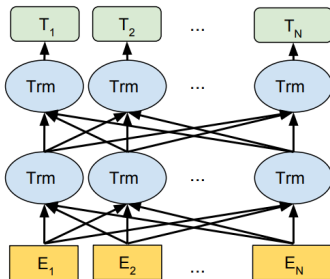


# BERT

## ARLMs vs. MLM



### Learning goals

- Understand the concept of self-supervision
- Gain ability to distinguish different types of language models

# AGAIN: WHAT IS A LANGUAGE MODEL?

- Statistical model that predicts text that fits well for a given context (typically also text)
- Auto-regressive LMs (ARLMs)
  - Predict one word that is highly likely given a prompt (previous words)
  - For predicting an entire text, repeat the process (i.e., extend the prompt with previously predicted words)
  - To predict a text from scratch, use an extra symbol <START> as the initial prompt

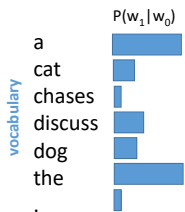
# ARLM: TOY EXAMPLE

---

**<START>**

**$w_0$**

# ARLM: TOY EXAMPLE

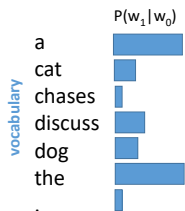


---

**<START>**

**$w_0$**

# ARLM: TOY EXAMPLE



---

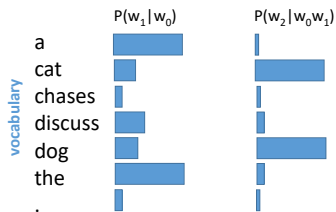
**<START>**

**the**

**$w_0$**

**$w_1$**

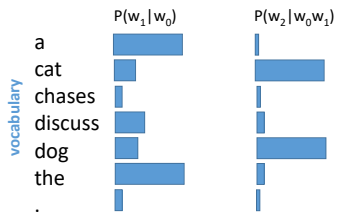
# ARLM: TOY EXAMPLE



---

**<START>**      **the**  
 **$w_0$**              **$w_1$**

# ARLM: TOY EXAMPLE



**<START>**

**$w_0$**

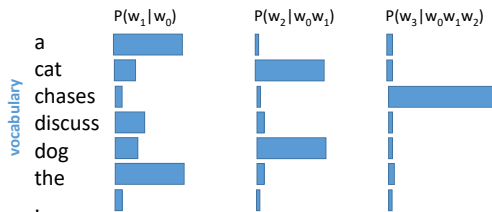
**the**

**$w_1$**

**dog**

**$w_2$**

# ARLM: TOY EXAMPLE



<START>

the

dog

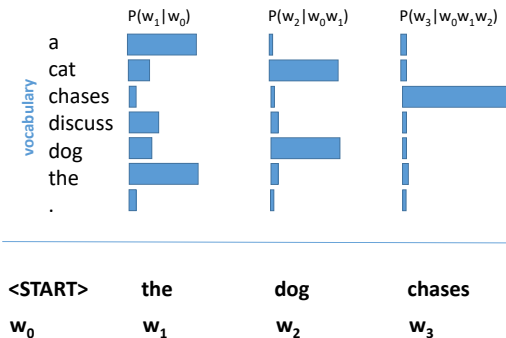
$w_0$

$w_1$

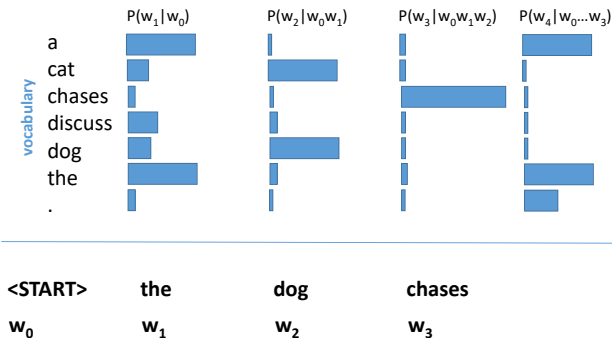
$w_2$



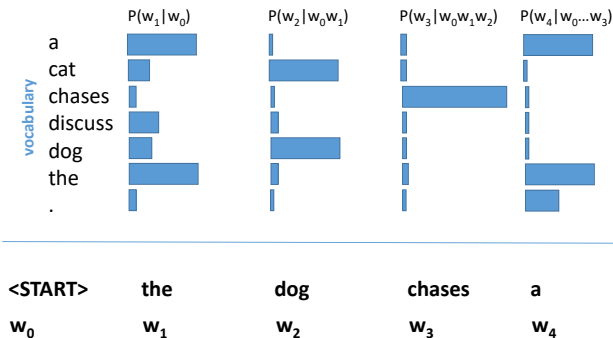
# ARLM: TOY EXAMPLE



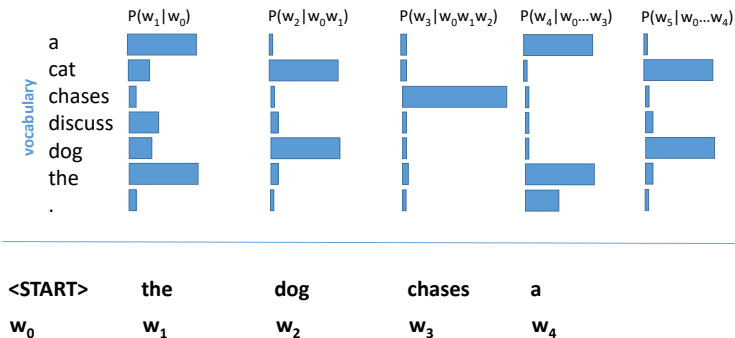
# ARLM: TOY EXAMPLE



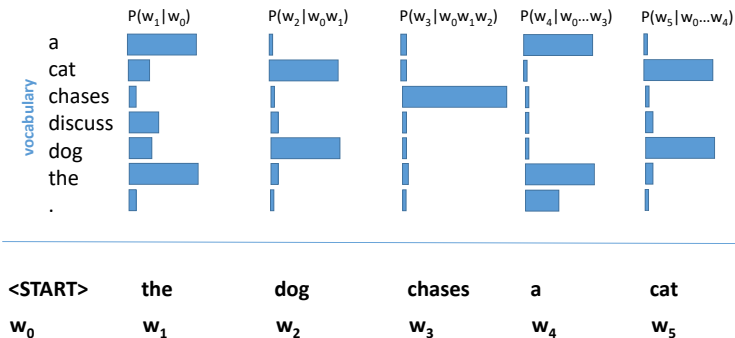
# ARLM: TOY EXAMPLE



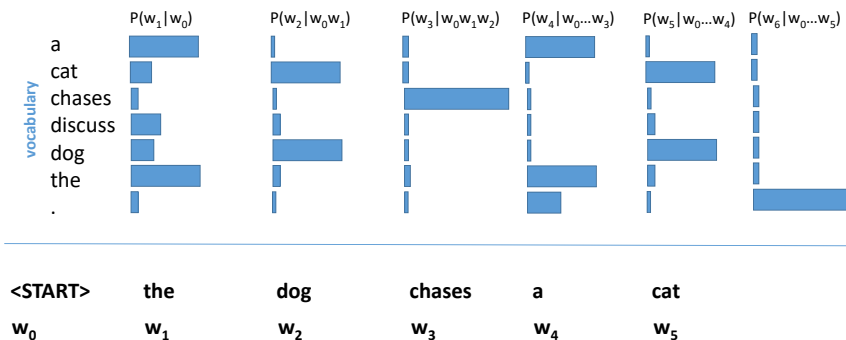
# ARLM: TOY EXAMPLE



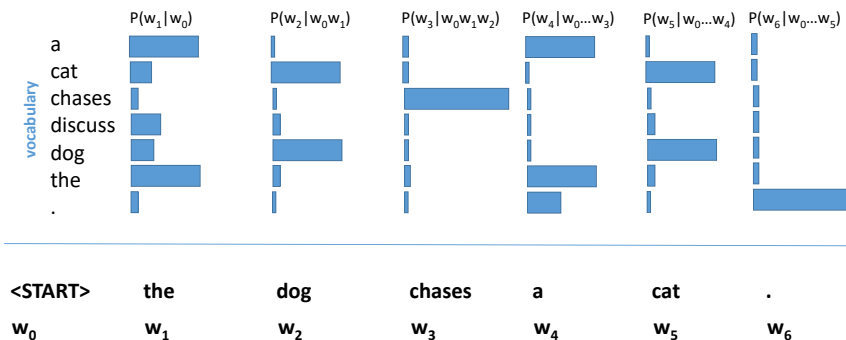
# ARLM: TOY EXAMPLE



# ARLM: TOY EXAMPLE



# ARLM: TOY EXAMPLE



# ARLM: PROBABILISTIC INTERPRETATION

- Gives an estimate for the probability of a sentence using conditional probabilities
- In general:

$$P(A \cap B) = P(B) \cdot P(A|B)$$

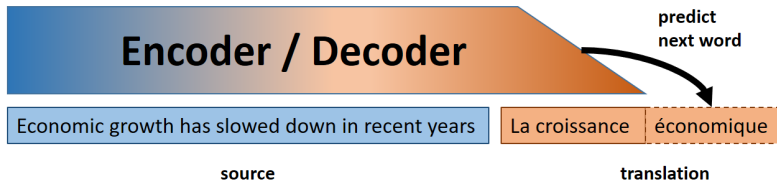
- For  $P(\text{sentence})$ :

$$\begin{aligned} &P(w_1, w_2, \dots, w_n | w_0) \\ &= P(w_1 | w_0) \cdot P(w_2 | w_0, w_1) \cdot \dots \cdot P(w_n | w_0, \dots, w_{n-1}) \end{aligned}$$



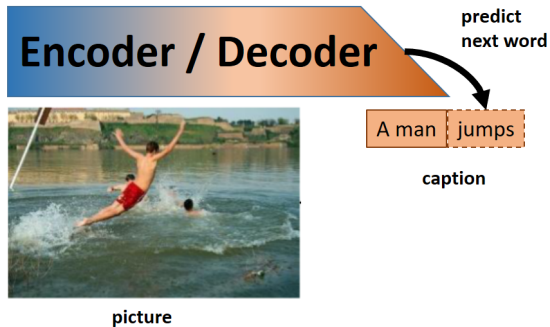
# EXAMPLES OF ARLMS (1)

## Neural Machine Translation:



# EXAMPLES OF ARLMS (1)

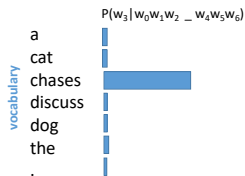
## Image Captioning:



# MASKED LANGUAGE MODELS (MLM)

- We have seen auto-regressive LMs
  - context: previous words
  - predict: next word
- Another type: *Masked* LMs (MLMs)
  - context: surrounding words
  - predict: masked word

# MLM: TOY EXAMPLE



---

<START>	the	dog	<MASK>	a	cat	.
$w_0$	$w_1$	$w_2$		$w_4$	$w_5$	$w_6$

12

# MLM: PROBABILISTIC INTERPRETATION

- Estimates  $P(w_i | w_0, w_1, w_{i-1}, \dots, w_{i+1}, w_n)$
- No “clean” estimate for  $P(sentence)$ , as

$$P(w_1, w_2, \dots, w_n | w_0)$$

$$\neq P(w_1 | w_0, w_2, \dots, w_n) \cdot P(w_2 | w_0, w_1, w_3, \dots, w_n)$$

- ARLMs are better than MLMs for generating texts
- Advantage of MLMs: Learning contextualized representations

# Self-supervised Learning



# DEFINITION

## *Unsupervised Learning:*

- No labels attached to the data
- Learn patterns / clusters from the features only

## *Supervised Learning:*

- (Gold) Labels attached to the data
- Learn from the association between features and labels

## **Self-Supervised Learning:**

- No *external* labels attached to the data  
→ Samples with suitable labels can be generated from the known structure of the data itself
- *Technically* supervised learning, but *no labeling effort* + simultaneous ability to generate massive amounts of labeled data points

# SELF-SUPERVISED OBJECTIVES

## Self-supervised objectives:

- Skip-gram objective (cf. word2vec)
- Language modeling objective
- *Masked language modeling (MLM)* objective
- ... and many more possibilities for text data