triplets	No	Knowledge Phrases
Explains	Sample area	Sample area	Model	Sample	Sample	Model	Sample	Model	Model, single samples if good	Model, single samples (when good)	Model	Model	Model
Focus	All features	All features	All features	All sample (locates class)	All sample (locates class)	Neuron	Layer	Partial Computation	Partial Computation	Partial Computation	No	Input X Layer	All weights
Use for…	Intuition	Intuition	Intuition	Intuition	Intuition	Intuition	Intuition	Evaluating hypotheses	Evaluating hypotheses, single samples if good	Evaluating hypotheses, single samples (when good)	Assert presence of/ extract knowledge	Locating computation/ knowledge	Blocking knowledge

Used creatively for attention relevance!

Not in here: pruning/ablations in general: knocking out whole sub-components to see if model needs them

Outline

- ✦ Introduction
- ✦ Methods and Concepts
  - Several black and white box methods, intuition vs hypotheses, observation vs intervention
  - Background interrupt: Classical NLP tasks
- ✦ Friends
  - Mechanistic interpretability; Formal analysis; Extraction
- ✦ Conclusion

I keep hearing about mechanistic interpretability

Mechanistic Interpretability

Circuits

Individual subcomponents of the transformer computation, ideally with intuitice meaning - e.g. induction heads

Induction Heads

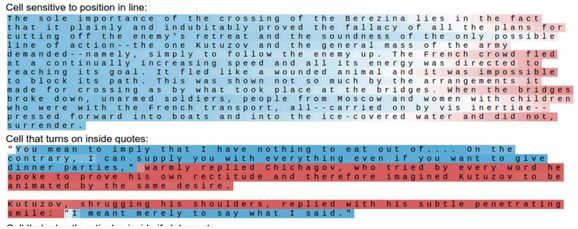
specific “circuits” that perform:

$$[A][B] \ldots [A] \rightarrow [B]$$

the “attention circuit” (KQ) finds [A], and then sends [B] (VO)

Polysemanticity/Privileged Basis

Features tracked by the partial computations are not necessarily aligned with the standard basis: single neuron not always fun to follow (contrast with e.g. Karpathy RNN demo)



Path/attribution patching

Similar to causal tracing, but with finer grained/more deliberate interventions

Mechanistic Interpretability

Induction Heads

Provided intuition: $[A][B] \ldots [A] \rightarrow [B]$

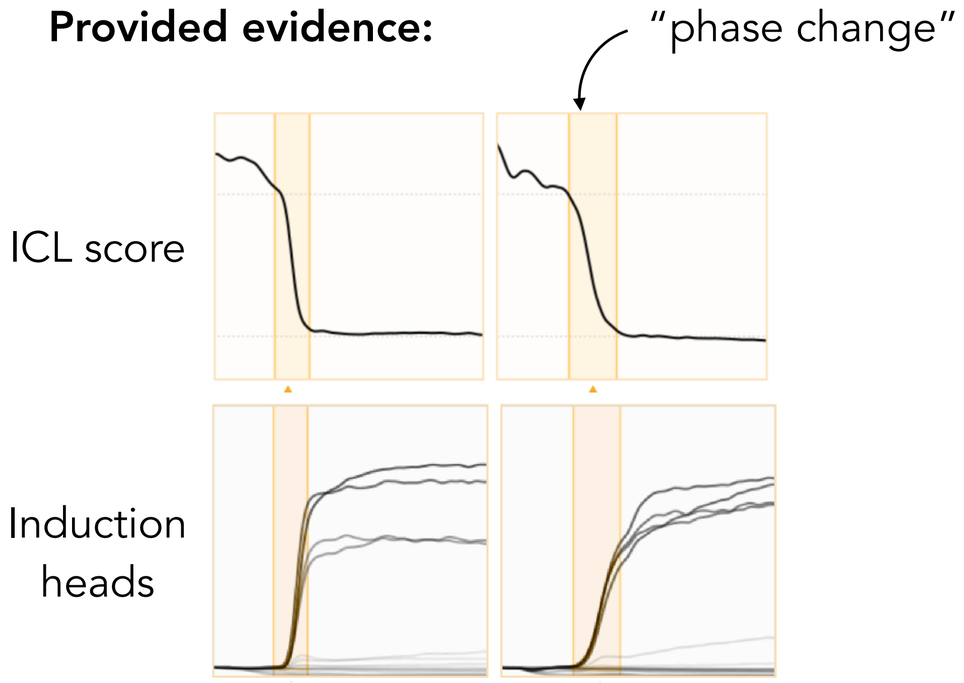
Provided definition: Heads for which, on input sample x, attention from position i tends to positions j for which $x_j = x_{i+1}$

Claim: Models that do in-context learning learn induction heads, and these heads are very helpful for ICL

Provided definition: In context learning score:

$$\frac{1}{N} \sum_{n \in [N]} L(x_{500} \,|\, x_{\leq 500}) - L(x_{50} \,|\, x_{\leq 50})$$

Provided evidence:



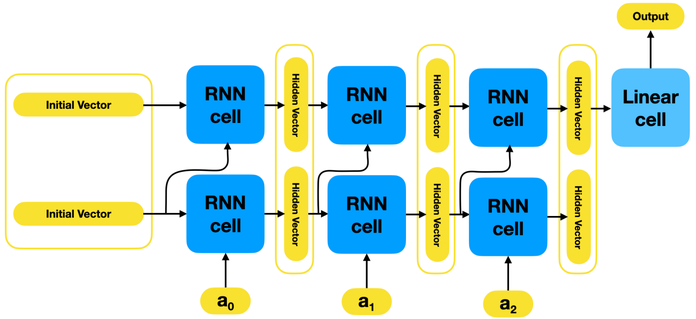
ablations too - manipulating time of IH appearance, removing IHs

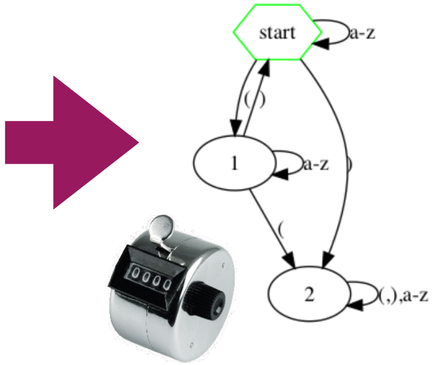
In-context learning and induction heads, Olsson et al, 2022

Recent investigation: What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation, Singh et al, 2024

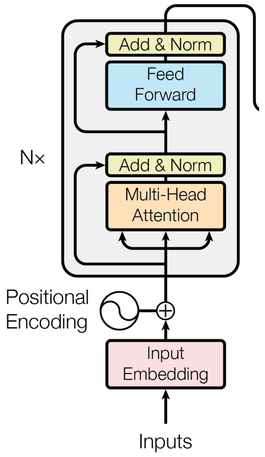
Can’t we just turn the networks into interpretable models?

Extraction

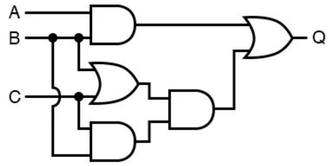




Extraction of rules from discrete-time recurrent neural networks,
Omlin and Giles, 1996
Extracting automata from recurrent neural networks using queries and counterexamples,
Weiss et al, 2018
Extracting context-free grammars from recurrent neural networks using tree-automata learning and A* search,
Barbot et al, 2021



?



RASP

$$\exists i\,[\mathsf{X}(\mathrm{a}, i) \wedge \mathsf{X}(\mathrm{b}, i+1)]$$

Proposes relevant model (RASP):
Thinking like transformers, Weiss et al, 2021

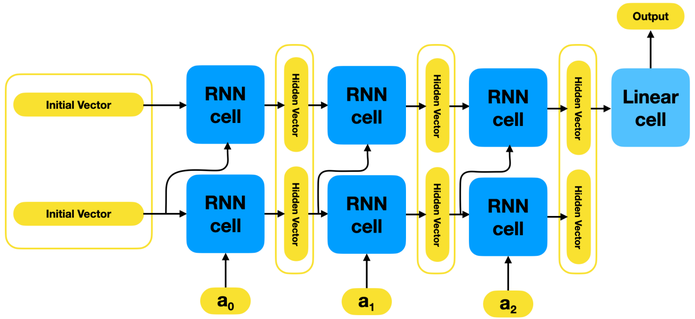
Proposes relevant playground (Tracr):
Compiled transformers as a laboratory for interpretability, Lindner et al, 2023

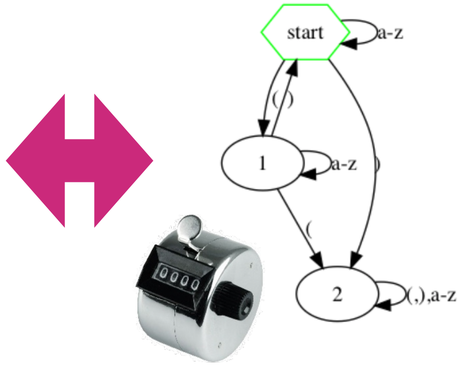
Proposes inherently interpretable models:
Learning transformer programs, Friedman et al, 2023

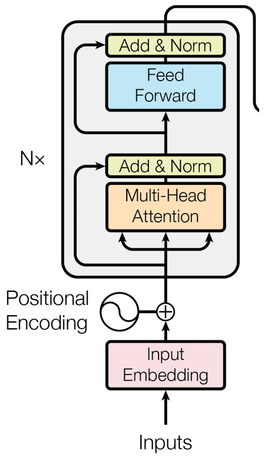
What about some theory

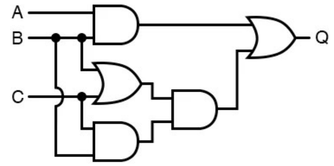
Formal Analysis/Expressivity

Understanding NN power with the help of formal tasks









RASP

$$\exists i\,[\mathsf{X}(\mathrm{a}, i) \wedge \mathsf{X}(\mathrm{b}, i+1)]$$

For example:

RNNs can learn regular languages!

Transformers struggle, but can learn shortcuts…

RNNs can't practically reverse or copy

Transformers can

Insights could inspire architecture design

A survey of neural networks and formal languages: Ackerman and Cybenko, 2020
Formal language theory meets modern NLP, Merrill, 2021
Formal aspects of language modeling, Cotterell et al, 2023
Transformers as recognizers of formal languages: a survey on Expressivity, Strobl et al, 2023

Conclusion

NNs have taken over, but **we don’t understand them**

A huge variety of methods to attempt explaining decisions has sprung up in response

We must be careful to not fool ourselves when working with them: **we must not draw conclusions from intuitions alone**

Good understanding facilitates **adversarial input** discovery, **model editing,** …

Nice resource (though doesn’t cover everything here):
Interpretable Machine Learning, Christoph Molnar, 2023
https://christophm.github.io/interpretable-ml-book/