Multimodal Entity Synonym Set Expansion and Visually-Synonyms-Aware Fine-Tuning

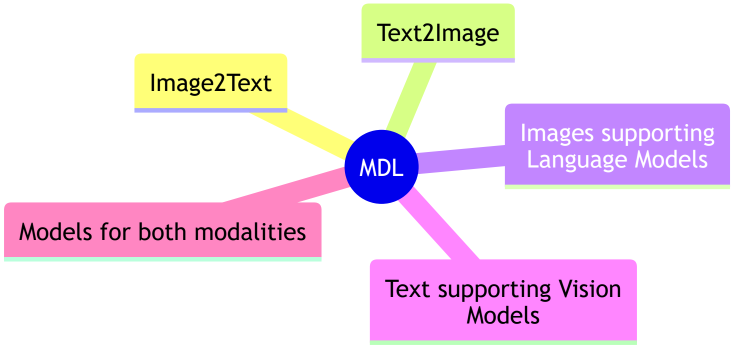
Junjie Chen

10/14/2023

Intro: Multimodal Deep Learning

From: (LMU Munich, Germany)

Research Direction



MDL

Image2Text

Text2Image

Images supporting Language Models

Models for both modalities

Text supporting Vision Models

Read more about Multimodal Deep Learning

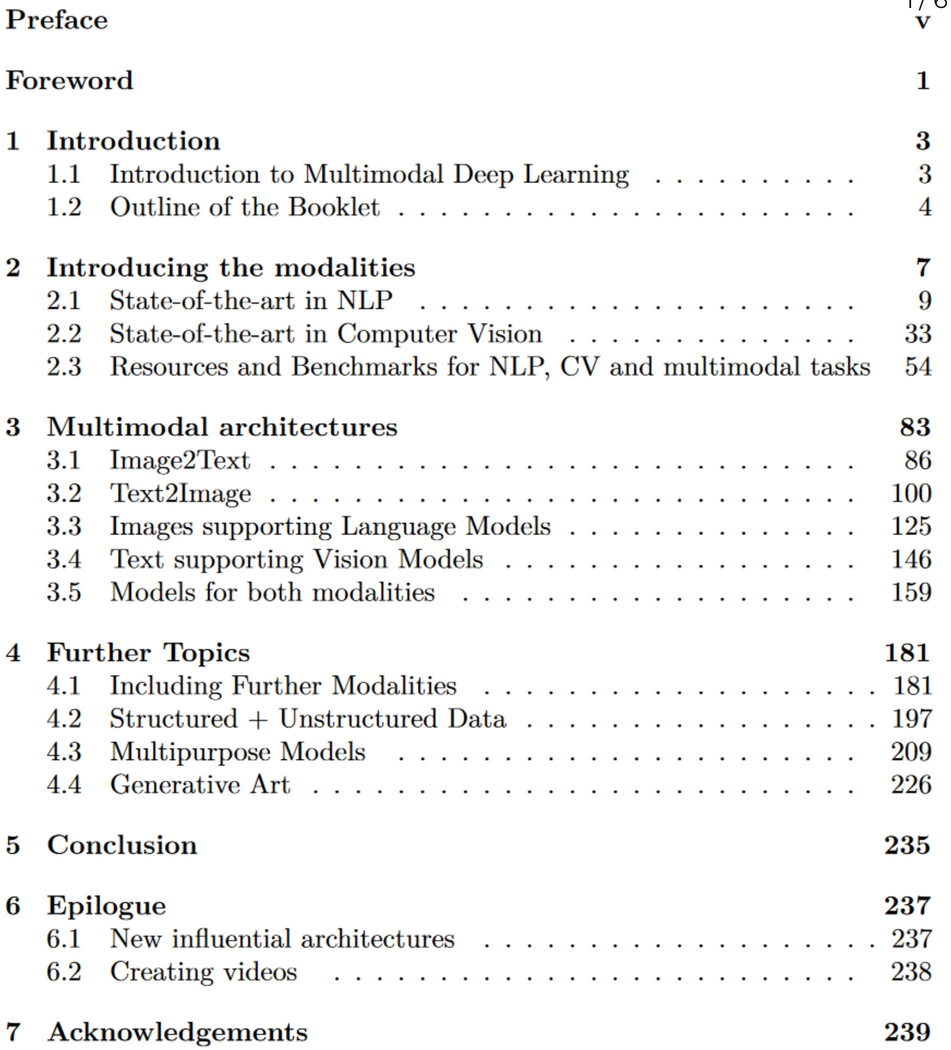


Table of contents

1. Intro: Multimodal Deep Learning
2. Words In (Non-Symbolic) Contexts
3. Vokenization^1
4. Code Analysis...

Words In (Non-Symbolic) Contexts

Symbol Grounding Problem [1]:

- It asserts that it is not possible to understand the meaning (semantics) of a word by just looking at other words because words are essentially meaningless symbols.
- It is possible to understand the meaning only if **the word is put in a context**, **a perceptual space**, other than that of written language: the word **must be grounded in non-symbolic representations, like images**, for example.
- ChatGPT大模型技术争议与符号奠基问题

[1] (Harnad, S. (1990). The symbol grounding problem. 42(1-3):335–346.) ↩

Vokenization[1]

Voken:

we assume a text corpus where each token is aligned with a related image. Hence, these images could be considered as visualizations of tokens and we name them as 'vokens'.

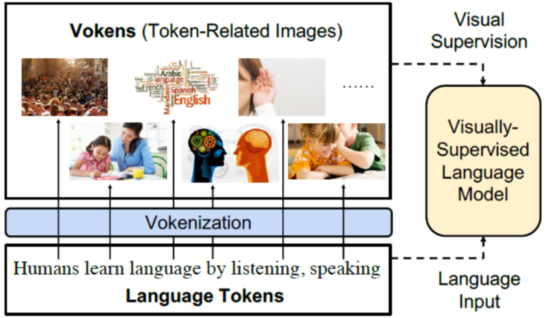

Vokens (Token-Related Images)
Arabic
language
French
Spanish
English
......
Visual Supervision
Visually-Supervised Language Model
Vokenization
Humans learn language by listening, speaking
Language Tokens
Language Input

FIGURE 3.41: From Tan and Bansal (2020). Visually supervised the language model with token-related images, called Vokens.

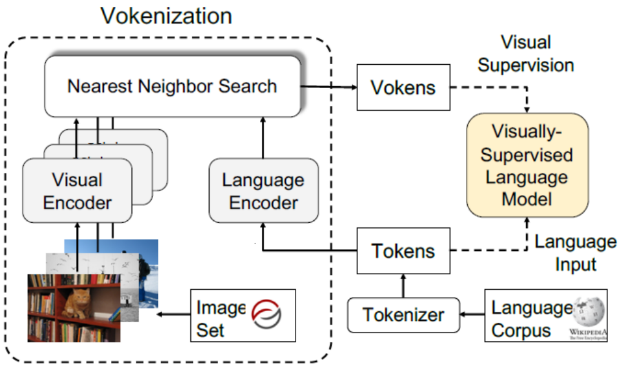


FIGURE 3.43: From Tan and Bansal (2020). The Vokenization process. A contextualized image (visual token, Voken) is retrieved for every token in a sentence and with this visual token, visual supervision is performed.

1. Tan, H. and Bansal, M. (2020). Vokenization: Improving language understanding with contextualized, visual-grounded supervision. ↩

Code Analysis...