Analysis of Neural Machine Translation

Lecture # 9

Hassan Sajjad and Fahim Dalvi
Qatar Computing Research Institute, HBKU

Analysis of Neural MT

- Neural MT achieves **state-of-the-art** performance for several language pairs - Almost every major tech company has switched over to this new paradigm

Analysis of Neural MT

- Neural MT achieves **state-of-the-art** performance for several language pairs - Almost every major tech company has switched over to this new paradigm
- However, we have seen that a lot of the details so far exist because *"they work"*
- Little research has gone into finding out what these models actually learn about source and target languages

Analysis of NMT

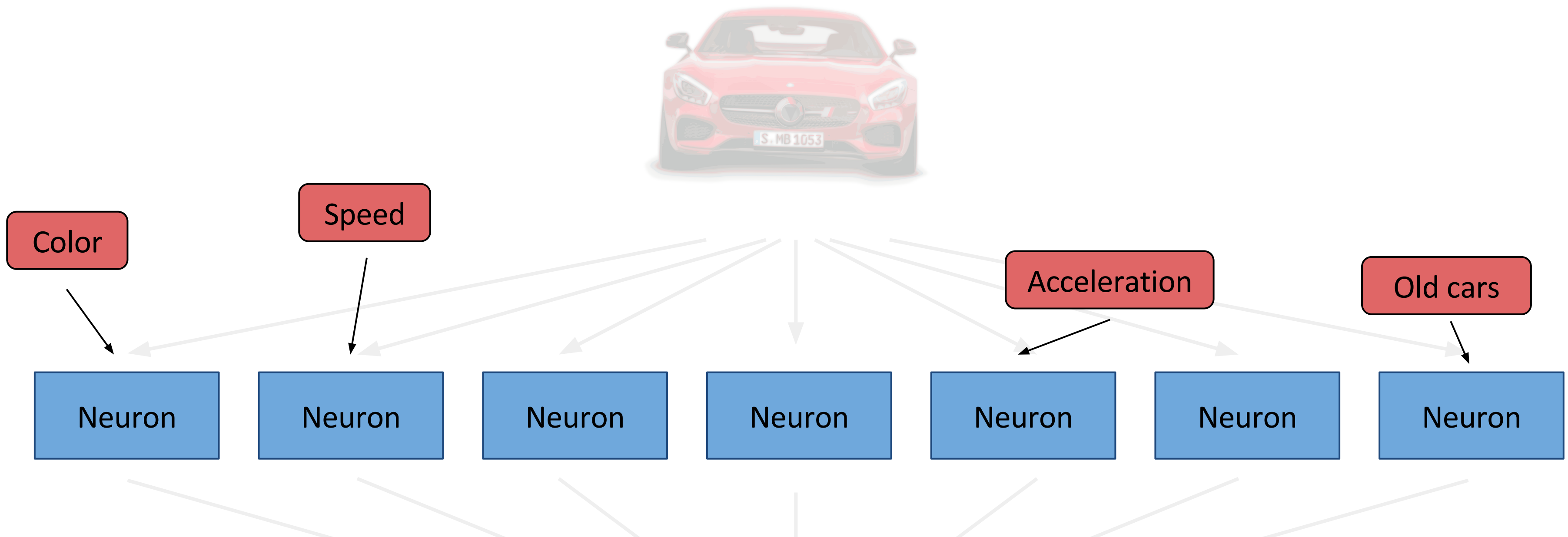
Activations

Analysis of Neural MT

Today, we will look at methods to probe and peek into these models - and see what they are actually learning!

Activation Analysis

Let’s start at the level of neurons:



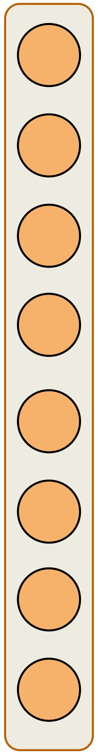
We said earlier that as a simplification, each **neuron** learns to look at a particular feature of the input (the *car* in this case)

Activation Analysis

Let's start at the level of neurons:

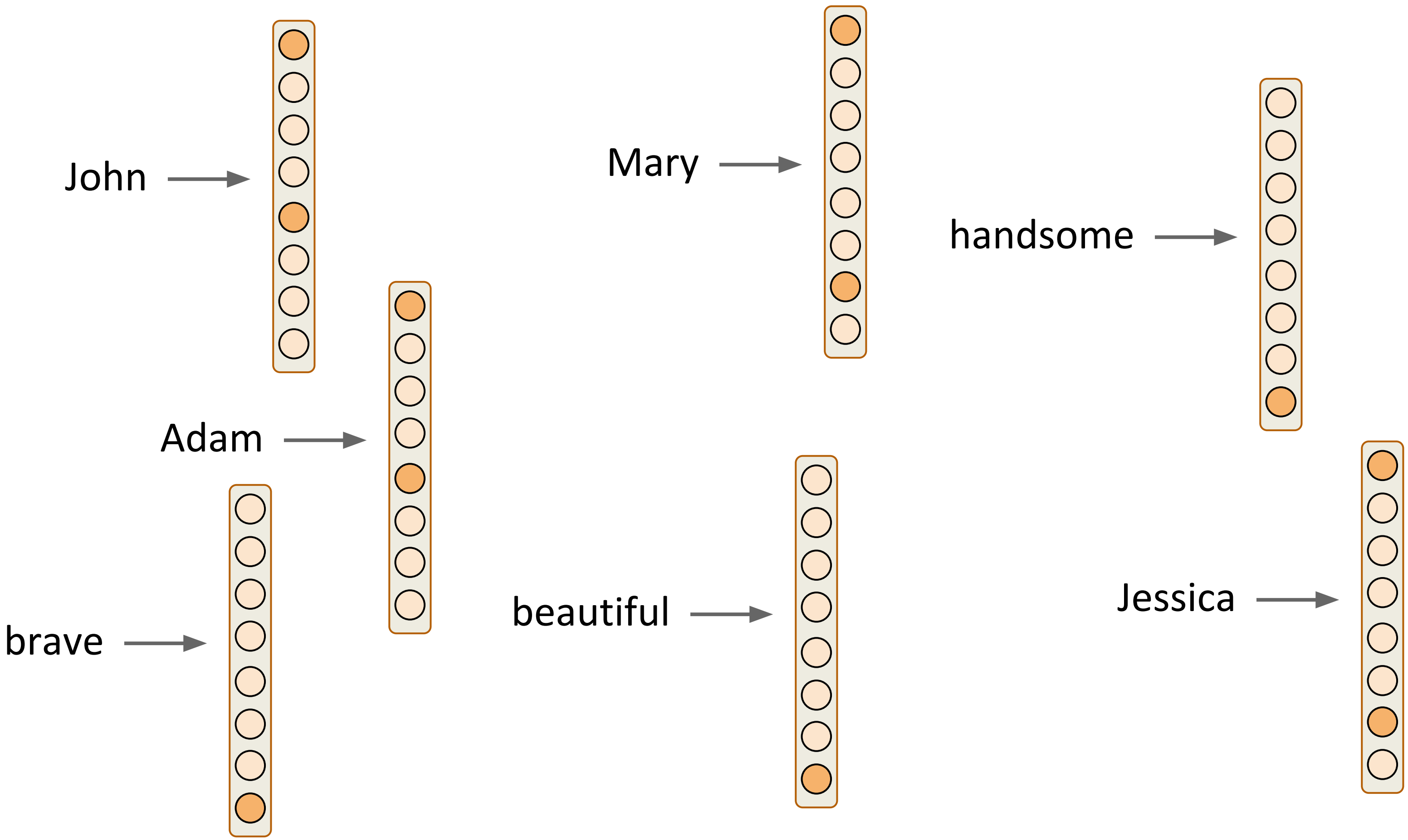
Idea: Given a trained model, if we can change only *one* feature, we can look at how the neurons react to figure out which neurons are responsible for that feature!

Activation Analysis



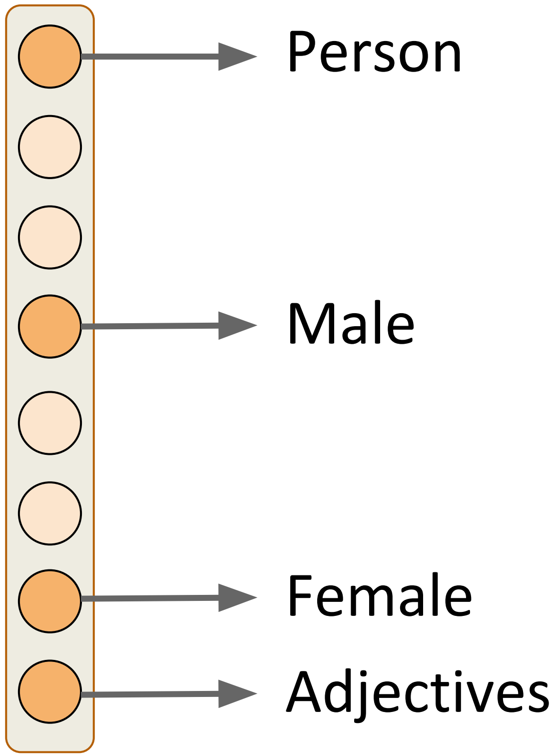
Consider a single layer from some trained neural network

Activation Analysis



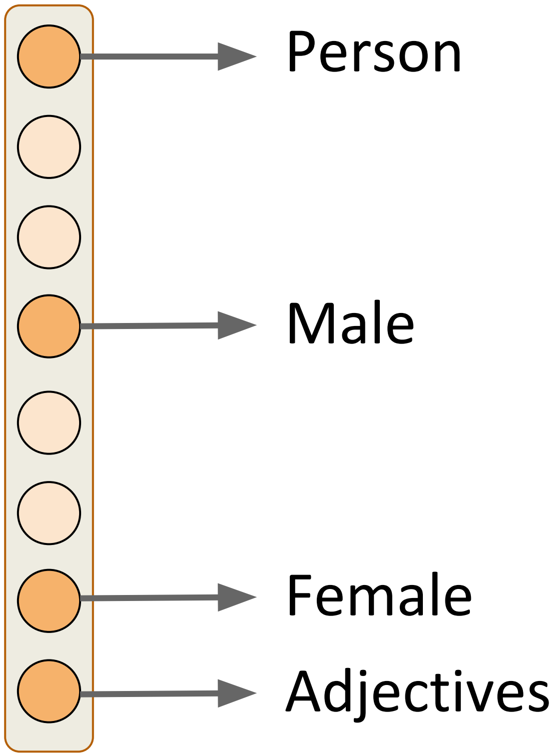
Can you find patterns in the above activations?

Activation Analysis



Some pattern finding can help us detect certain neurons that handle certain features!

Activation Analysis

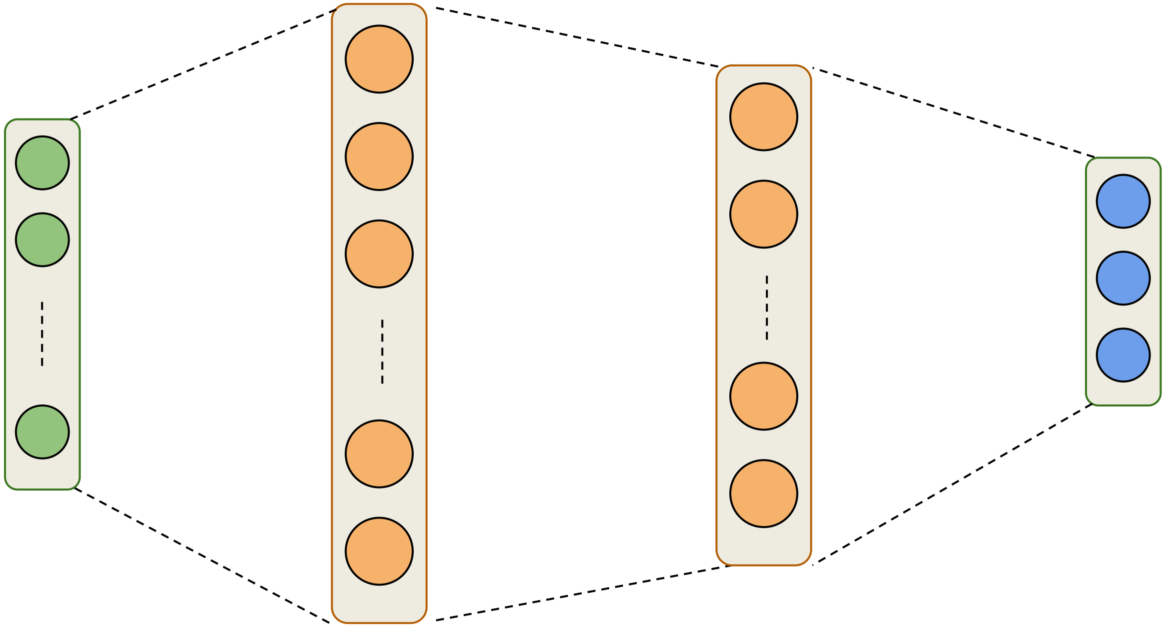


Also importantly, we can find out if certain features have no effect on the activations!

Analysis of NMT

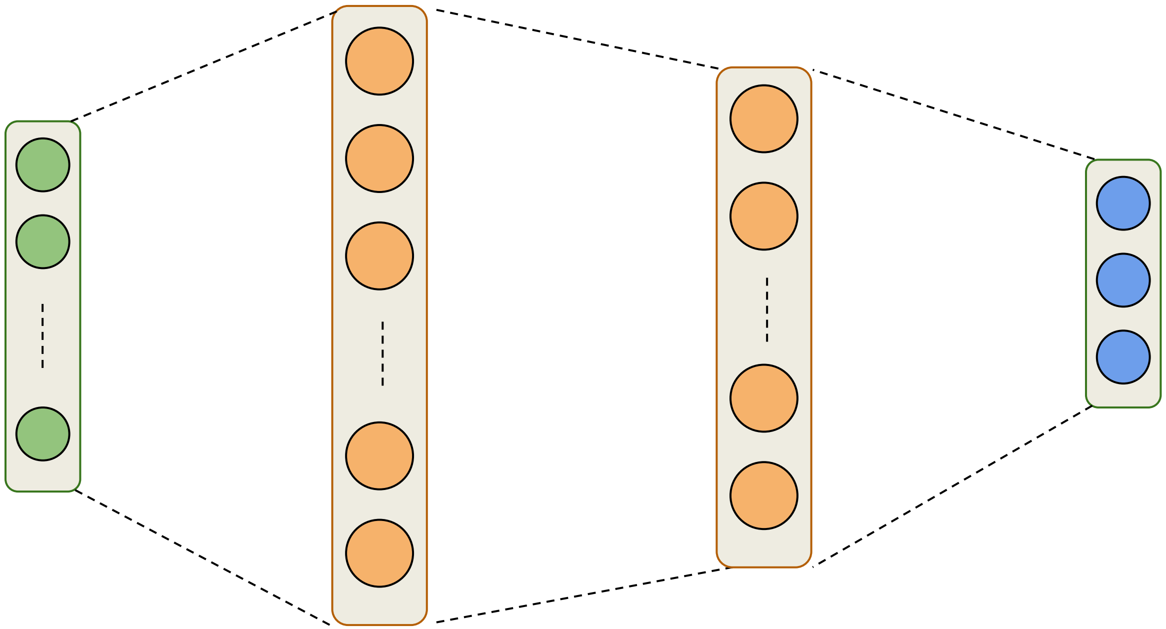
Saliency Detection using gradients

Saliency Detection



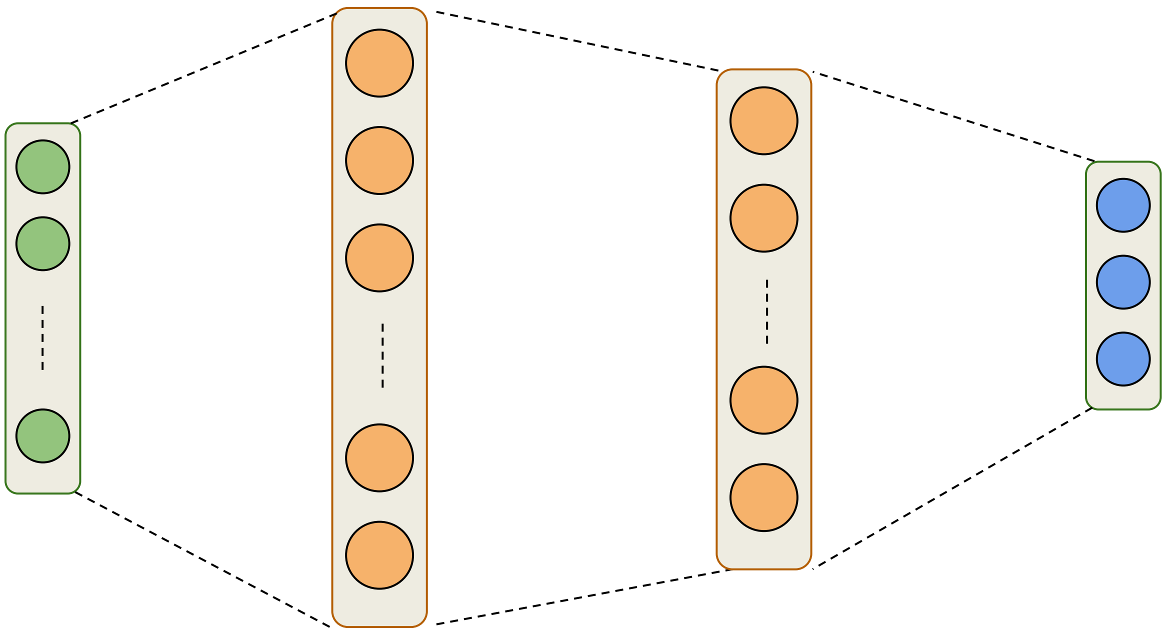
Usually, we backpropagate the gradients to the parameters to improve our model

Saliency Detection



Q: What if we backpropagate into the inputs themselves?

Saliency Detection

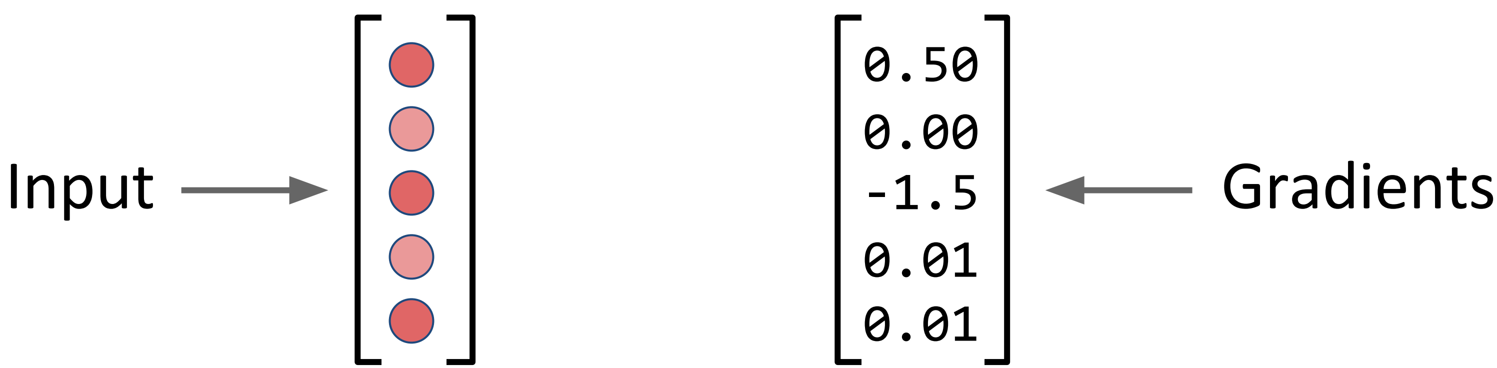


Q: What if we backpropagate into the inputs themselves?
A: The gradient will tell us about the importance of certain input features!

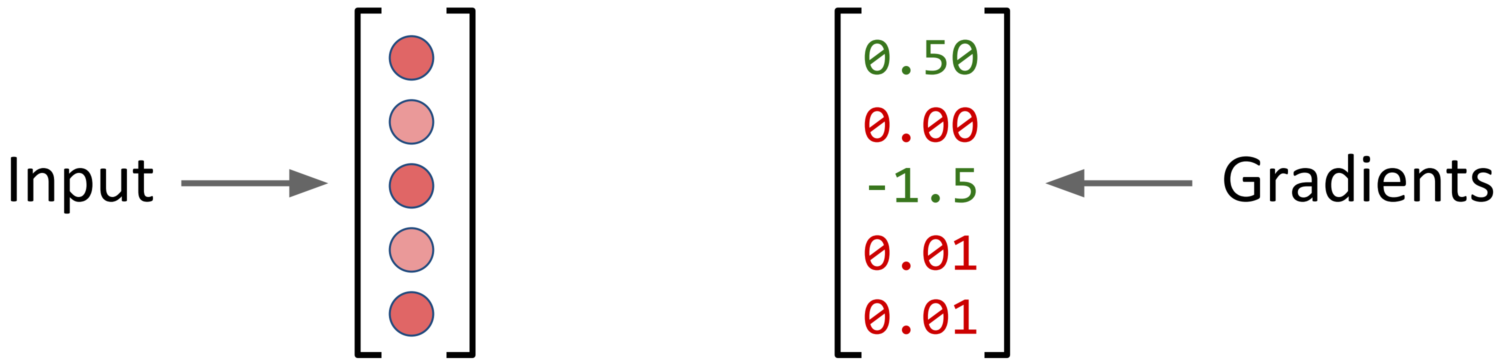
Saliency Detection

Consider that we have a trained network - and we now want to find out which features in our input are most important for prediction. Let our input consists of 5 features.

If we perform a forward pass using a single example, and the backpropagate the error all the way back to the example, we will have 5 gradients - one for each feature



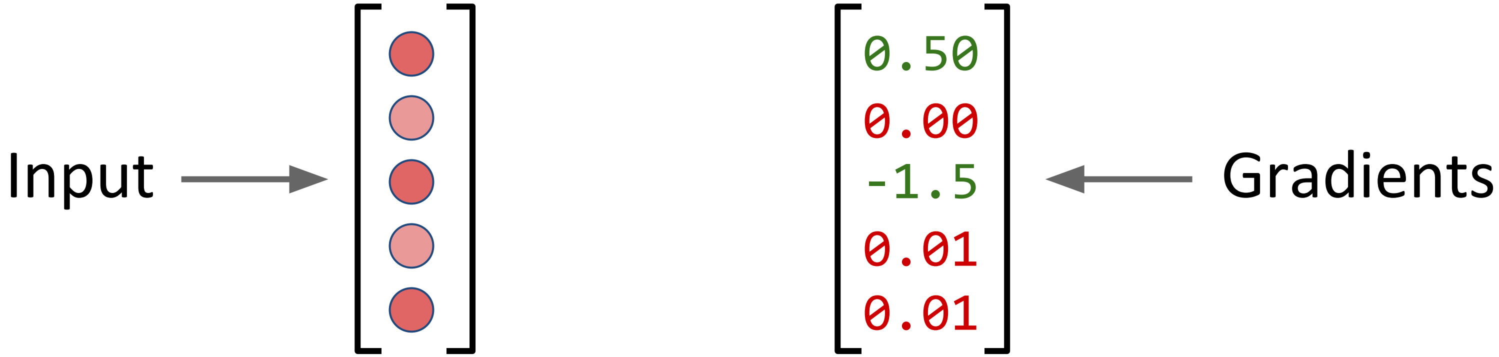
Saliency Detection



Consider the intuitive reasoning behind gradients:
A **high value** means changing the corresponding feature will **result in a change** in the loss.

A **low value** means changing the corresponding feature **will not result in any significant change** in the loss!

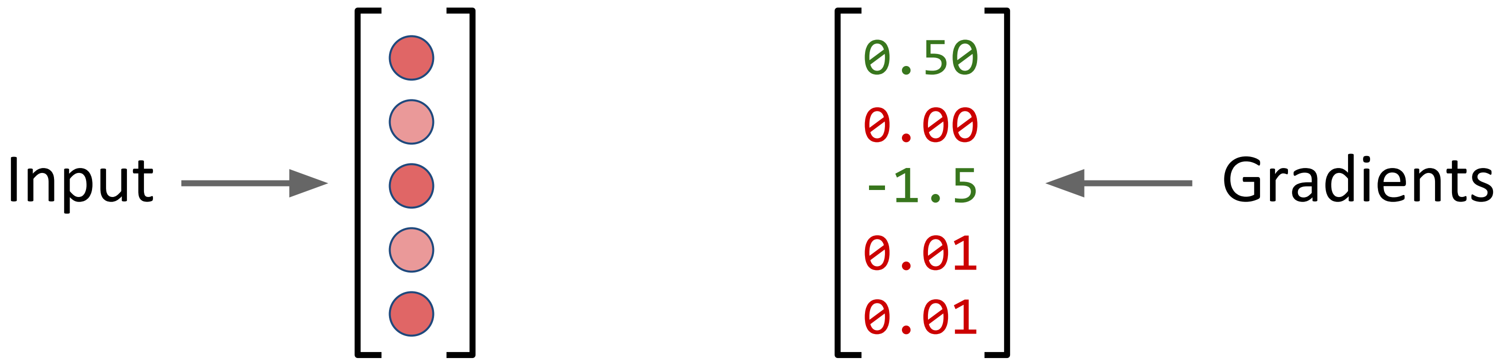
Saliency Detection



Let us perform this backpropagation for all examples, and take the average of the gradients for all examples.

If the average value is high for a feature, we can say it's important (or formally, **salient**)

Saliency Detection



Let us perform this backpropagation for all examples, and take the average of the gradients for all examples.

If the average value is low for a feature, we can say it's not very important

Saliency Detection

This technique is also generally called “saliency detection” in a lot of literature. There are some variations - but the basic idea remains the same.

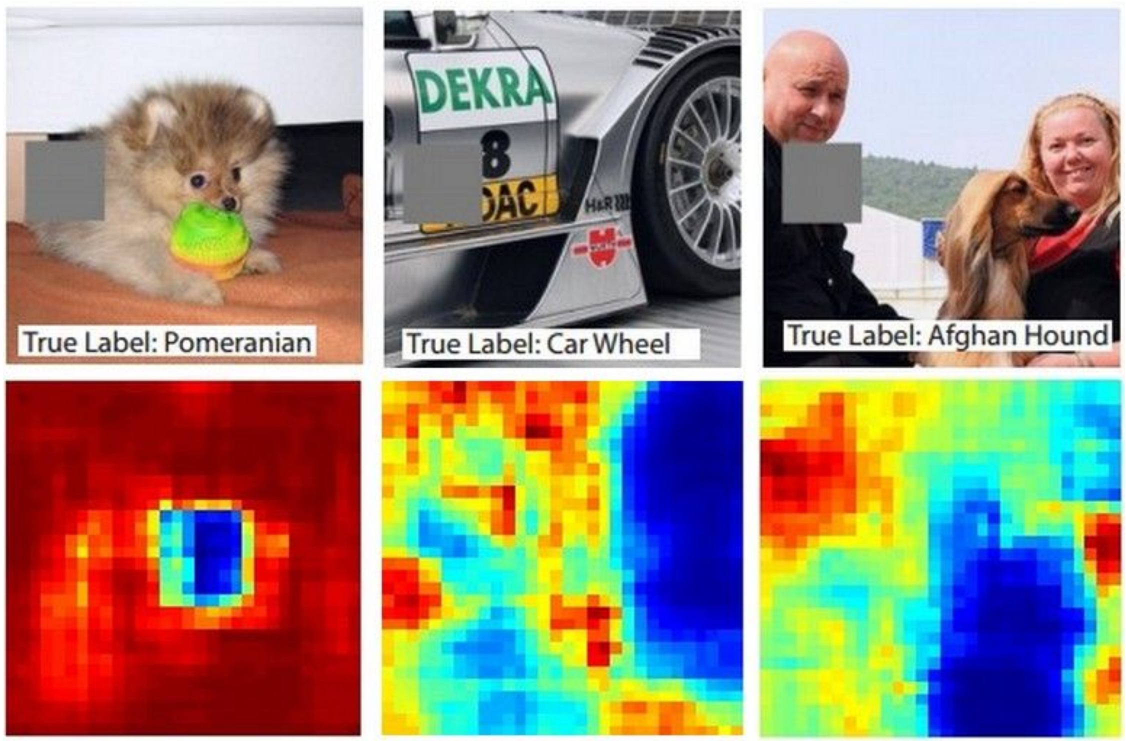
A slightly different but related technique is called “Layer-wise relevance propagation”, which uses the forward pass activation values to figure out which input features are important!

Analysis of NMT

Input Masking

Input Masking

Very simple technique that works well for images: hide part of the input image and compare the loss at the end



http://cs231n.github.io/understanding-cnn/

Analysis of NMT

Extrinsic Evaluation

Extrinsic Evaluation

Let's evaluate representations extrinsically!

- Shi et al. (2016) made the first attempt in analyzing the encoder's ability to learn source language syntax
- Belinkov et al. (2017) analyzed the encoder and decoder in learning morphology
- Belinkov et al. (2017b) analyzed the encoder in learning semantics of a language

Questions to Answer

- What do these models learn about **language phenomena**, such as morphology, semantic, and syntax?
- What is the **effect of word representations**, such as words and characters, on learning?
- What is the **role of encoder and decoder** in understanding the language?
- How does the **target language affect** the overall learning of the network?

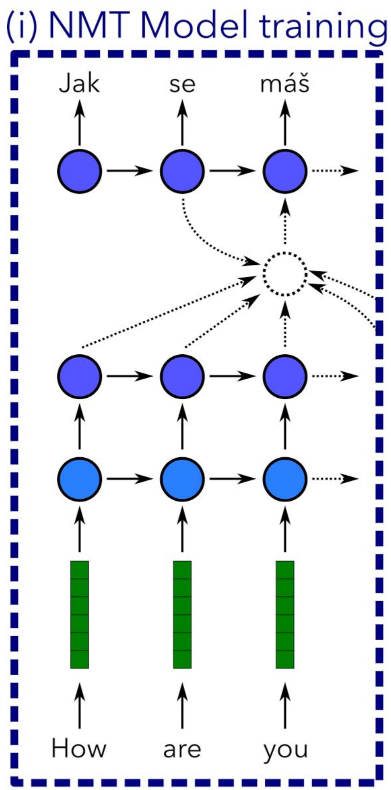
Methodology

Intuition

- Every word is represented as a dense vector in various layers of the network
- Do these dense vectors have information about linguistic properties of a word?
- Let’s take these dense vectors and evaluate their quality against language processing tasks, such as POS tagging, semantic tagging

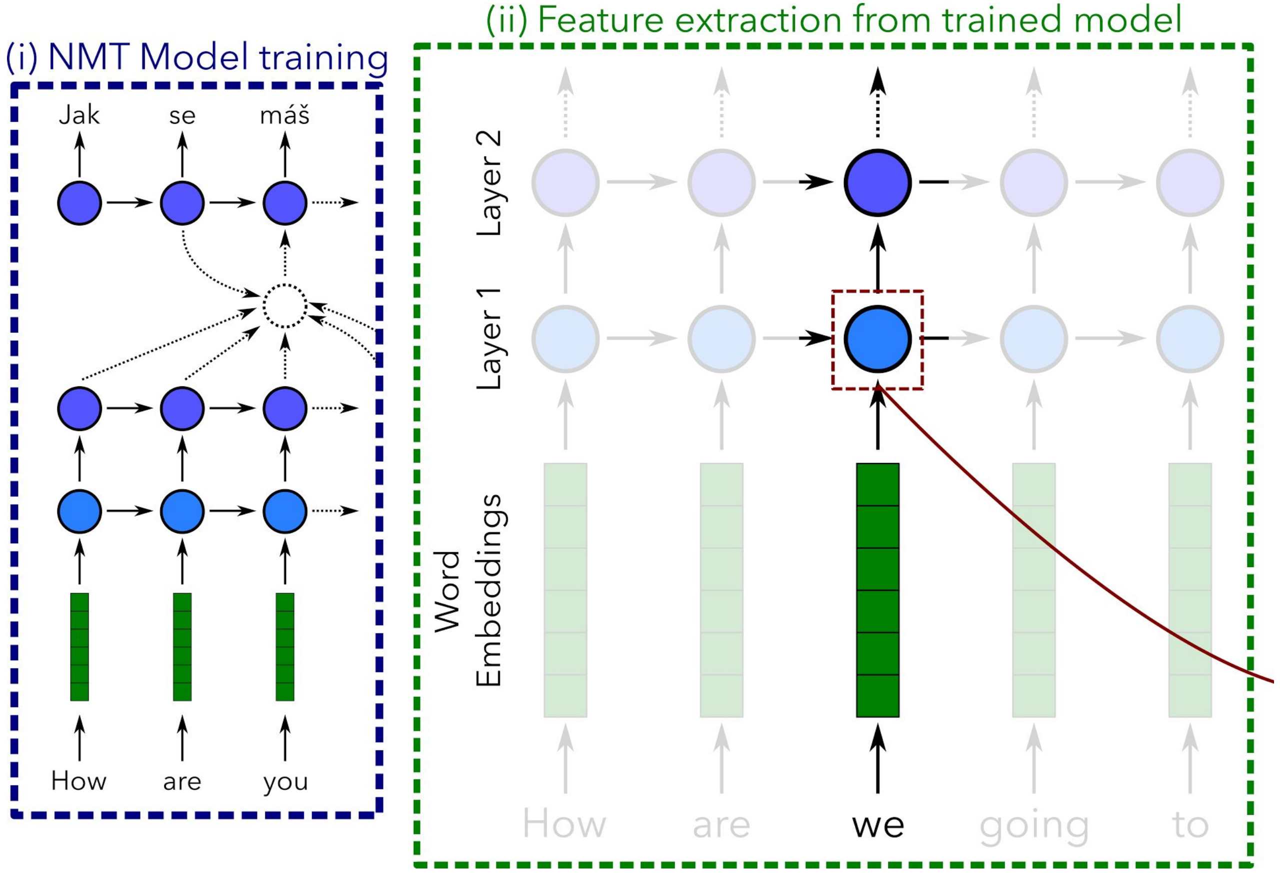
Methodology

1. Train an NMT system



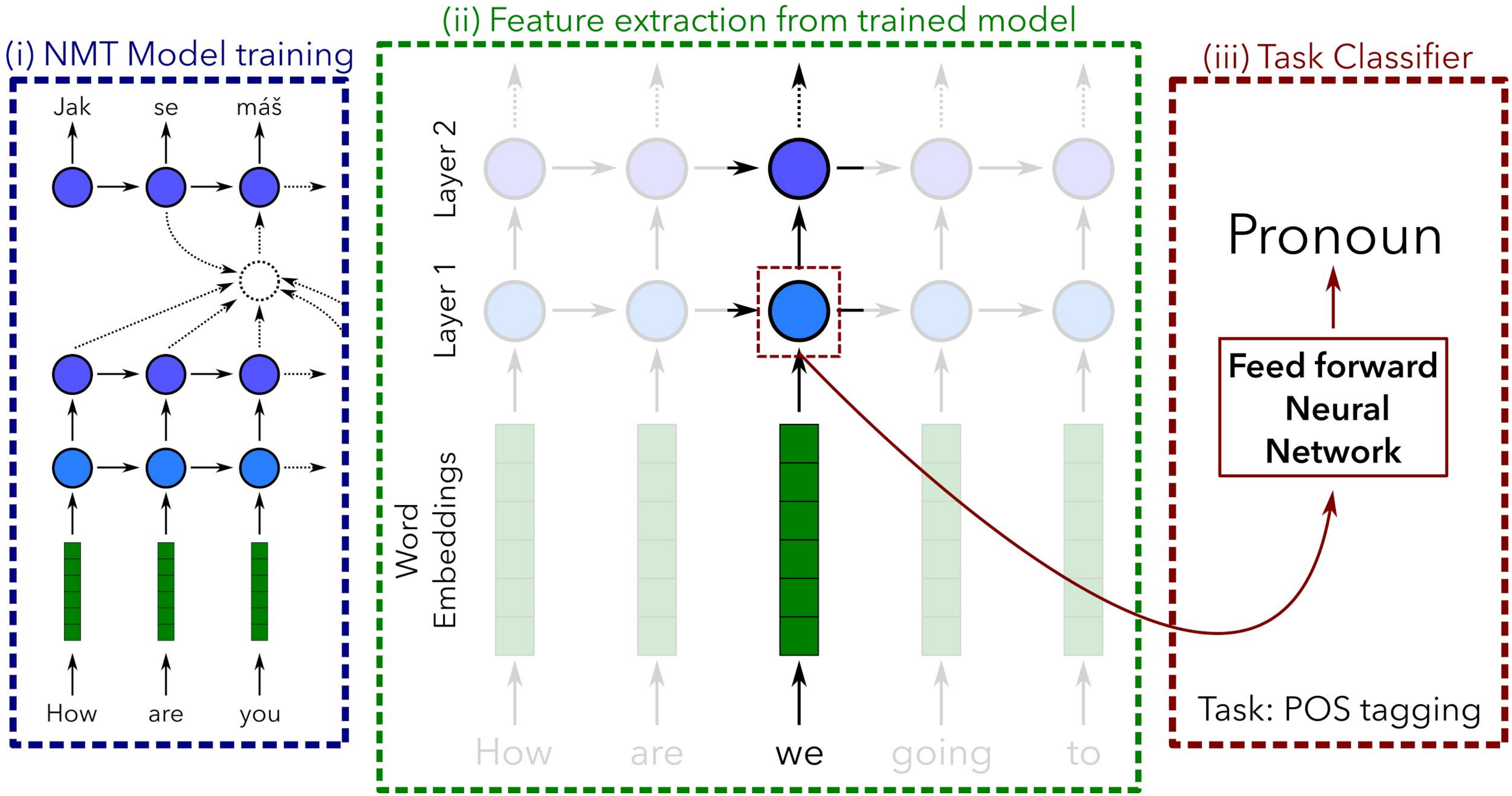
Methodology

2. Extract feature representations using the trained model



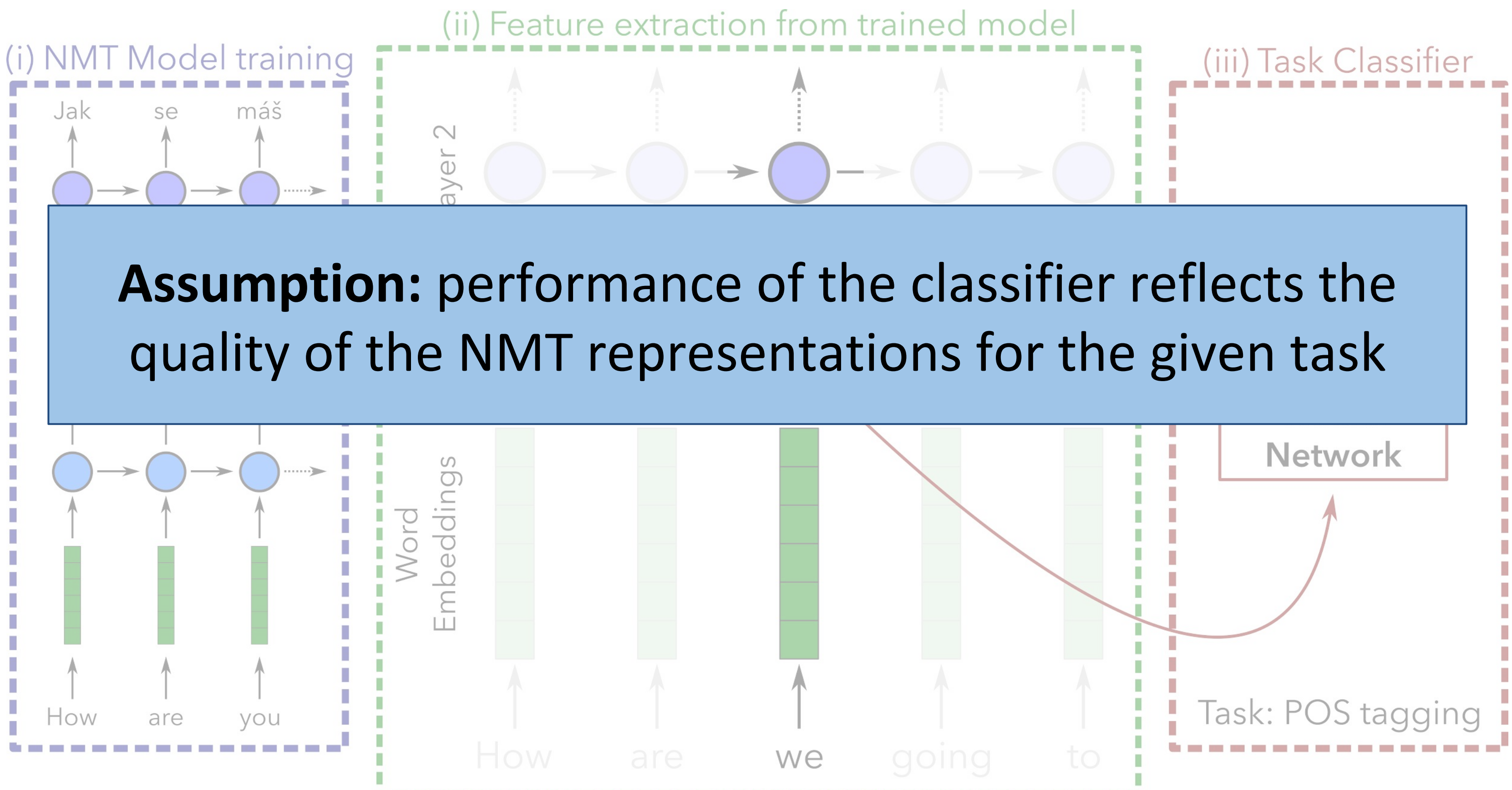
Methodology

3. Evaluate quality of features on an extrinsic task



Methodology

3. Evaluate quality of features on an extrinsic task



Assumption: performance of the classifier reflects the quality of the NMT representations for the given task

Methodology

Given an annotated corpus, say POS tagged:

- input every sentence of the corpus to the NMT trained model
- do a forward pass
- extract word representations corresponding to that word from a layer
- use it as a feature in an external classifier
- train the classifier
- predict test set
- evaluate using gold annotations

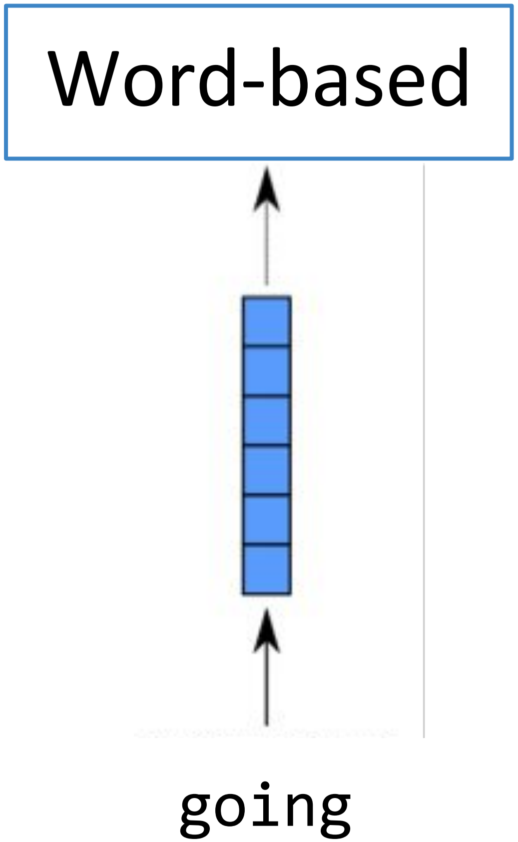
Experimental Setup

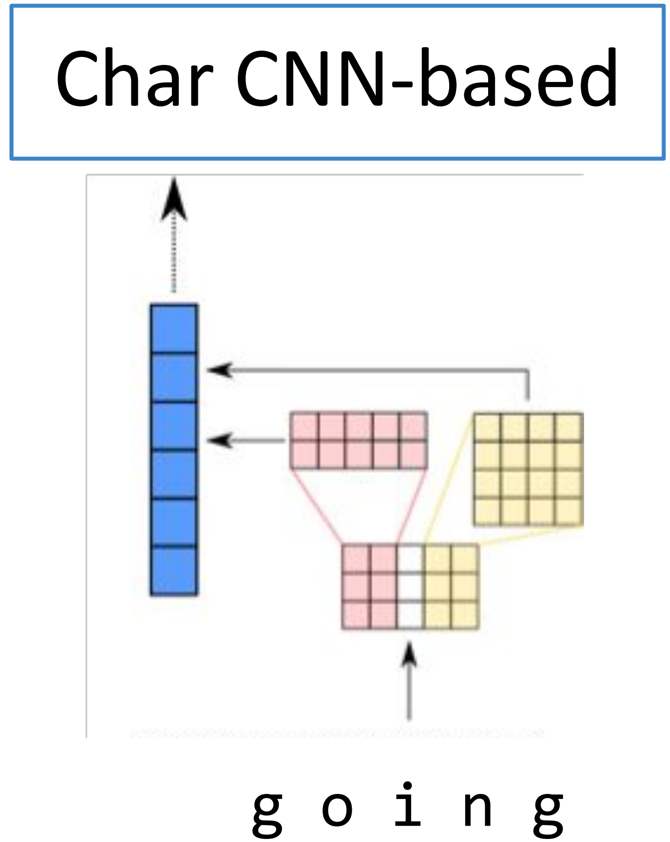
- Tasks
  - part of speech tagging (“runs” = verb)
  - morphological tagging (“runs”= verb, present tense, 3rd person singular)
- Languages
  - Arabic-,German-,French-,Czech-English

Effect of Word Representations

- Let’s start with the analysis of word representations learned on the encoder side

Effect of Word Representations





Effect of Word Representations

- Given a word-based and a char-CNN based NMT model
  - extract features of words in the POS-tagged corpus
  - train a classifier separately for word-based features and for char-CNN features
  - evaluate them against gold POS tags

Effect of Word Representations

- Overall, both representations are richer in capturing morphological information of words (accuracy around 90% in most of the cases)

Note: we are looking at source language morphology!

	Pred	BLEU
	Word/Char	Word/Char
Ar-En	89.62/95.35	24.7/28.4
Ar-He	88.33/94.66	9.9/10.7
De-En	93.54/94.63	29.6/30.4
Fr-En	94.61/95.55	37.8/38.8
Cz-En	75.71/79.10	23.2/25.4

Effect of Word Representations

- Character-based models are better in learning morphology of language (95.35 vs. 89.62 for Ar-En)

	Pred	BLEU
	Word/Char	Word/Char
Ar-En	89.62/95.35	24.7/28.4
Ar-He	88.33/94.66	9.9/10.7
De-En	93.54/94.63	29.6/30.4
Fr-En	94.61/95.55	37.8/38.8
Cz-En	75.71/79.10	23.2/25.4

Effect of Word Representations

- Difference of word vs. char based accuracies increases for morphologically rich languages (see **Ar** and **Cz** results)

	Pred	BLEU
	Word/Char	Word/Char
Ar-En	89.62/95.35	24.7/28.4
Ar-He	88.33/94.66	9.9/10.7
De-En	93.54/94.63	29.6/30.4
Fr-En	94.61/95.55	37.8/38.8
Cz-En	75.71/79.10	23.2/25.4

Effect of Word Representations

- POS tagging accuracy is independent of the quality of machine translation system (see Ar-He has very low BLEU score but still achieve good POS accuracy)

	Pred	BLEU
	Word/Char	Word/Char
Ar-En	89.62/95.35	24.7/28.4
Ar-He	88.33/94.66	9.9/10.7
De-En	93.54/94.63	29.6/30.4
Fr-En	94.61/95.55	37.8/38.8
Cz-En	75.71/79.10	23.2/25.4

Effect of Encoder Depth

What kind of information is stored at different layers of the model?

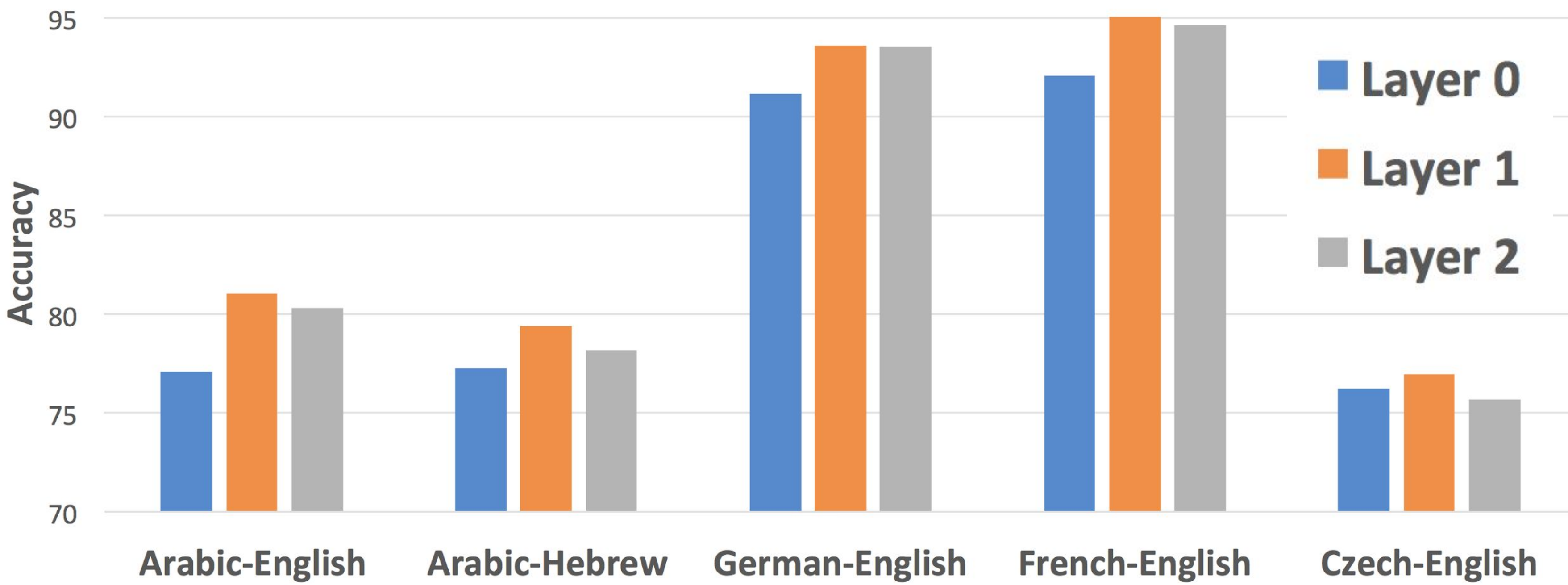
Effect of Encoder Depth

- Neural models can be very deep
  - Google translate: 8 encoder/decoder layers
  - Zhou et al. 2016: 16 layers
- What kind of language information is learned at each layer?
- Let’s analyze a 2-layer encoder
  - extract representations from different layers for training the classifier

Effect of Encoder Depth

$$Layer1 > Layer2 > Layer0$$

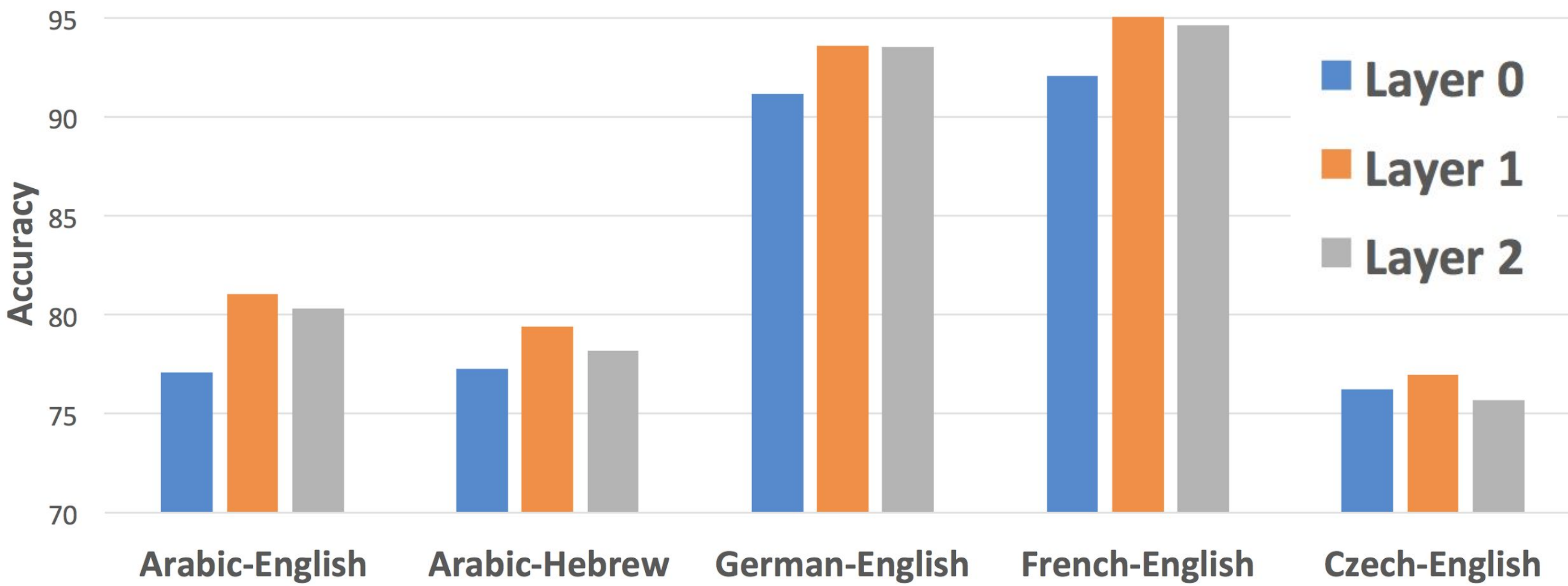
But, deeper models translate better!



Effect of Encoder Depth

$$Layer1 > Layer2 > Layer0$$

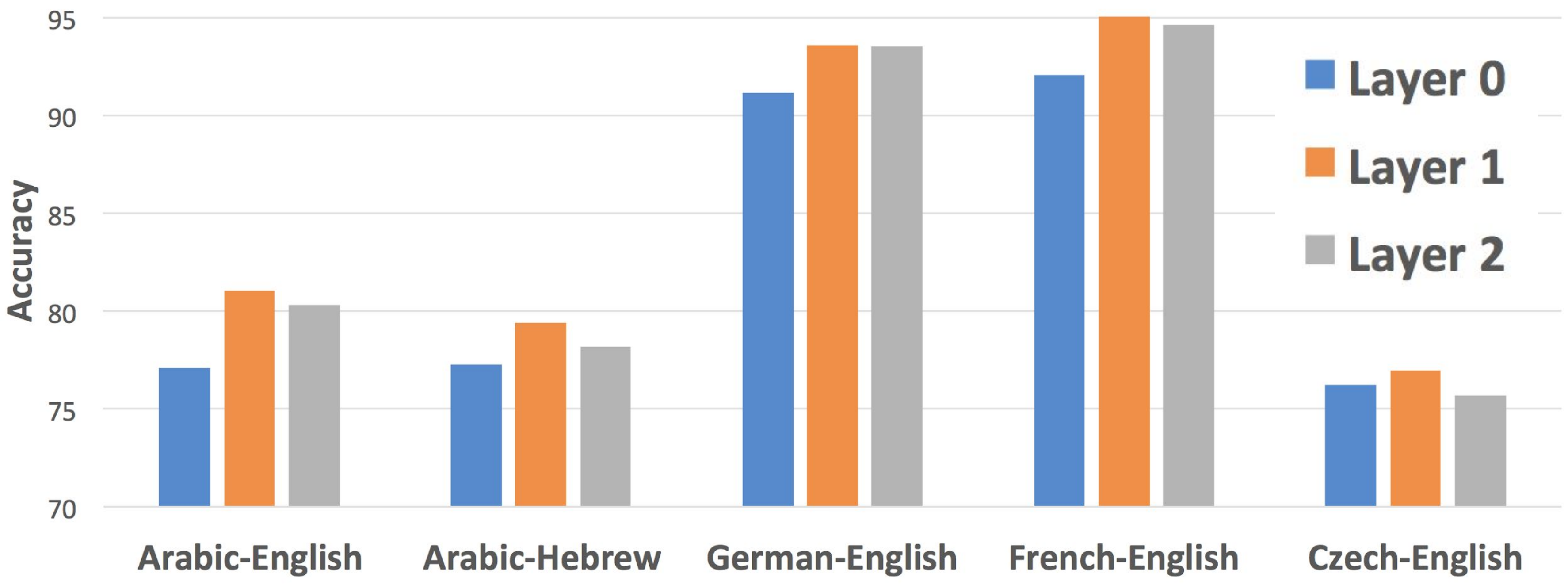
What do higher layers learn better?



Effect of Encoder Depth

$$Layer1 > Layer2 > Layer0$$

What do higher layers learn better?



Effect of Target Language

Does the target language affect source-side representations?

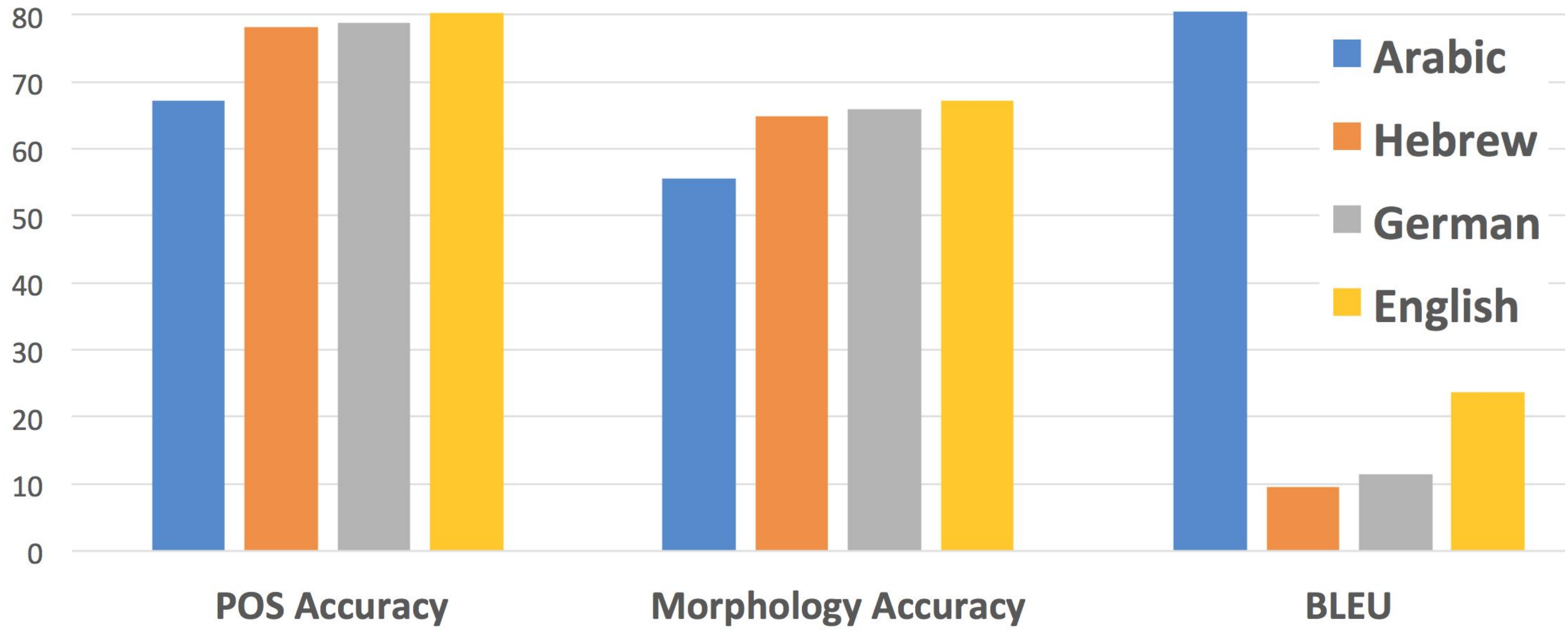
Effect of Target Language

Does the target language affect source-side representations?

- Fix source language and train NMT models on different target languages

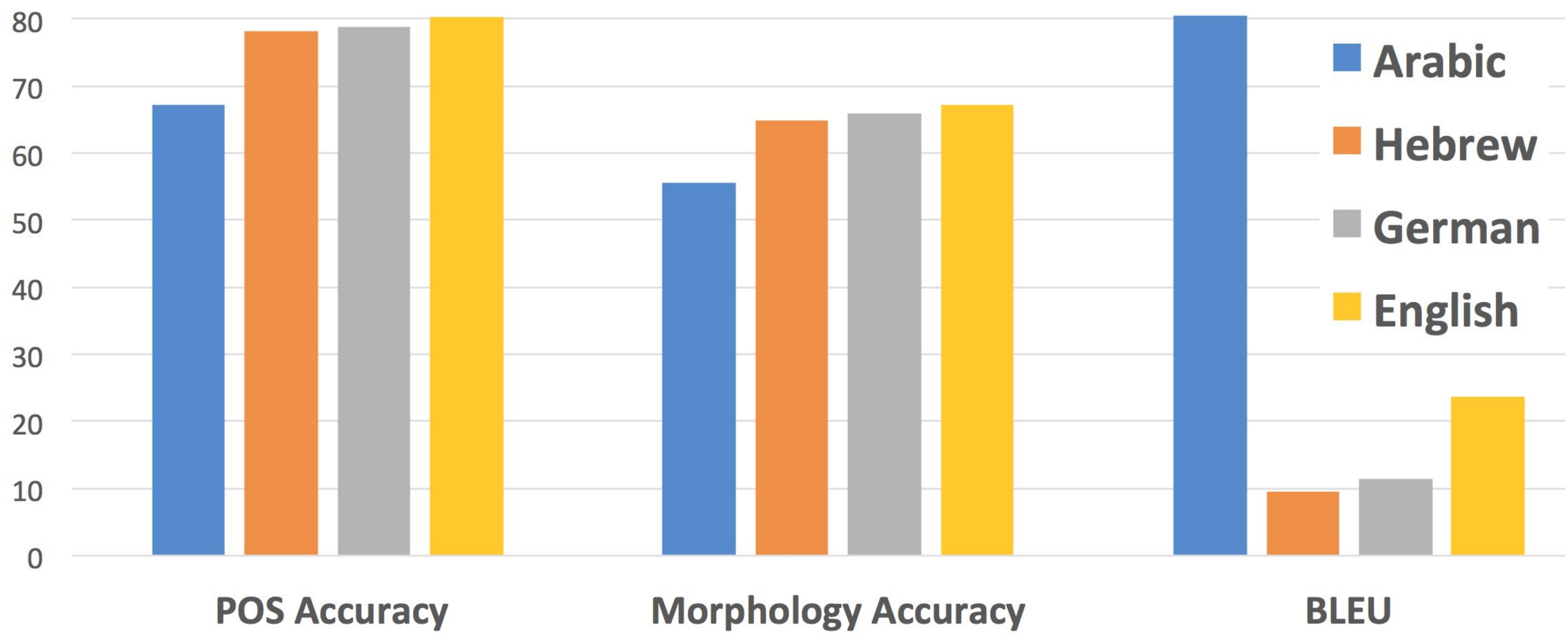
Effect of Target Language

- Arabic ⟶ other languages
- Different morphology of source-target ⟶ better source side representations



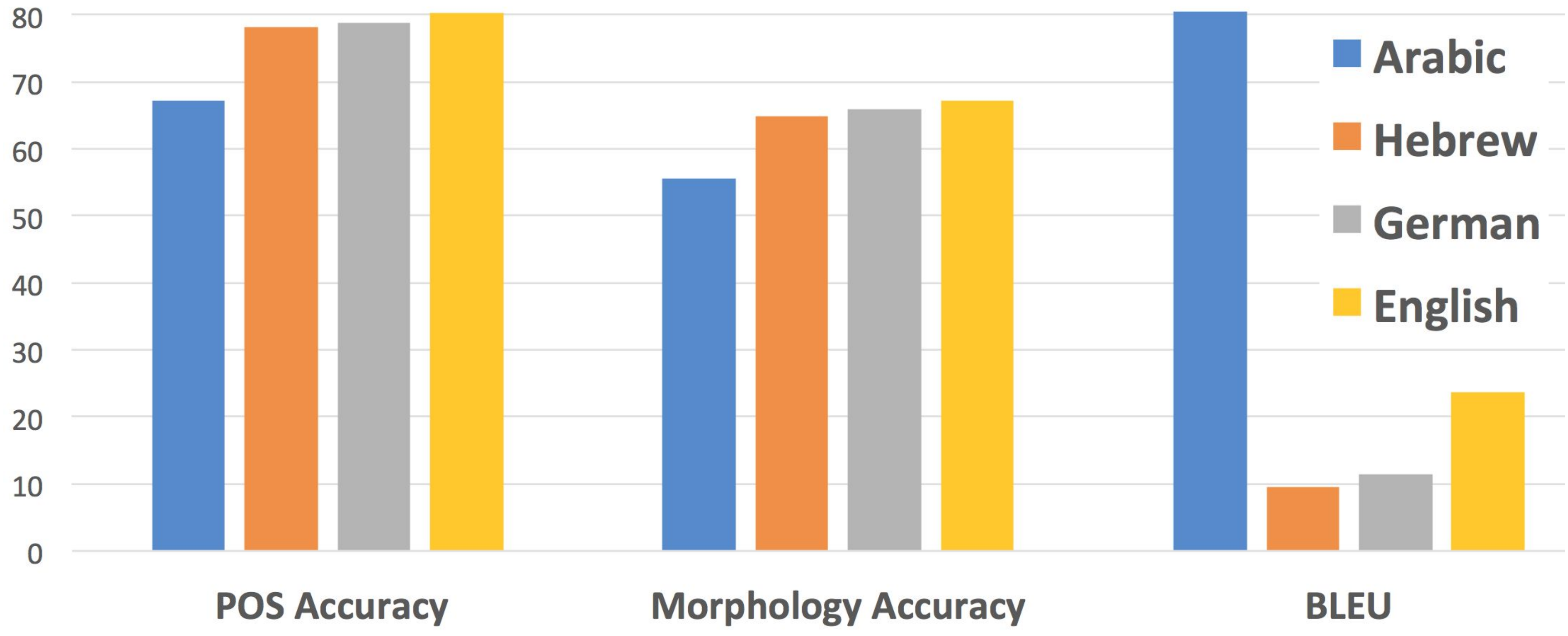
Effect of Target Language

- Arabic → Arabic translation
- Better BLEU but limited learning of morphology



Effect of Target Language

- Arabic $\longrightarrow$ Arabic translation
- Better BLEU but limited learning of morphology
- Easier task, but more of a memorization task



Extrinsic Evaluation Summary

- Neural MT representations contain useful information about word forms, semantics and syntax

- Lower layers focus on word-level features, such as part of speech tagging while higher layers learn more abstract phenomena, such as semantics and syntax

Summary

There are many ways to look into neural networks - none of them are perfect, but a combination of them may help us better understand what is going on.

Summary

There are many ways to look into neural networks - none of them are perfect, but a combination of them may help us better understand what is going on.

Understanding these networks is essential for us to build better and more efficient models!