Language Technology and Artificial Intelligence
Trends and perspectives

Benjamin Roth
Digitale Textwissenschaften

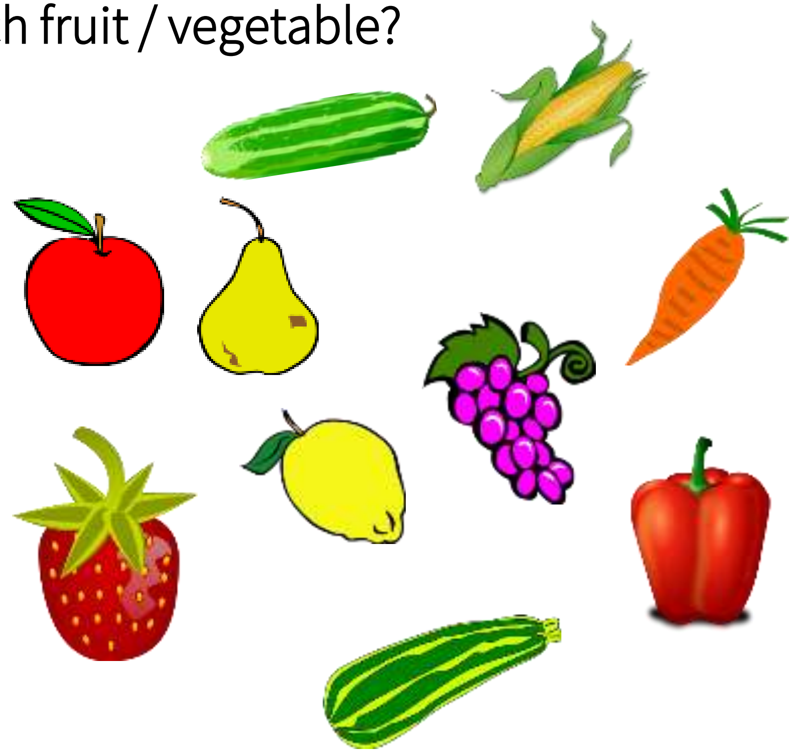
Universität Wien

Neural networks
& Pre-trained language models

Classical Machine Learning: Predict Output Classes from Input Features

Camera

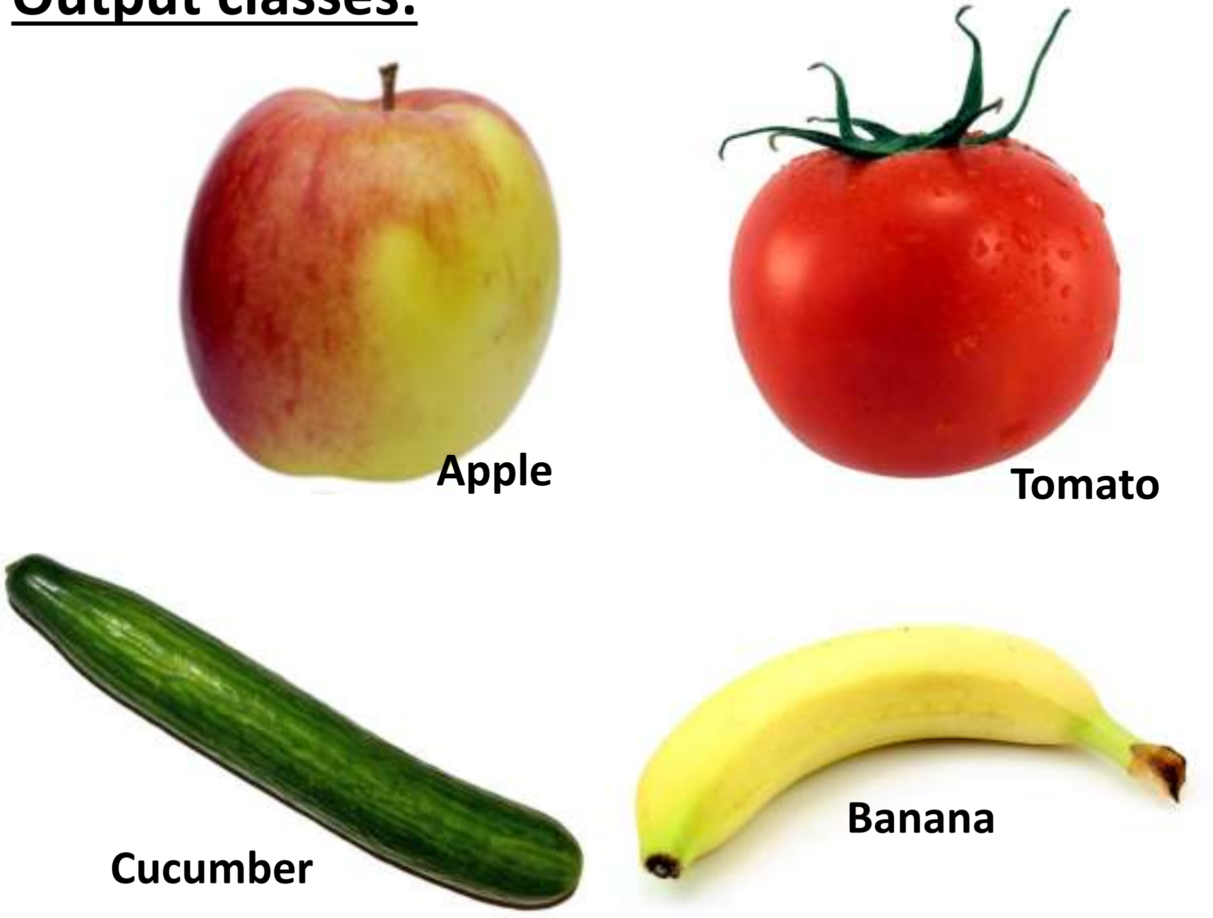
Which fruit / vegetable?



what are good features?

Predict Output Classes from Input Features

Output classes:



Features:

- **red**
- **yellow**
- **green**
- **round**
- **long**

Predict Output Classes from Input Features

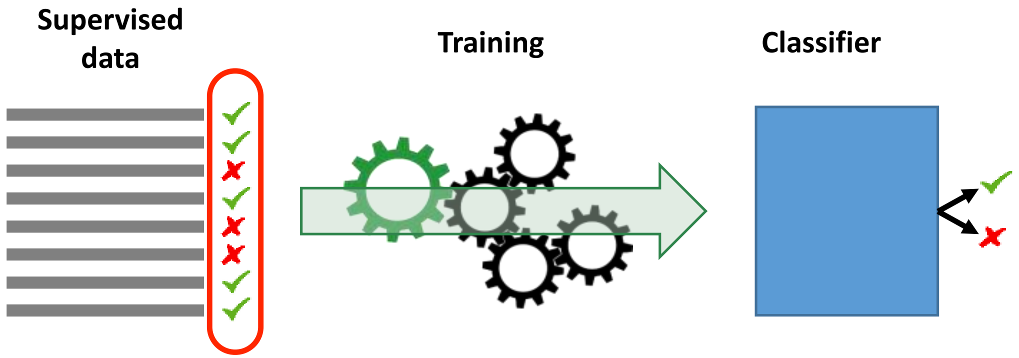
- We can represent inputs and outputs as vectors
  - **Input:** values of input features (``redness'', roundness, …)
  - **Output:** probability of every output class

$$\begin{array}{l} \text{red} \\ \text{yellow} \\ \text{green} \\ \text{round} \\ \text{long} \end{array} \begin{bmatrix} 0.9 \\ -1 \\ 0.1 \\ 0.8 \\ 0.2 \end{bmatrix} \xrightarrow{\text{Prediction}} \begin{bmatrix} 0.15 \\ 0.8 \\ 0.05 \\ 0 \end{bmatrix} \begin{array}{l} \text{apple} \\ \text{tomato} \\ \text{cucumber} \\ \text{banana} \end{array}$$

Predicted class

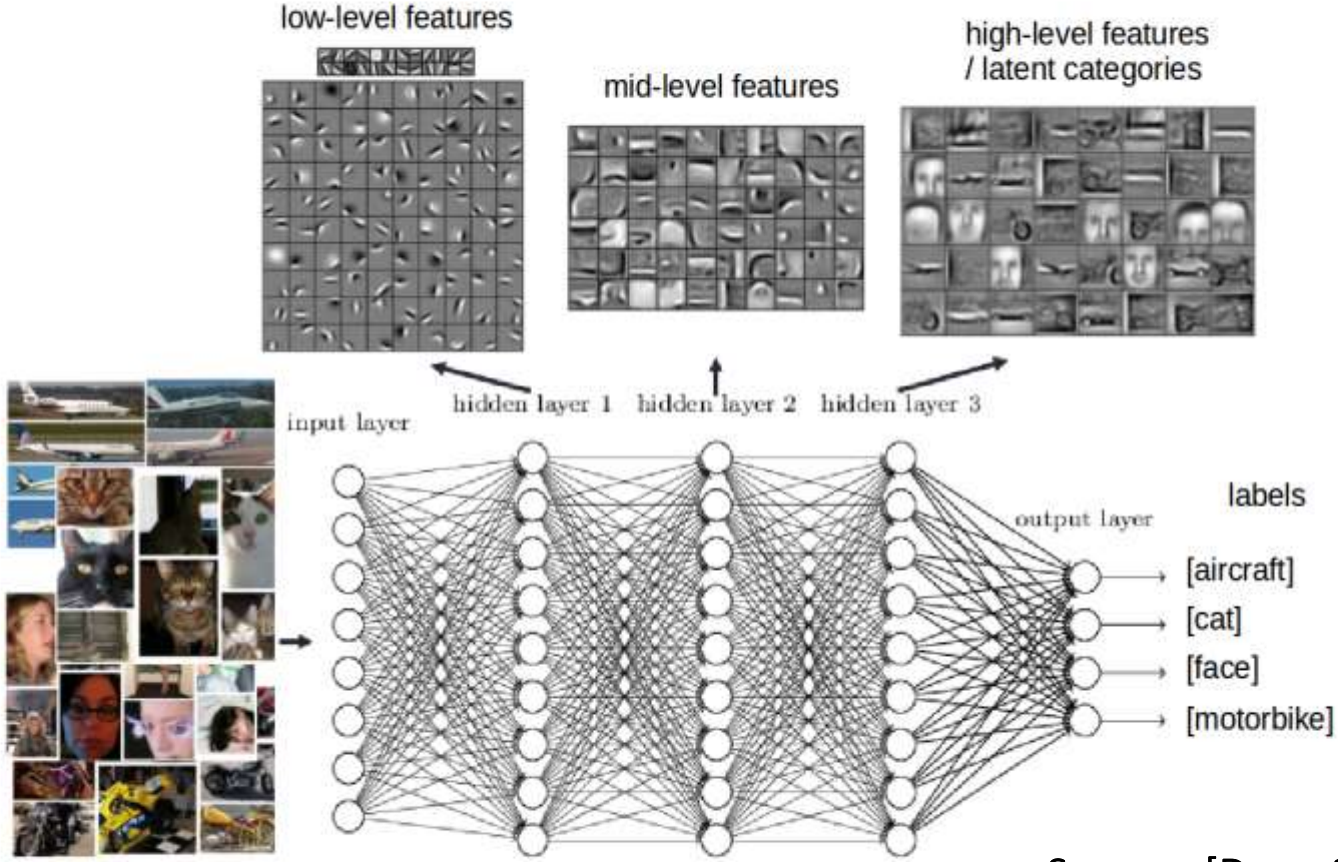
Training & Evaluation

- The training algorithm needs
  - **examples** to learn from
  - a way to **measure** whether it makes progress
- Annotation/labels:
  - Experts look at inputs, and **annotate** the correct output **labels**
  - Sometimes labels can be obtained as a side-product of some activity (Users move emails in SPAM folder)



Representation learning = deep learning = neural networks

- **Raw input** instead of defined feature representation:
  - Images: Pixels
  - Text: Sequence of words or characters
- **Learn higher-level abstractions**

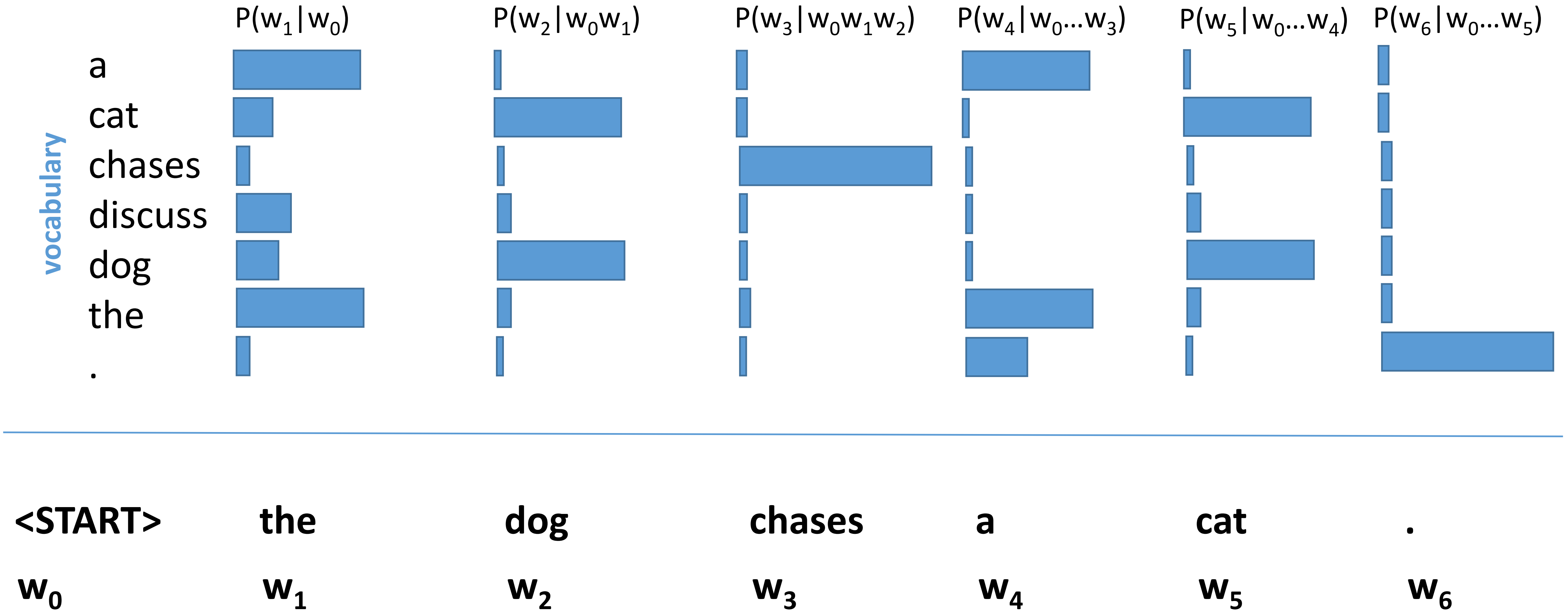


Source: [Deng 2009; Lee 2009]

What is a (statistical) language model? (LM)

- Statistical model that **predicts text** that fits well for a given **context** (typically also text)
  - 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

Language model (toy example)



Demo: GPT-3

- Popular language models:
  - BERT, trained on 3300M words (Wikipedia+BooksCorpus)
  - GPT-3, trained on 500000M words (CommonCrawl+Webtext2+Books+Wikipedia)
- GPT-3 [Brown 2020], https://openai.com/api/
- GPT-J, https://huggingface.co/EleutherAI/gpt-j-6B

Overview Documentation Examples Playground

Upgrade Help B Digital Philology University of Vienna

Save View code Share ···

My name is Benjamin Roth, I am a researcher at the University of Vienna, specializing in Machine Learning and Digital Philology. Today I will talk about trends and perspectives of language technology and artificial intelligence.

Generate 40

Save View code Share ···

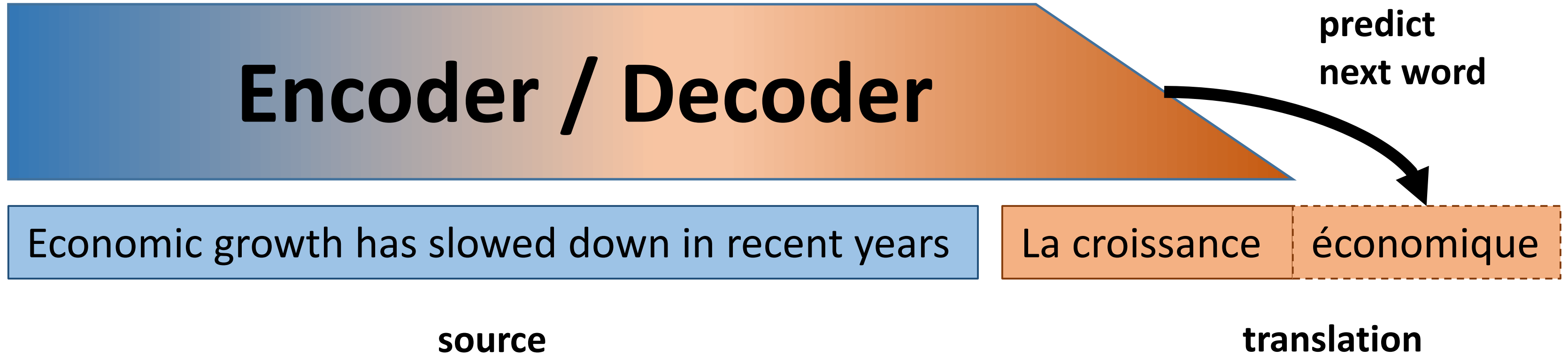
My name is Benjamin Roth, I am a researcher at the University of Vienna, specializing in Machine Learning and Digital Philology. Today I will talk about trends and perspectives of language technology and artificial intelligence.

First I will make a brief survey of trends in the development of language technology. I will look at the development of algorithms for language generation, language translation, parsing and language understanding. I will then say a few words about the future of artificial intelligence.

Language technology can be seen as the field of applications of artificial intelligence to language. It is not surprising that the development of language technology is driven by progress in artificial intelligence. The number of applications of language technology has increased dramatically in the last few years. The application areas cover important aspects of our society like education, commerce, business, government and entertainment.

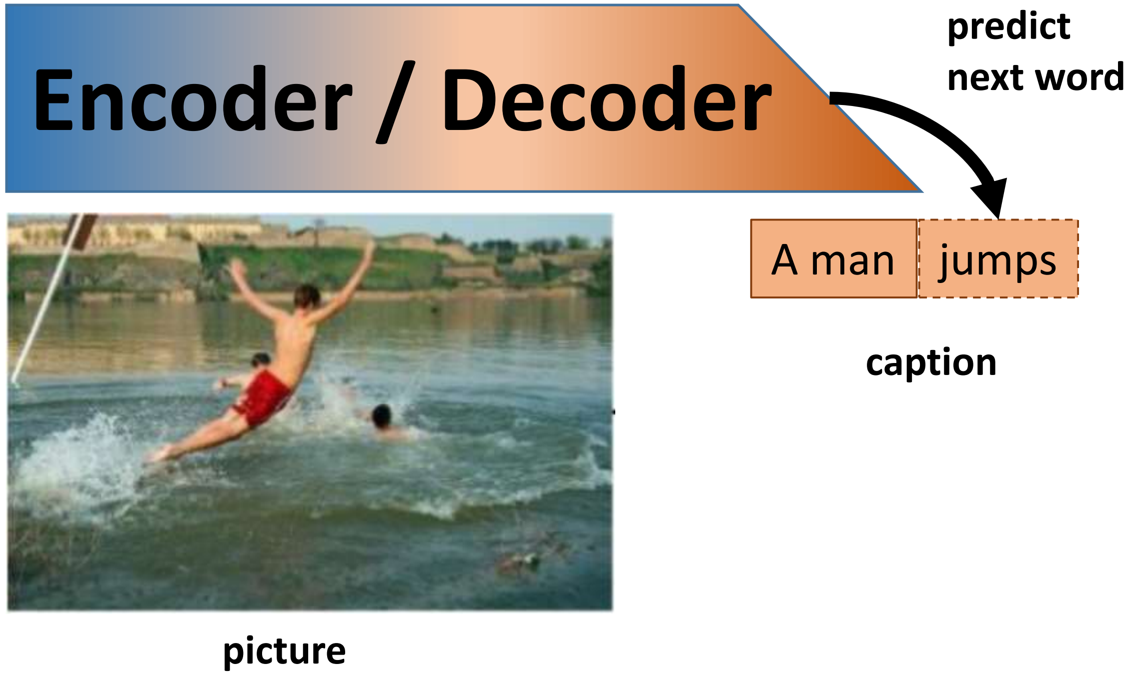
Generate 165

Neural Machine Translation



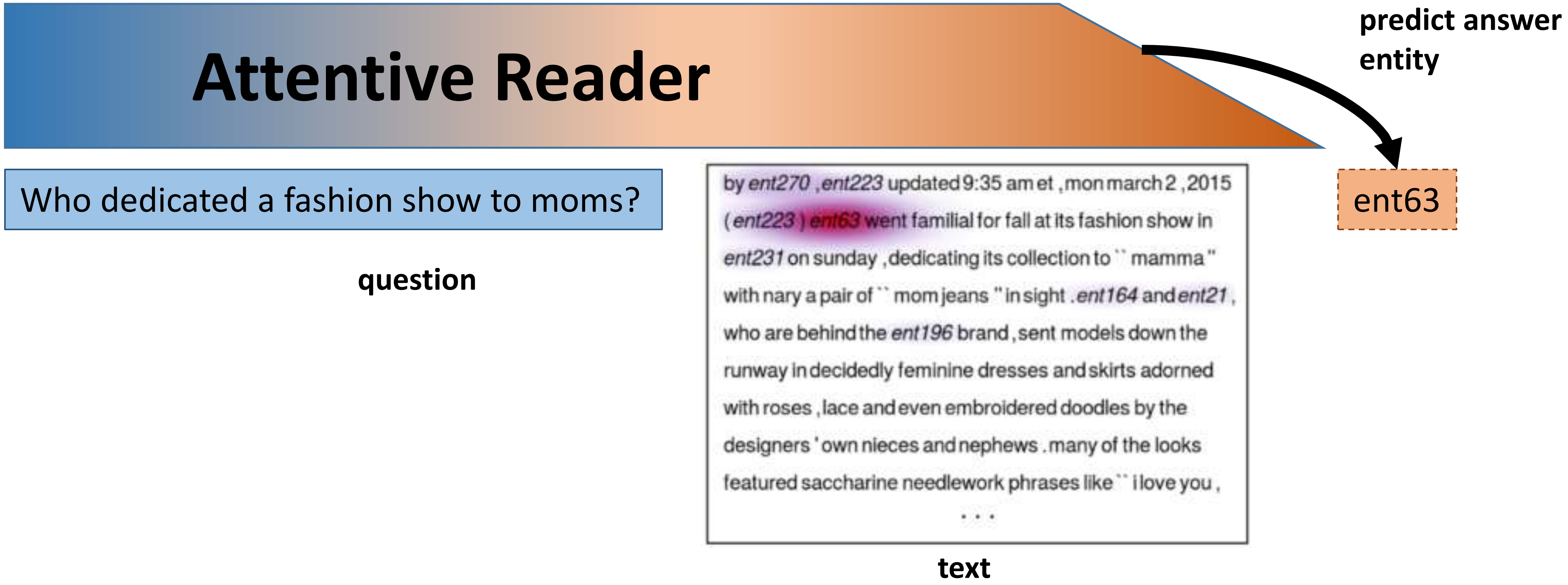
[Sutskever 2014; Bahdanau 2015; Vaswani 2017 …]

Image captioning



[Kiros 2014; Mao 2014; Xu 2015;…]

Question Answering



[Hermann 2015, Seo 2017, …]

Deep learning limitations (and how to overcome them)

- Lack of training data
  - → domain adaptation, transfer learning [Howard & Ruder 2018]
  - → unsupervised pre-training [Devlin 2018, Brown 2020]
- Difficulty to leverage human expertise
  - → combine with rule-based systems, weak supervision [Ratner 2017, Sedova 2021]
- Lack of insight
  - → Automated explanations [Poerner 2018, Sydorova 2019]

Perspectives & open problems

Statistical models = intelligence?

- **No.** ([Bender & Koller, 2020], tradition of [Searle, 1980])
  - Just repetition of already observed superficial patterns.
  - You wouldn't rely on a language model (or a parrot) on advice for fighting a bear.
  - **But:** The models **do** generalize to some degree.
- **Why not?**
  (e.g. Glue Benchmark [Wang 2019], tradition of [Turing, 1950])
  - Ultimately we need to rely on observed behaviour.
  - Define a task, and measure success rate on examples not seen in training.
  - Intelligence is task-specific, not a global property.

Transparent and explainable predictions

- Why is a husky classified as a wolf?
  (LIME [Ribeiro 2016])
- Why is a social media post classified as hate speech?
  (Hatecheck [Röttger 2021])
- Why is a loan approved or rejected?

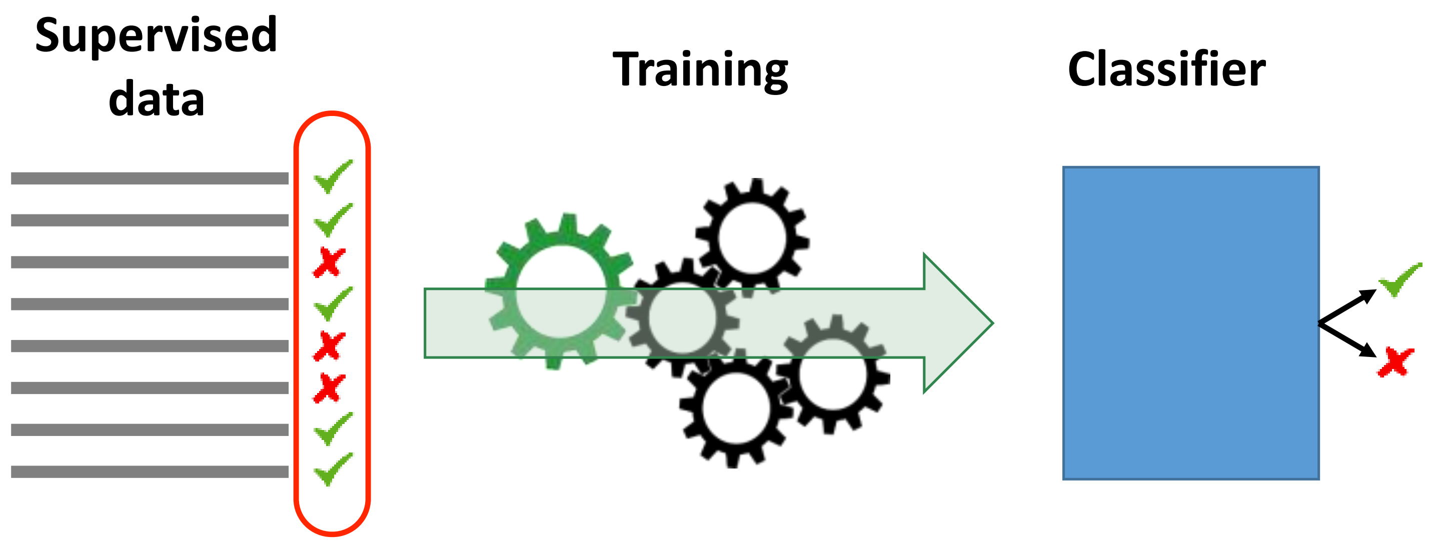
→ Which explanations methods are reliable? [Poerner 2018, Sydorova 2019]

(a) Husky classified as wolf (b) Explanation

- **Right to explanation** (EU GDPR Recital 71):
  - ``[safeguards include …] the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision.''

Expert knowledge or annotated data?

- Supervised training: learn a function that maps an input to an output based on labeled examples



But what if labels are not available?

→ Combine machine learning with expert knowledge encoded in rule-based systems! [Sedova 2021]

knodle

Outlook

Thank you!
Questions?

- Modern neural language models
  - leverage enormous amounts of data
  - on some tasks achieve performance not thought possible before
- These are exciting times!
- Now, we need to think more about
  - transparency, accountability and fairness
  - what we understand by *intelligence*
  - how to flexibly approach problems without training data

References

- [Bahdanau 2015] Dzmitry Bahdanau, Kyunghyung Cho, Yoshua Bengio. Neural machine translation by jointly learning to align and translate.
- [Bender&Koller 2020]Bender, Emily M., and Alexander Koller. "Climbing towards NLU: On meaning, form, and understanding in the age of data." ACL 2020.
- [Brown 2020] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. Language Models are Few-Shot Learners. NeurIPS 2020.
- [Cybenko 1989] G. Cybenko Approximations by superpositions of sigmoidal functions
- [Deng 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. CVPR 2009.
- [Devlin 2018] J Devlin, MW Chang, K Lee, K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding.

References

- [Hermann 2015] Karl Moritz Hermann, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, Phil Blunsom. Teaching Machines to Read and Comprehend.
- [Howard & Ruder 2018] Jeremy Howard, Sebastian Ruder. Universal Language Model Fine-tuning for Text Classification
- [Hornik 1991] Kurt Hornik. Approximation Capabilities of Multilayer Feedforward Networks
- [Kiros 2014] Ryan Kiros, Ruslan Salakhutdinov, Rich Zemel. Multimodal Neural Language Models.
- [Lee 2009] Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng. Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations.
- [Mao 2014] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, Alan Yuille. Deep Captioning with Multimodal Recurrent Neural Networks (m-RNN).

References

- [Perez 2018] Guillermo Valle-Pérez, Chico Q. Camargo, Ard A. Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions.
- [Poerner 2018] Nina Poerner, Benjamin Roth, Hinrich Schütze. Evaluating neural network explanation methods using hybrid documents and morphological agreement.
- [Radford 2019] A Radford, J Wu, R Child, D Luan, D Amodei, I Sutskever. Language Models are Unsupervised Multitask Learners.
- [Ratner 2017] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, Christopher Ré. Snorkel: Rapid Training Data Creation with Weak Supervision.
- [Röttger 2021] Paul Röttger, Bertram Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, Janet B. Pierrehumbert. HateCheck: Functional Tests for Hate Speech Detection Models.
- [Ribeiro 2016] Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Model-agnostic interpretability of machine learning."

References

- [Searle 1980] John Searle. Minds, Brains, and Programs.
- [Sedova 2021] A Sedova, A Stephan, M Speranskaya, B Roth. Knodle: Modular Weakly Supervised Learning with PyTorch.
- [Sydorova 2019] Alona Sydorova, Nina Poerner, Benjamin Roth. Interpretable question answering on knowledge bases and text.
- [Seo 2017] Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, Hannaneh Hajishirzi. Bidirectional Attention Flow for Machine Comprehension.
- [Sutskever 2014] Ilya Sutskever, Oriol Vinyals, Quoc V. Le. Sequence to Sequence Learning with Neural Networks.
- [Strubell 2019] Strubell, Emma, Ananya Ganesh, and Andrew McCallum. "Energy and policy considerations for deep learning in NLP."

References

- [Turing 1950] Turing, Alan. "Computing Machinery and Intelligence". In: Mind (October 1950).
- [Vaswani 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin . Attention is all you need.
- [Wang 2018] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). GLUE: A multi-task benchmark and analysis platform for natural language understanding.
- [Wang 2021] Wang, Ben and Komatsuzaki, Aran. GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model.
- [Xu 2015] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention.