

Multimodal Natural Language Processing

Info 159/259 - April 20, 2021

Jon Gillick

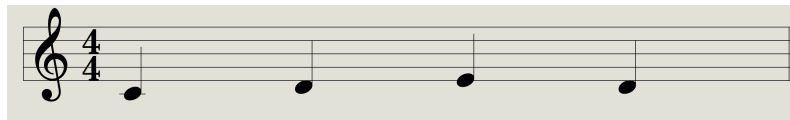
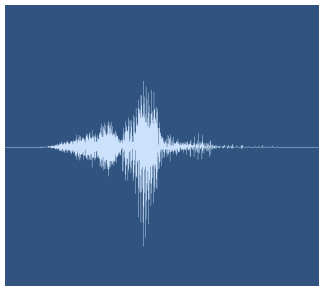
UC Berkeley
School of Information

NLP = Processing* Language with Computers

- Language is written...
- Language is spoken (e.g. speech recognition)...
- Language is sung (e.g. singing transcription)...
- Language is visual (e.g. sign language)...
- Language is tactile (e.g. Braille)...
- Language is **situated in the real world...**

NLP = Processing* Language with Computers

- Sometimes the information we are interested in is not written down in words.
- Multimodal NLP reasons about text while taking into account other modalities like image, audio, etc.



Today's Topics - Multimodal NLP

- Image Captioning (Text & Image)
- Lyrics Information Processing (Text & Music)

Today's Topics - Multimodal NLP

- **Image Captioning** (Text & Image)
- Lyrics Information Processing (Text & Music)

Image Captioning

- Automatically describing the content in an image

Image Captioning

- Automatically describing the content in an image

Tagging



Dog

Grass

Image Captioning

- Automatically describing the content in an image

Tagging



Dog

Grass

Captioning



Three dogs are playing in the grass

Image Captioning

- Automatically describing the content in an image



Input (X)



Three dogs play in the grass

Label (y)

Image Captioning

- Automatically describing the content in an image
- Modeling: $P(y \mid X)$



Input (X)



Three dogs play in the grass

Label (y)

Visual Features

- Pixel Intensities (Integers 0-255)



100% 50% 0% 0%

Visual Features

- Pixel Intensities (Integers 0-255)
- 3 Channels for Color



width height color

```
▶ doggos = cv2.imread('doggos.png')  
print(doggos.shape); print('\n')  
print(doggos)
```

(392, 606, 3)

[[[175 158 145]
[151 132 116]
[134 113 95]
...
[26 28 26]
[34 36 34]
[47 49 48]]

Height

Width

RGB (Red, Green, Blue)

CNN's for Image Processing

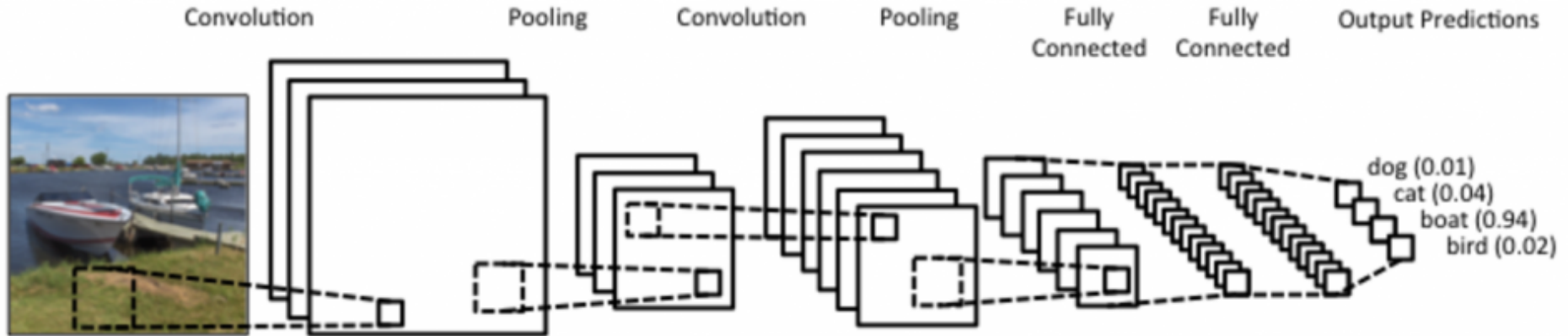
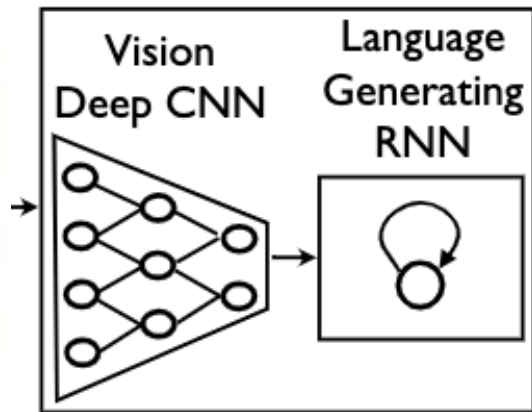


Image Captioning with a Multimodal Network



Three dogs play in the grass

Seq2Seq (Encoder-Decoder Framework)

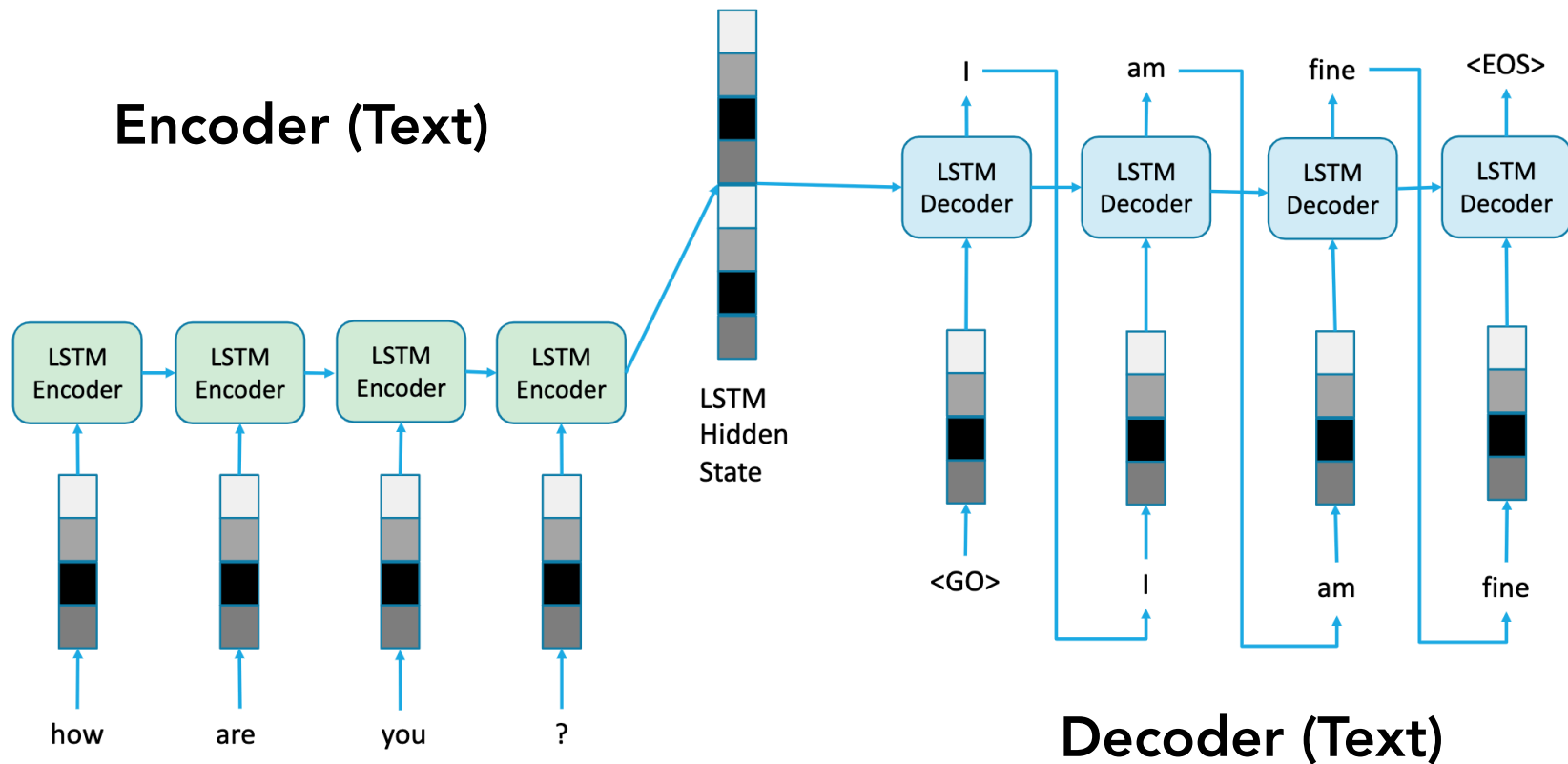


Image Captioning with a Multimodal Network

- Encoder-Decoder Framework (like Sequence to Sequence)

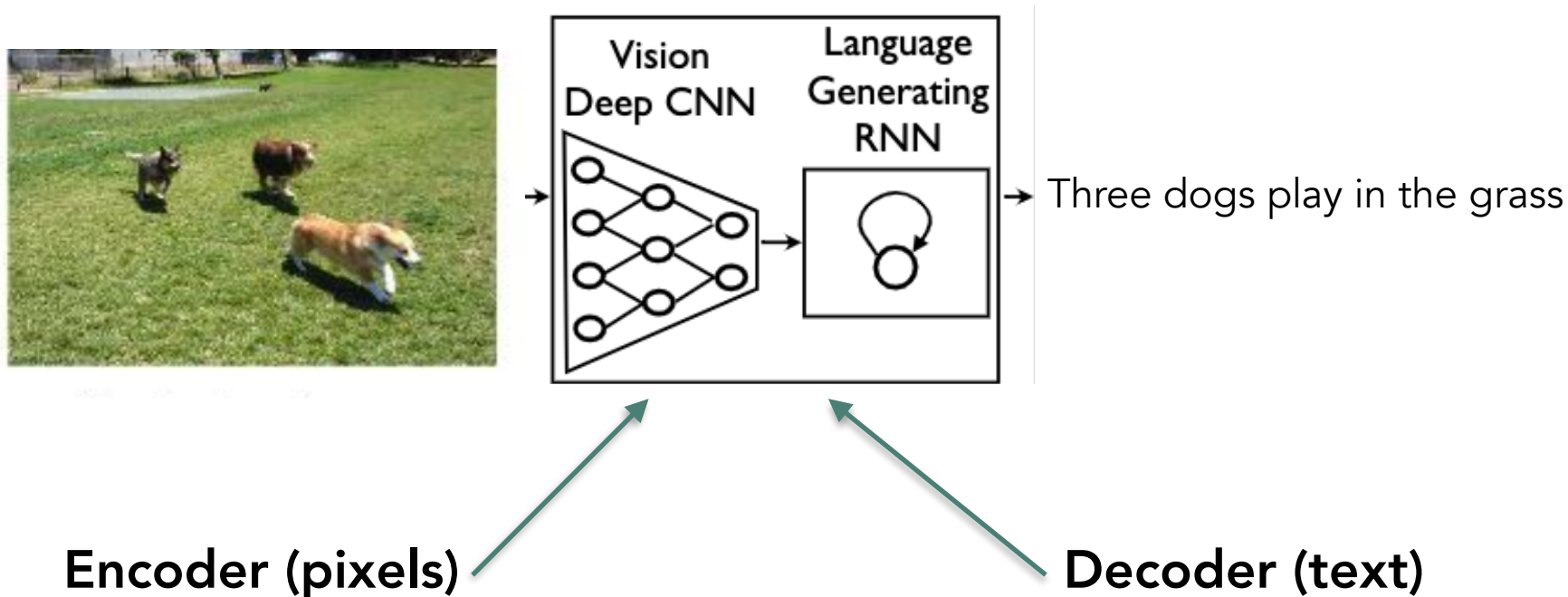
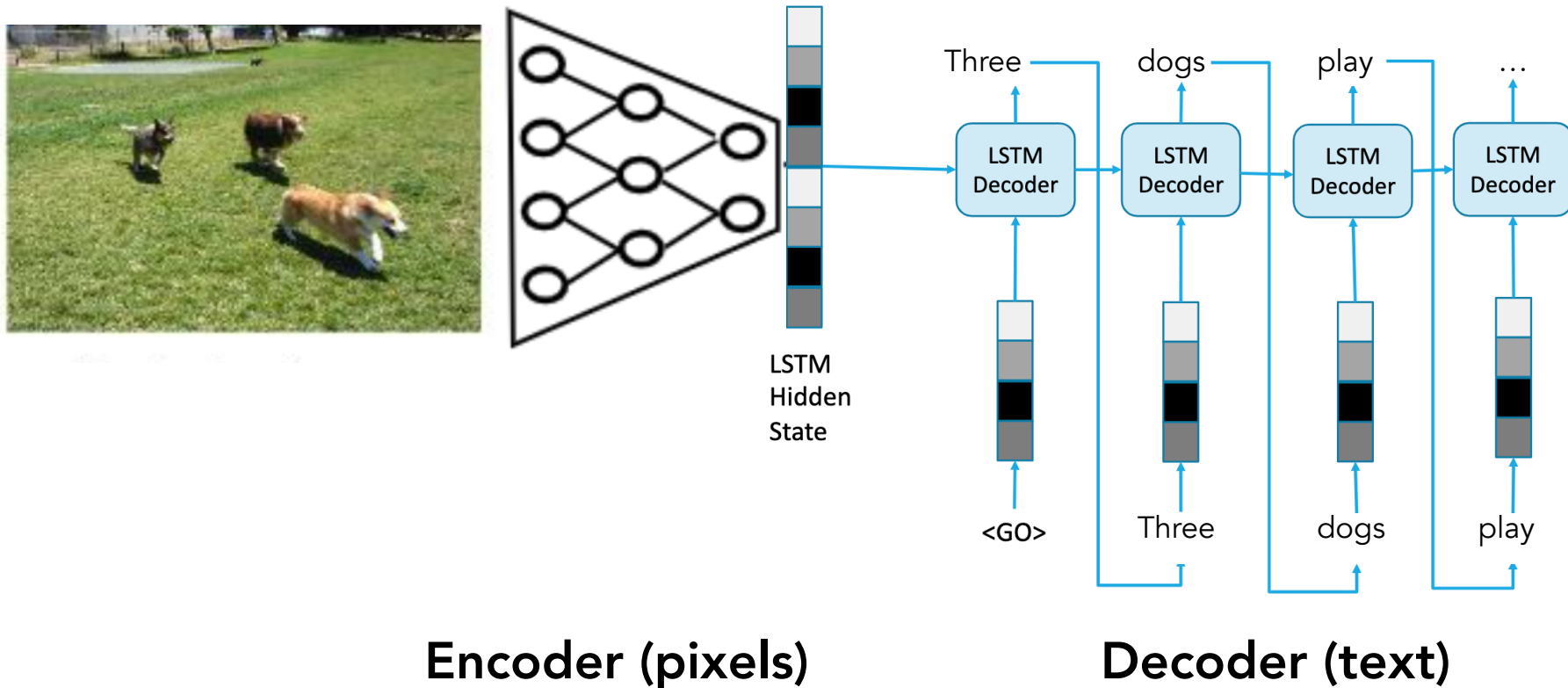


Image Captioning with a Multimodal Network



Data for Image Captioning

- Annotations here are captions written by crowdworkers
- “Each image has been annotated by labelers with 5 sentences that are relatively visual and unbiased”
- “Ground Truth” here is defined by what a collection of crowdworkers decided to write as captions.

Dataset name	size		
	train	valid.	test
Pascal VOC 2008 [6]	-	-	1000
Flickr8k [26]	6000	1000	1000
Flickr30k [33]	28000	1000	1000
MSCOCO [20]	82783	40504	40775
SBU [24]	1M	-	-

Image Captioning Results



Figure 5. A selection of evaluation results, grouped by human rating.

Image Captioning: Evaluation

- **BLEU Score** (also used in machine translation), measures n-gram overlap between predicted and reference captions
- **Ratings from crowdworkers** (e.g. "Describes without errors", "Describes with minor errors", "Unrelated to the image", etc.

Today's Topics - Multimodal NLP

- Image Captioning (Text & Image)
- **Lyrics Information Processing (Text & Music)**

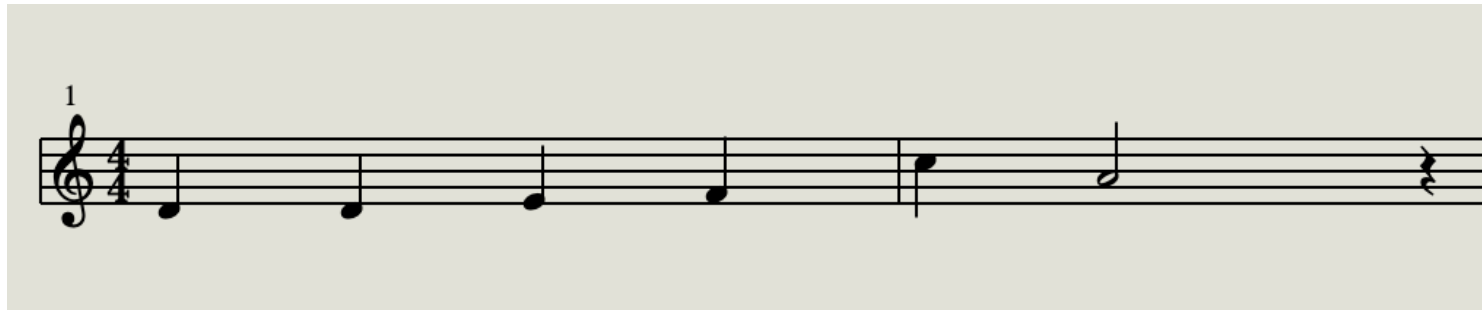
Jointly Modeling Melodies and Song Lyrics

- Motivation: Creativity Support Tools (e.g. suggestions for songwriters)
- Melody-Conditioned Lyrics Generation
- Lyric-Conditioned Melody Generating

Melody-Conditioned Lyrics Generation

- The linguistic context here is given by music
- We need to account for: the meaning of the words
- We need to account for: the fit with the melody

Fitting words with a melody



6 notes

Natural: I walked a- lone. This road

6 syllables

Awkward: Proceed to an unknown tomorrow

9 syllables

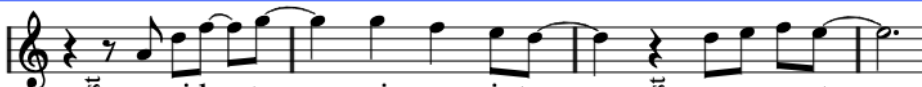
Aligning syllables with musical scores

Digital musical score data with syllables

Melody	Syllable
	Rest na ni ka ta ri na i to Rest o mo o ta

Aligning syllables with musical scores

Digital musical score data with syllables

Melody 

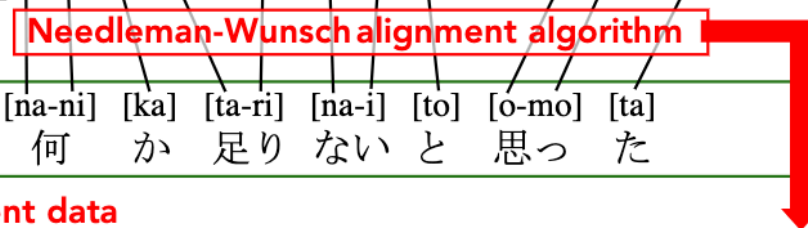
Syllable Rest na ni ka ta ri na i to Rest o mo o ta

Lyric text data with syllable and boundary

Syllable [na-ni] [ka] [ta-ri] [na-i] [to] [o-mo] [ta]

Word <BOL> 何 か 足り ない と 思っ た

Needleman-Wunsch alignment algorithm



Melody-Lyric alignment data

Melody 

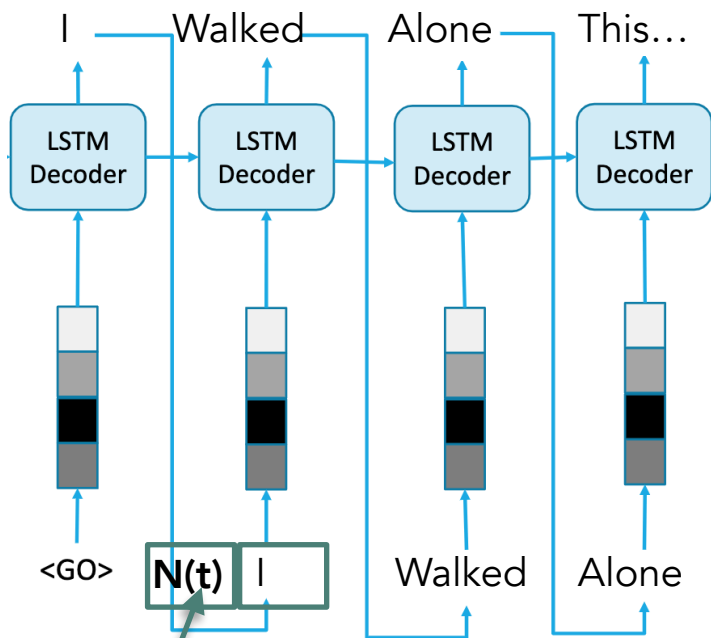
Syllable Rest na-ni ka ta-ri na-i to Rest o-mo ta

Word <BOL> 何 か 足り ない と Rest 思っ た

(some-thing) (FUNC)(enough) (not) (FUNC) (think) (FUNC)

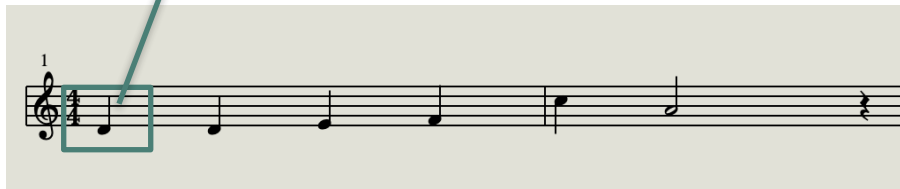
(I thought that something was missing)

Conditioning LSTM on Melody Features



$N(t)$: Melody feature at time t

- Embedding Vector for Melody
- Based on pitches, durations, syllables



Text generation in different (multimodal) contexts

- Chatbot

Context: "How are you"

- Image Captioning

Context:



- Lyrics Generation

Context:

