Learning with symbols within neural networks:
Motivating examples

Neuro-symbolic modeling

THE UNIVERSITY OF UTAH

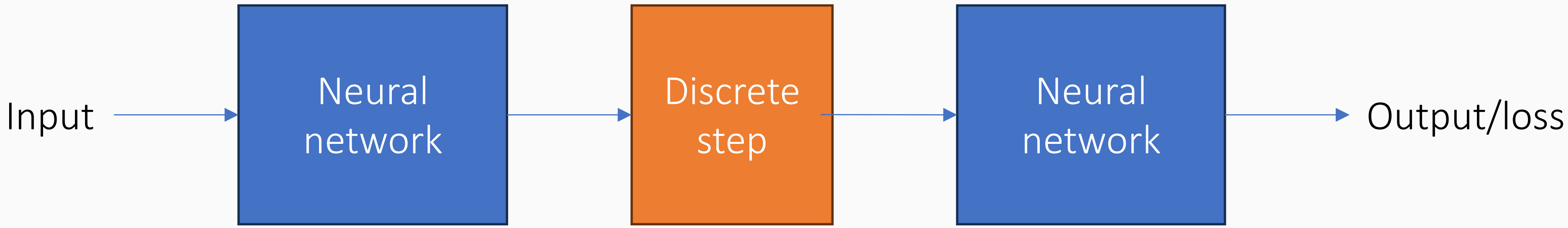
This lecture

- Motivating examples
- The straight-through estimator
- The Gumbel trick
- REINFORCE

(others if time permits)

Not all these approaches are always applicable

Neural networks containing discrete elements

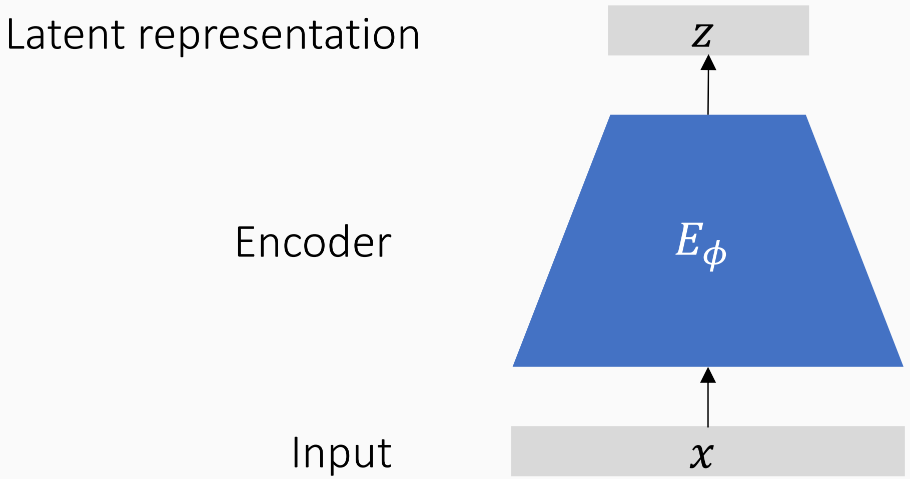


Let’s see some examples

Example 1: An autoencoder

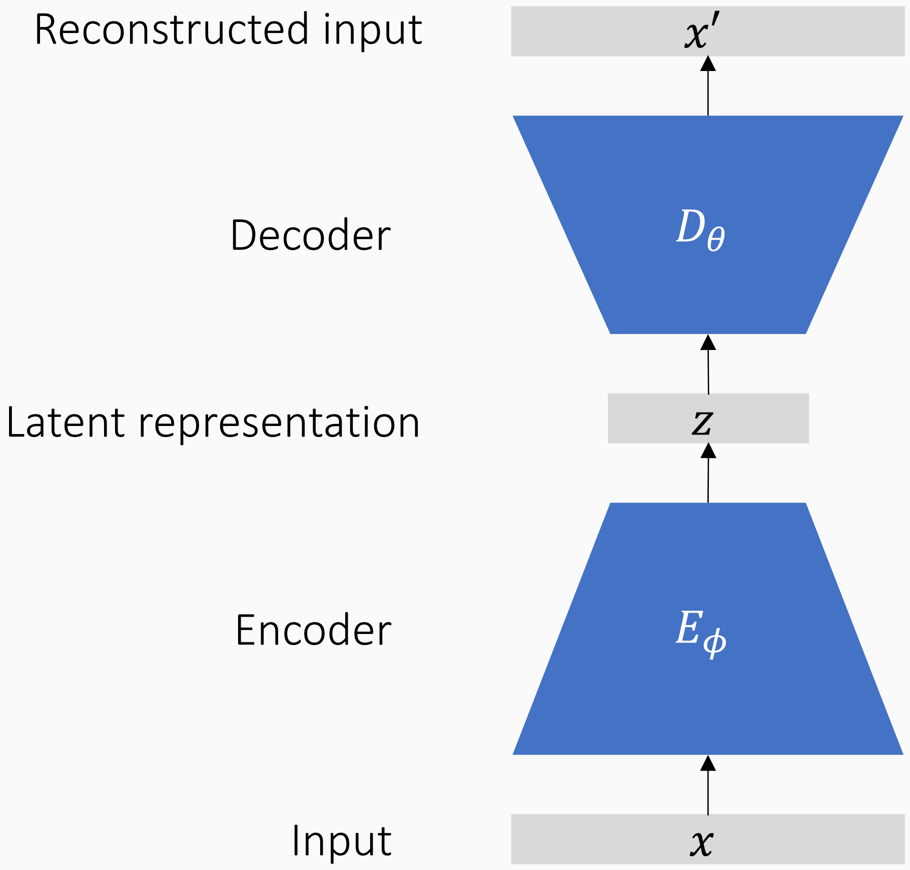
Input x

Example 1: An autoencoder



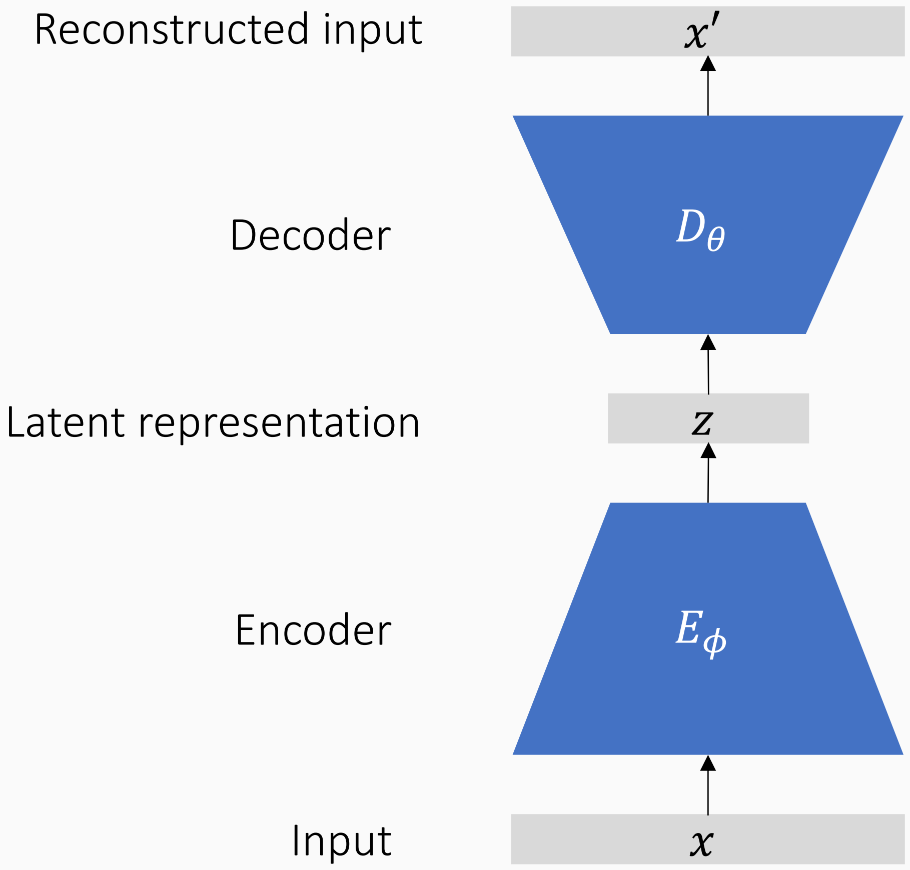
This could be any neural network architecture, parameterized by ϕ

Example 1: An autoencoder



This could be any neural network architecture, parameterized by θ

Example 1: An autoencoder

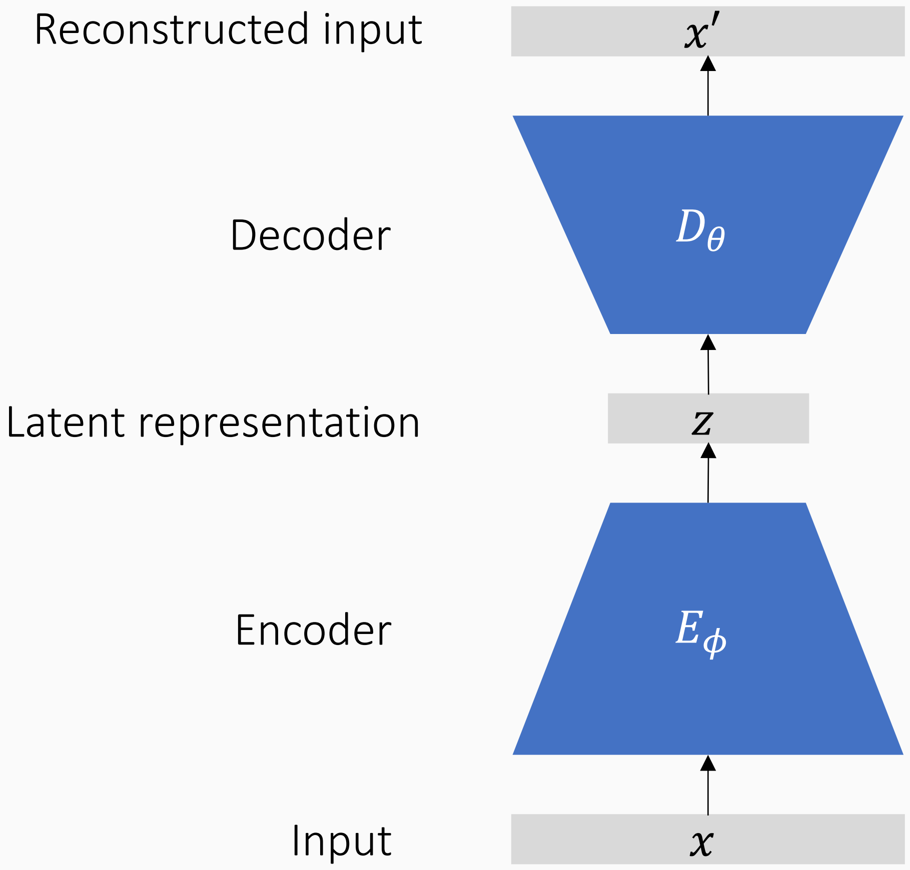


The goal of learning

To ensure that the reconstructed input x' is close to the original input x

If the encoder and decoder are differentiable and the latent representation is a real vector, we have an end-to-end differentiable function

Example 1: An autoencoder

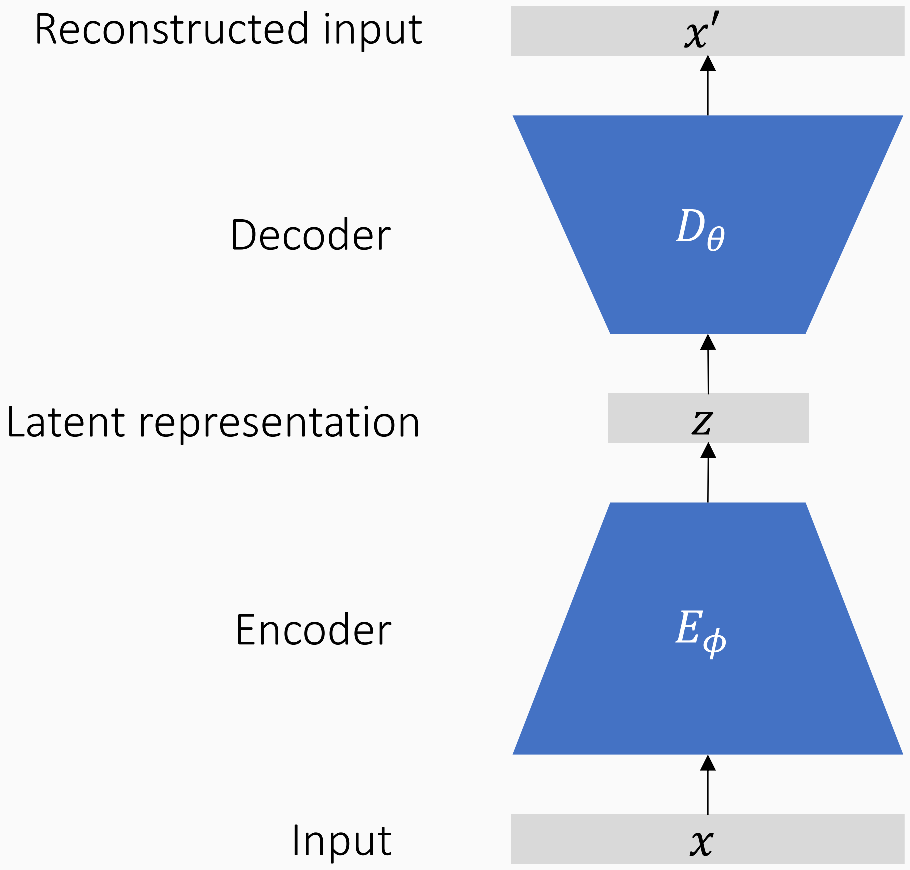


The goal of learning

To ensure that the reconstructed input x' is close to the original input x

The latent representation z can be seen as an *encoding* of the input that captures enough information to be able to reconstruct the input

Example 1: An autoencoder



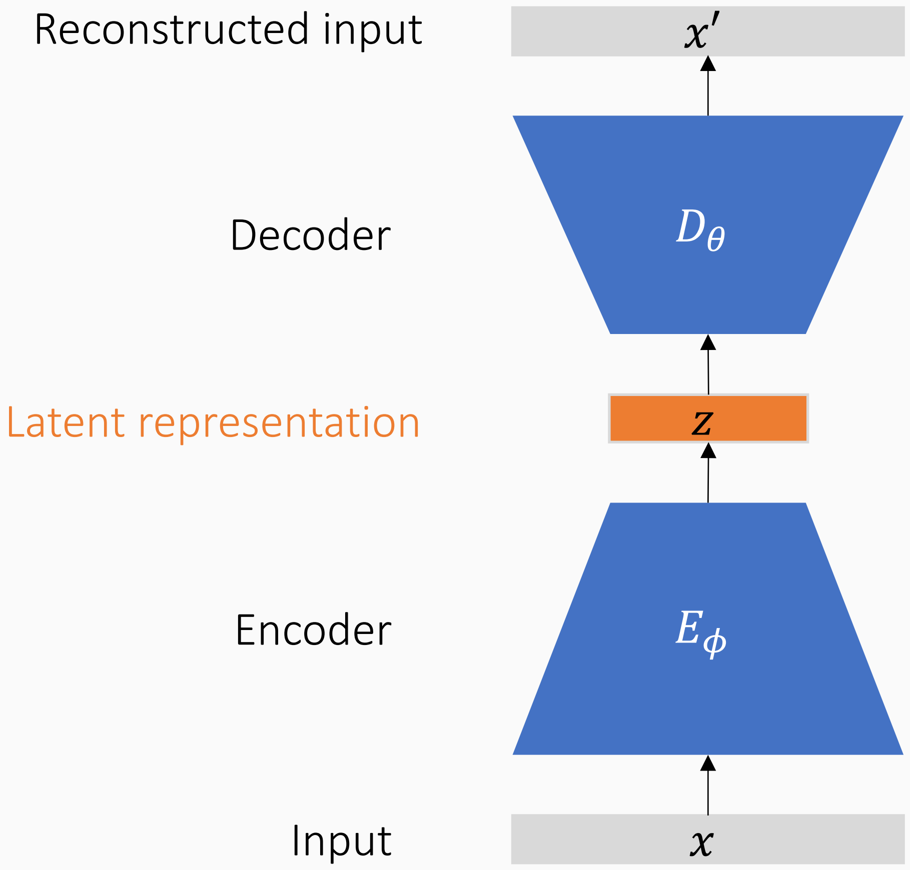
The goal of learning

To ensure that the reconstructed input x' is close to the original input x

The latent representation z can be seen as an *encoding* of the input that captures enough information to be able to reconstruct the input

Widely studied with several algorithms, models, etc

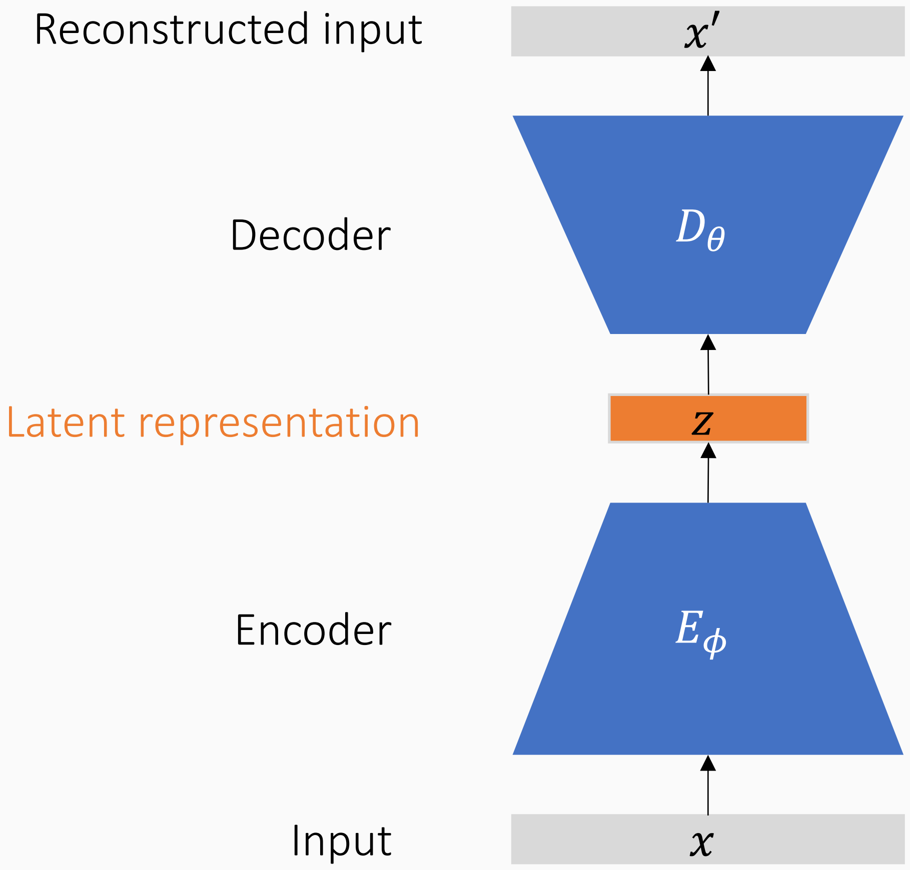
Example 1: An autoencoder



Suppose we want the latent representation to be discrete? E.g. a binary vector that is *sampled* from the sigmoids produced by the encoder

What might some advantages of such a discrete representation be?

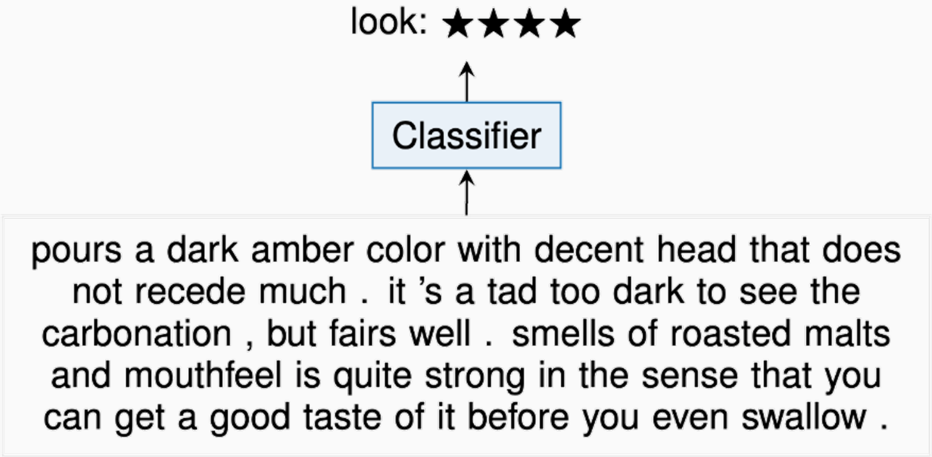
Example 1: An autoencoder



Suppose we want the latent representation to be discrete? E.g. a binary vector that is *sampled* from the sigmoids produced by the encoder

How does this affect learning?

Example 2: Text classification and rationales

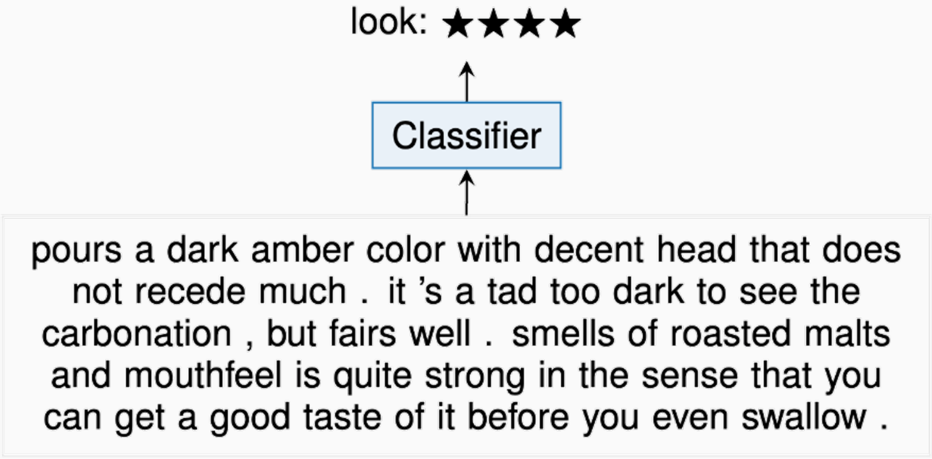


Suppose we have a text classification task

Given some text of a review, we need to assign it a rating for some aspect (here "look")

Example from Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963-2977. 2019.

Example 2: Text classification and rationales



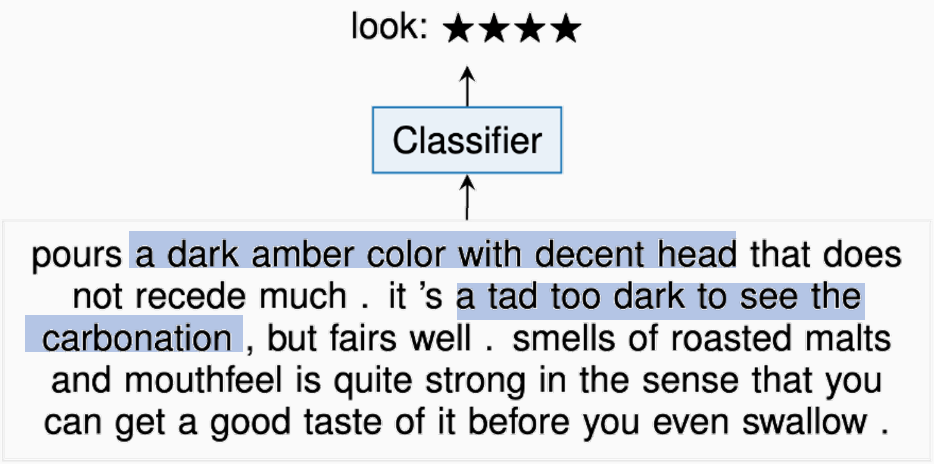
Suppose we have a text classification task

Given some text of a review, we need to assign it a rating for some aspect (here "look")

We can ask: Why did the classifier predict a four star rating for this example?

Example from Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963-2977. 2019.

Example 2: Text classification and rationales

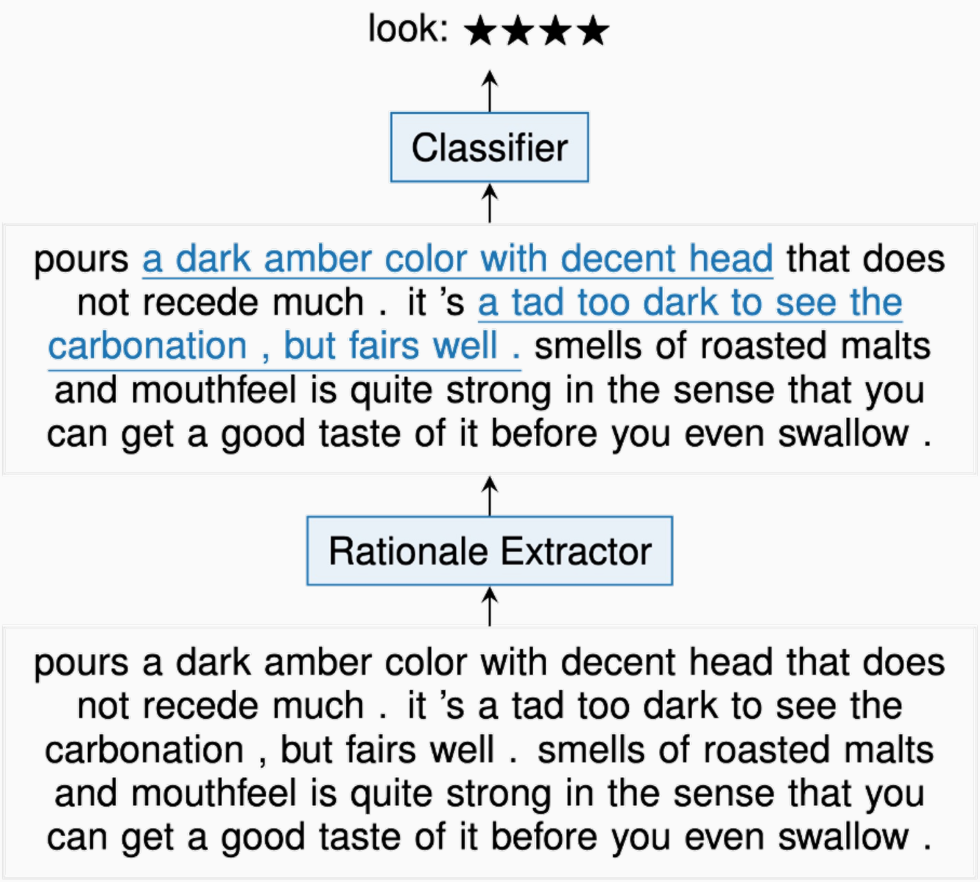


Why did the classifier predict a four star rating for this example?

One answer: Because of the highlighted words

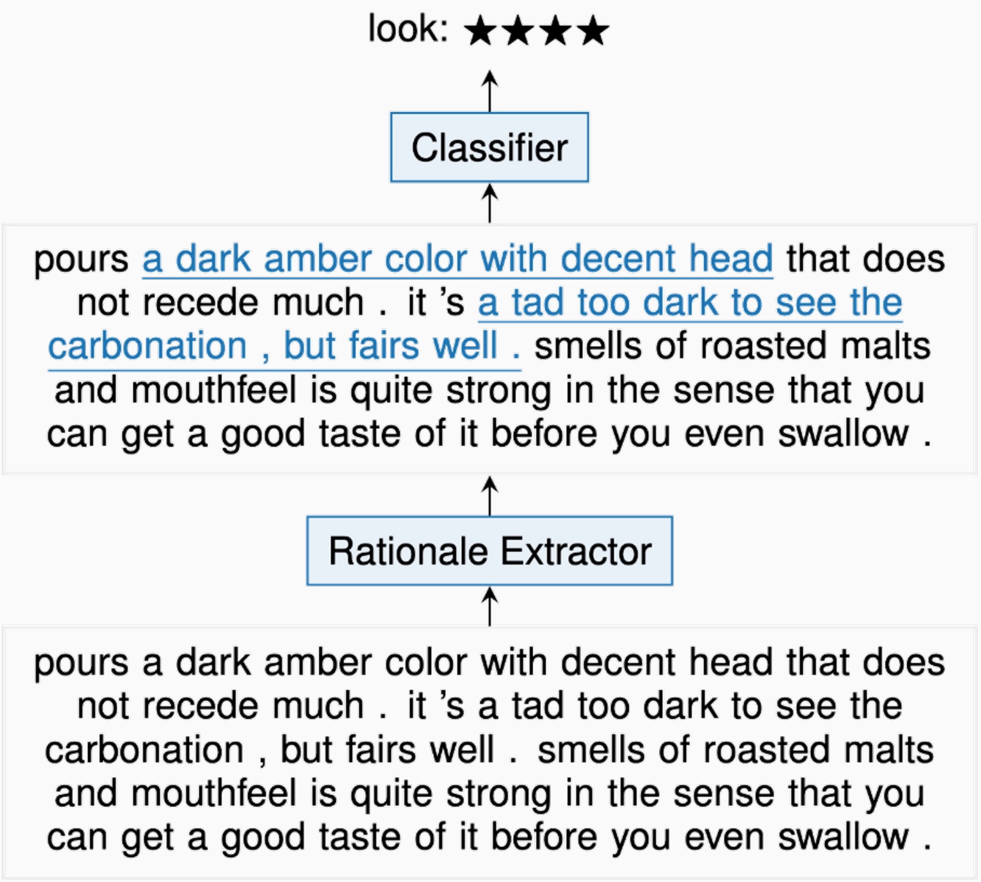
Example from Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963-2977. 2019.

Example 2: Text classification and rationales



Example from Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963-2977. 2019.

Example 2: Text classification and rationales



An alternative approach:

1. Identify the rationale (the highlighted words)

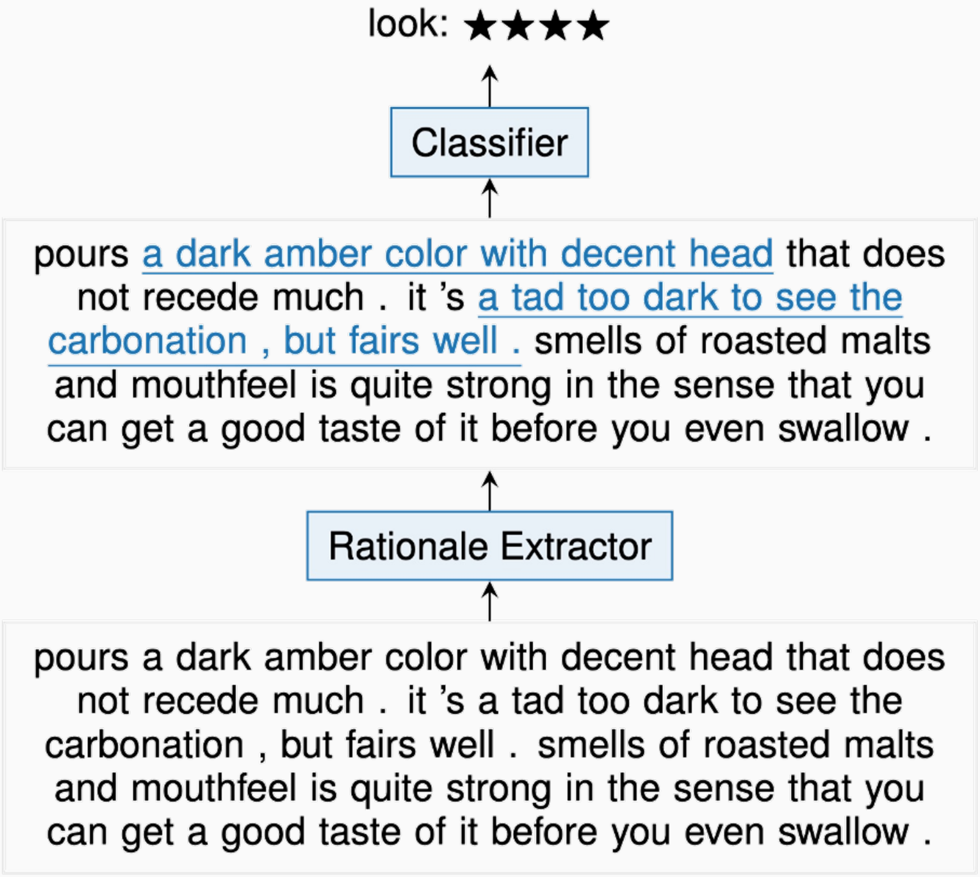
$$Z_i \mid x \sim Bernoulli\big(g_i(x, \phi)\big)$$

For each token, this represents whether the token is relevant or not

Each Z_i can be seen as a Boolean proposition

Example from Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963-2977. 2019.

Example 2: Text classification and rationales



An alternative approach:

1. Identify the rationale (the highlighted words)

$$Z_i \mid x \sim Bernoulli\big(g_i(x, \phi)\big)$$

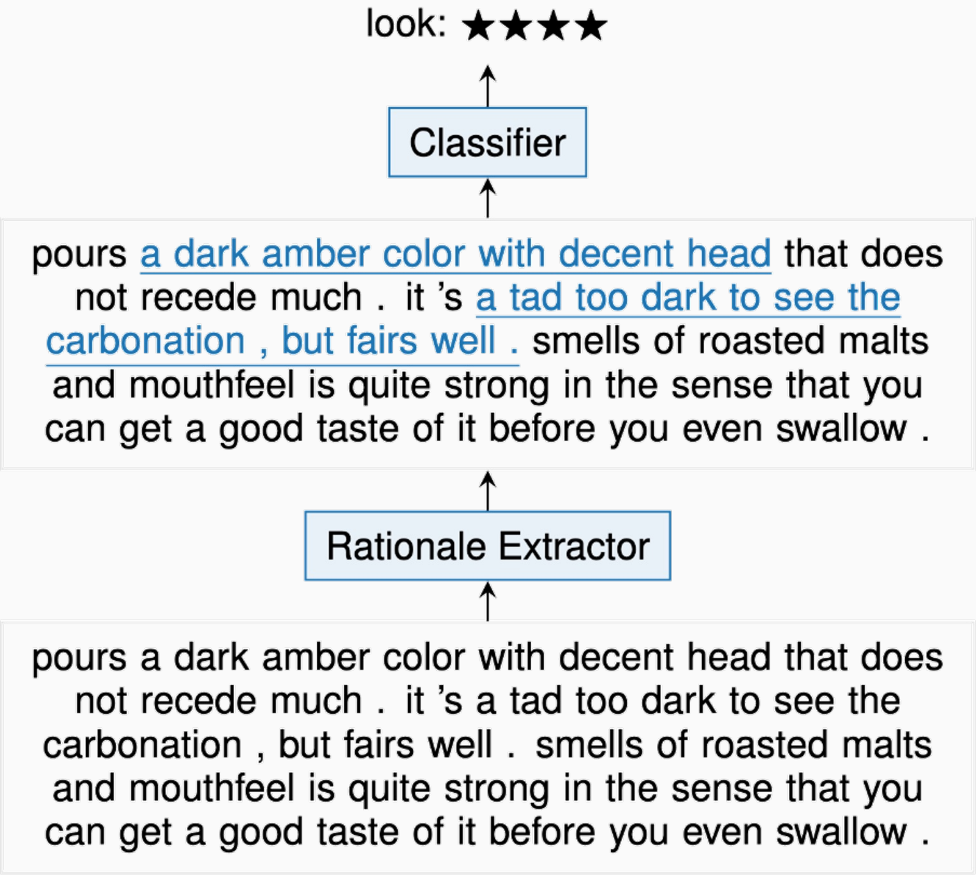
2. Use only the highlighted words as input to the classifier

$$Y \mid x \sim Categorical\big(f(x \odot Z, \theta)\big)$$

Mask out irrelevant words

Example from Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963-2977. 2019.

Example 2: Text classification and rationales



An alternative approach:

1. Identify the rationale (the highlighted words)

$$Z_i \mid x \sim Bernoulli\left(g_i(x, \phi)\right)$$

2. Use only the highlighted words as input to the classifier

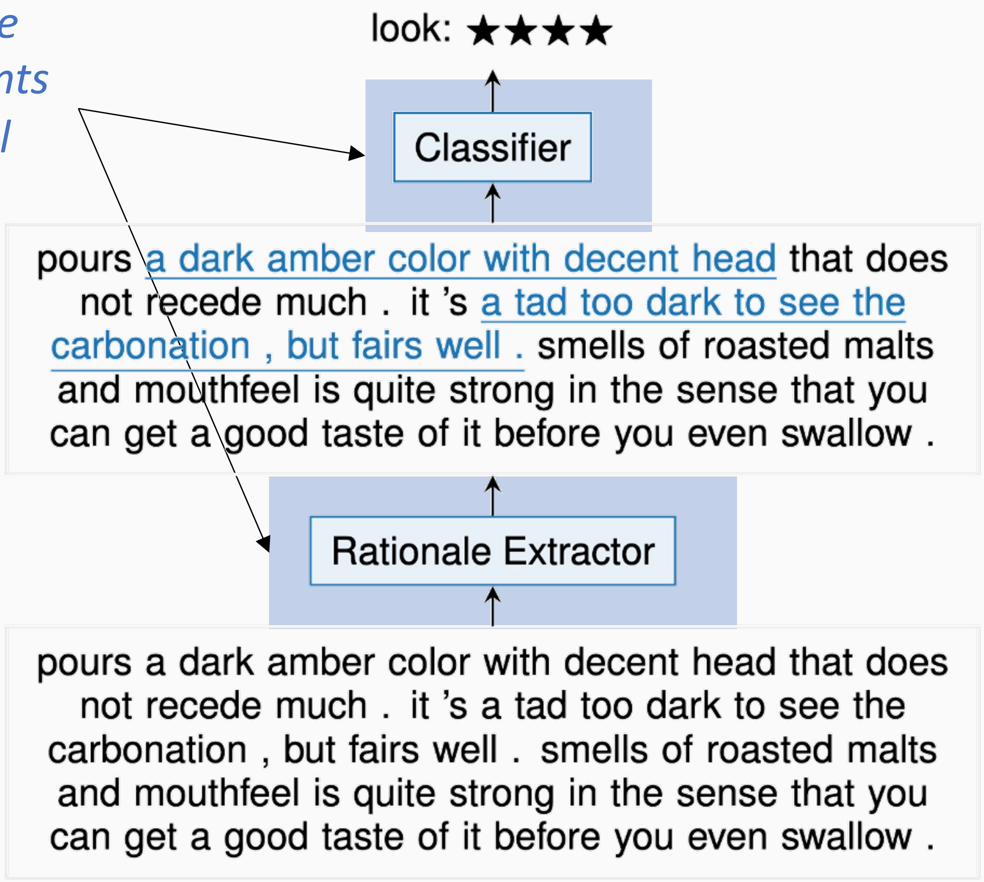
$$Y \mid x \sim Categorical\left(f(x \odot Z, \theta)\right)$$

Construct a categorical distribution over the remaining tokens

Example from Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963-2977. 2019.

Example 2: Text classification and rationales

Both these components are neural networks



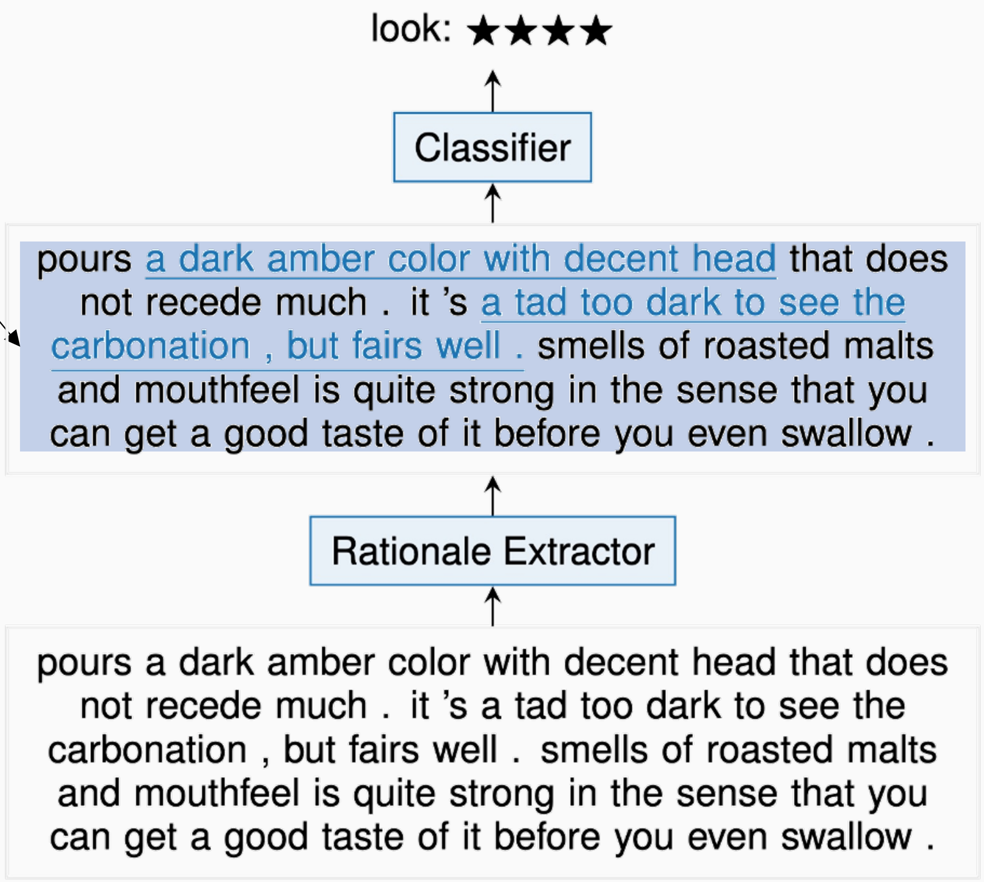
An alternative approach:

1. Identify the rationale (the highlighted words)
   $$Z_i \mid x \sim Bernoulli\big(g_i(x, \phi)\big)$$
2. Use only the highlighted words as input to the classifier
   $$Y \mid x \sim Categorical\big(f(x \odot Z, \theta)\big)$$

Example from Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963-2977. 2019.

Example 2: Text classification and rationales

But this step is discrete.



An alternative approach:

1. Identify the rationale (the highlighted words)

   $$Z_i \mid x \sim Bernoulli\big(g_i(x, \phi)\big)$$

2. Use only the highlighted words as input to the classifier

   $$Y \mid x \sim Categorical\big(f(x \odot Z, \theta)\big)$$

Example from Bastings, Jasmijn, Wilker Aziz, and Ivan Titov. "Interpretable Neural Predictions with Differentiable Binary Variables." In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2963-2977. 2019.

Other examples: Suppose we have …

…one neural component to permute a collection of inputs (e.g. sort it), and another network operating on the permutations

…one network identifying relevant parts of an image, and a different one using only the relevant components to make its prediction

…a black box program whose inputs are produced by a neural network, and whose outputs are consumed by another one

In all cases: How should we train the two networks?

Some approaches for addressing such problems

- The straight-through estimator
- The Gumbel trick
- Policy gradient and other RL approaches
- Differentiable structured layers

Not all these approaches are always applicable