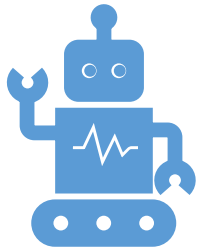
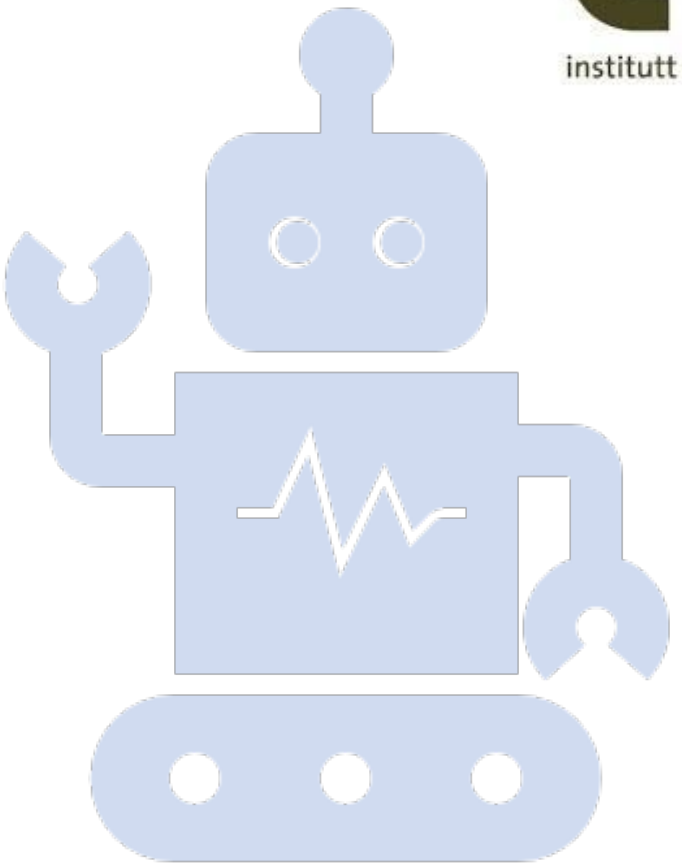
UiO : University of Oslo

institutt for informatikk



# Kunstig intelligens (AI) og store språkmodeller



Jan Tore Lønning
Institutt for informatikk/
Language Technology Group

# Kunstig intelligens

- Først inspirert av tenkning, logikk, regler
- *I X my Y* ➔ *Why do you X your Y?*
  - (Eliza, Joseph Weizenbaum, 1966)
- Store vyer, moderat suksess:
  - varierende interesse og finansiering
- Hvordan kan en maskin f.eks. gjenkjenne
  - ansikter?
  - trafikkskilt?
  - fingeravtrykk?

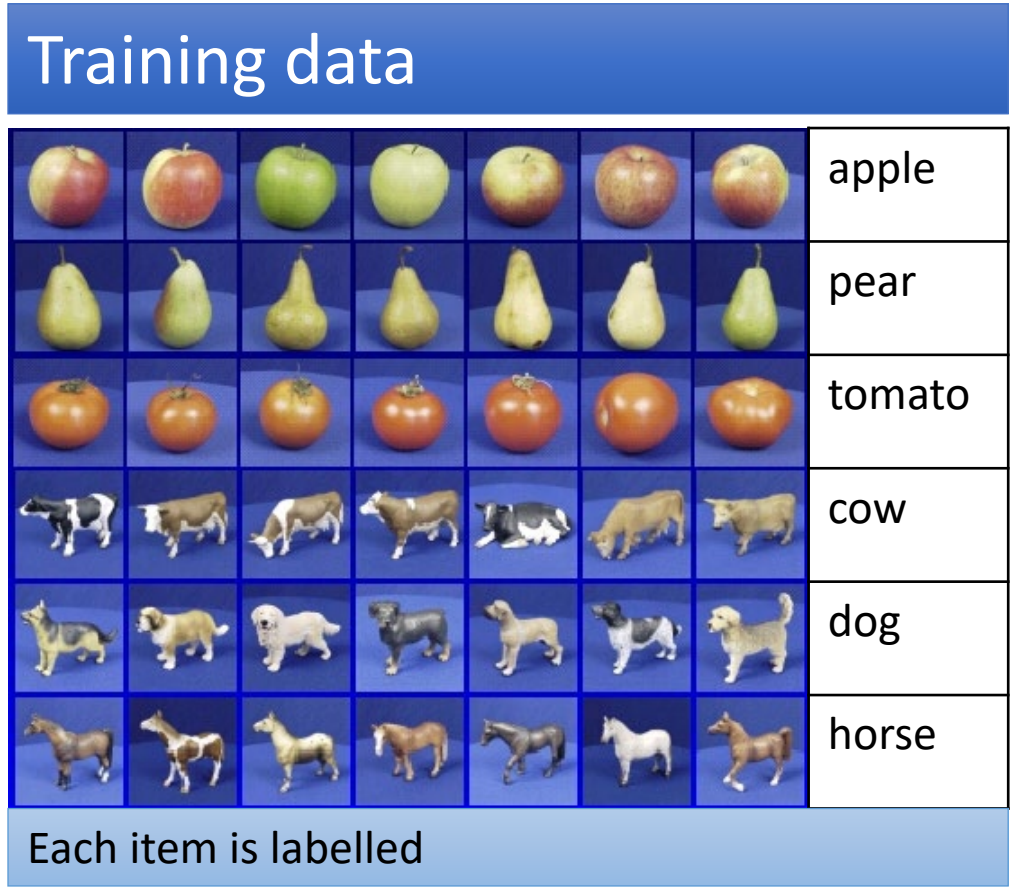
*…every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it…*

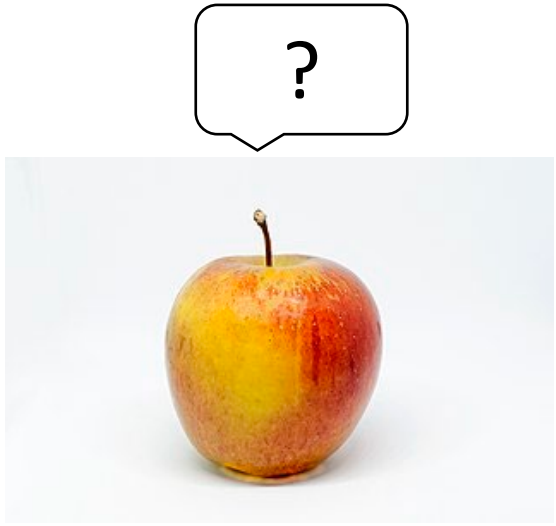
*…solve kinds of problems now reserved for humans…*

John McCarthy et. al. 1956

# Maskinlæring

= Overvåket ("supervised") læring

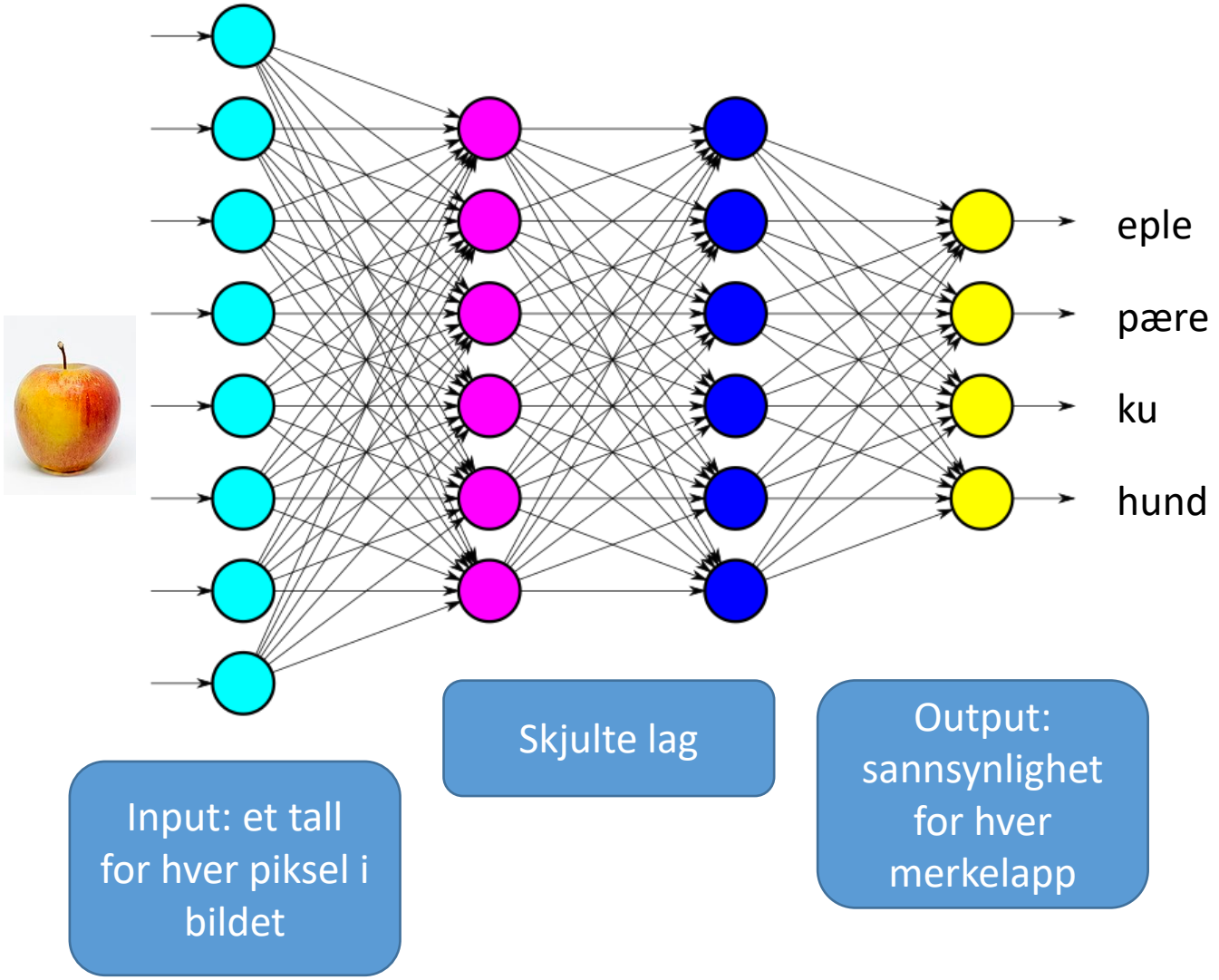



Oppgave: Riktig merkelapp til nye ting (bilder)

# Nevrale nettverk

- Kantene ("synapser"):
  - Vekter (tall) * signalene
  - Matematiske funksjoner på nodene
- Prediksjon:
  - Send tallene igjennom
  - Velg mest sannsynlige merkelapp
- Trening:
  - Prediker
  - Sammenlign med riktig lapp
  - Juster vektene automatisk




network from https://openclipart.org/detail/290666/neural-net-graph

# Nevrale nett

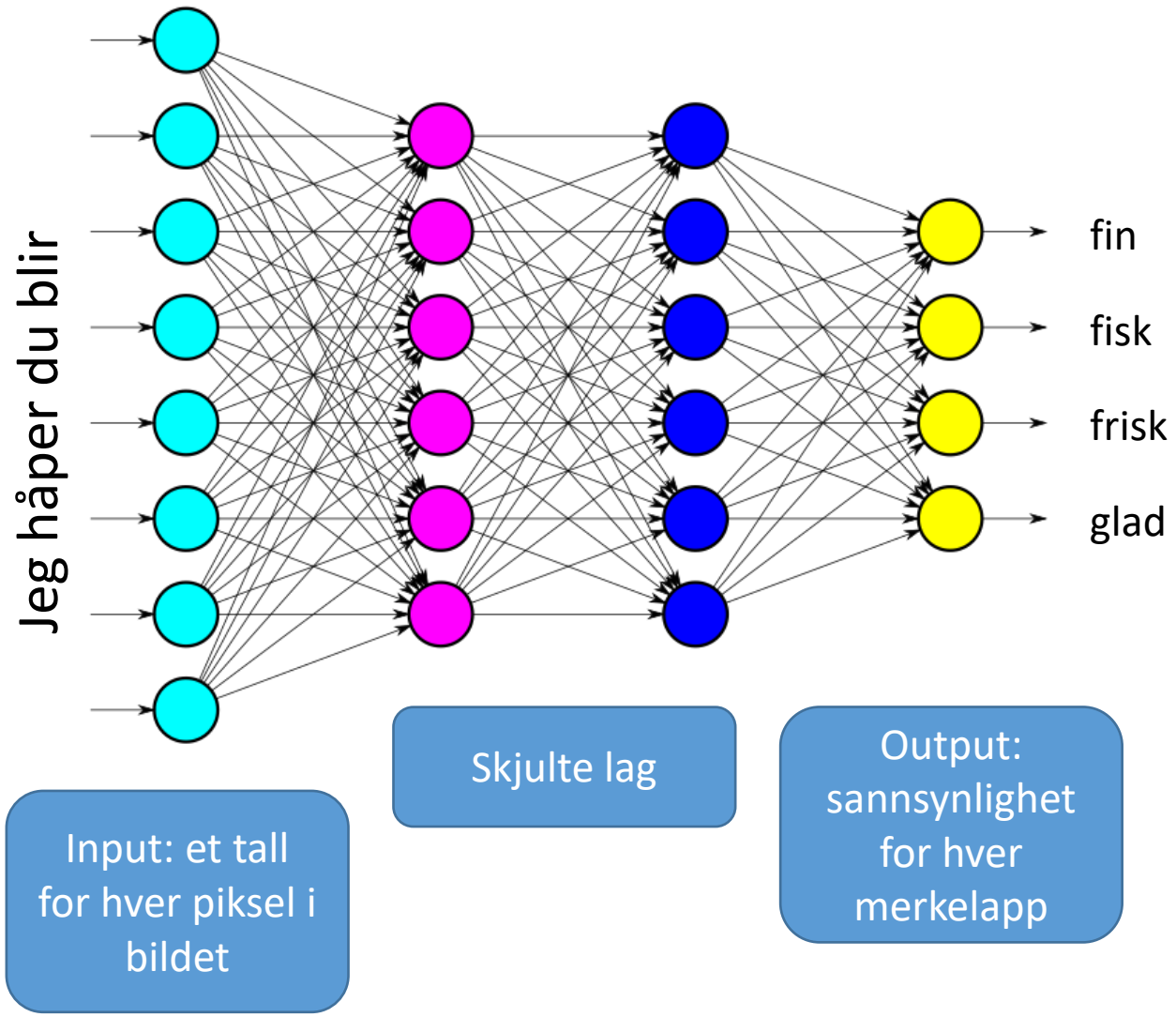
- 1940-tallet: start
- 1986: Viktig læringsalgoritme
- 2012: Billedklassifisering
  - Dype nettverk: mange lag
  - Feilrate 26%→3.6% på 4 år
  - Den nye AI-revolusjonen
- 2022: Tekst-til-bilderedigering
  - DALL-E-2: *Teddy bears shopping for groceries in ancient Egypt*
  - Midjourney

# Språkmodeller

- Talegjenkjenning:
  - *Jeg håper du snart blir …*
    - *fisk?*
    - *frisk?*
- Oversettelse:
  - *En rett linje*
  - *A corect? course? court? dish? law? plain? right? straight? line*
- Språkmodell:
  - sannsynlighetsfordeling over ordsekvenser
  - sannsynlighet for neste ord i en sekvens

# Nevral språkmodell

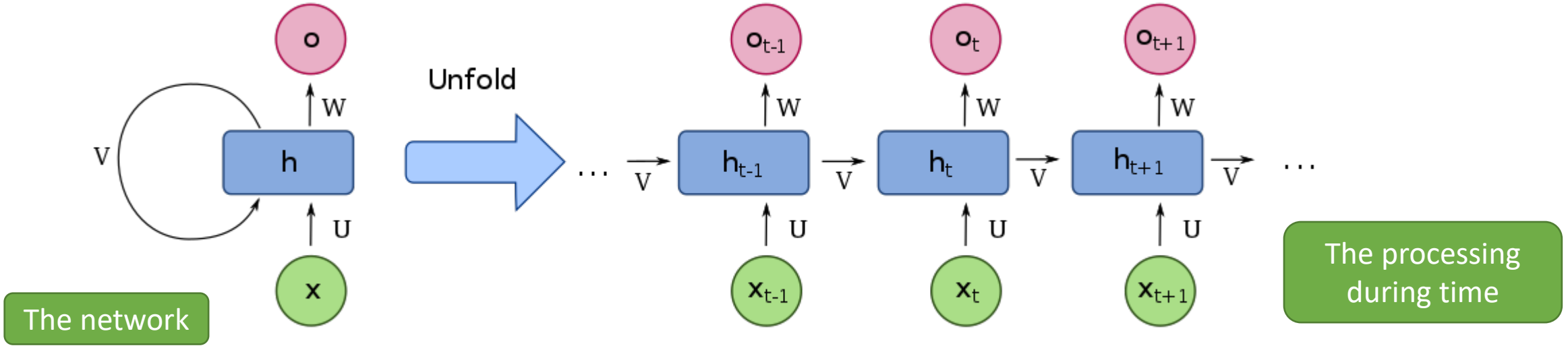
- Predikere neste ord
  - 50 000 forskjellige
- Input: representasjon av ordsekvens
- Et ord representeres av en tallvektor
  - (0.17, 0.48,..)
  - 50-300 dimensjoner
- Representasjonen læres av et nettverk
- Lignende ord får lignende representasjon
- "Selvovervåket"




network from https://openclipart.org/detail/290666/neural-net-graph

# Recurrent neural nets

- Model sequences/temporal phenomena
- A cell may send a signal back to itself – at the next moment in time




https://en.wikipedia.org/wiki/Recurrent_neural_network

# Tekstgenerering

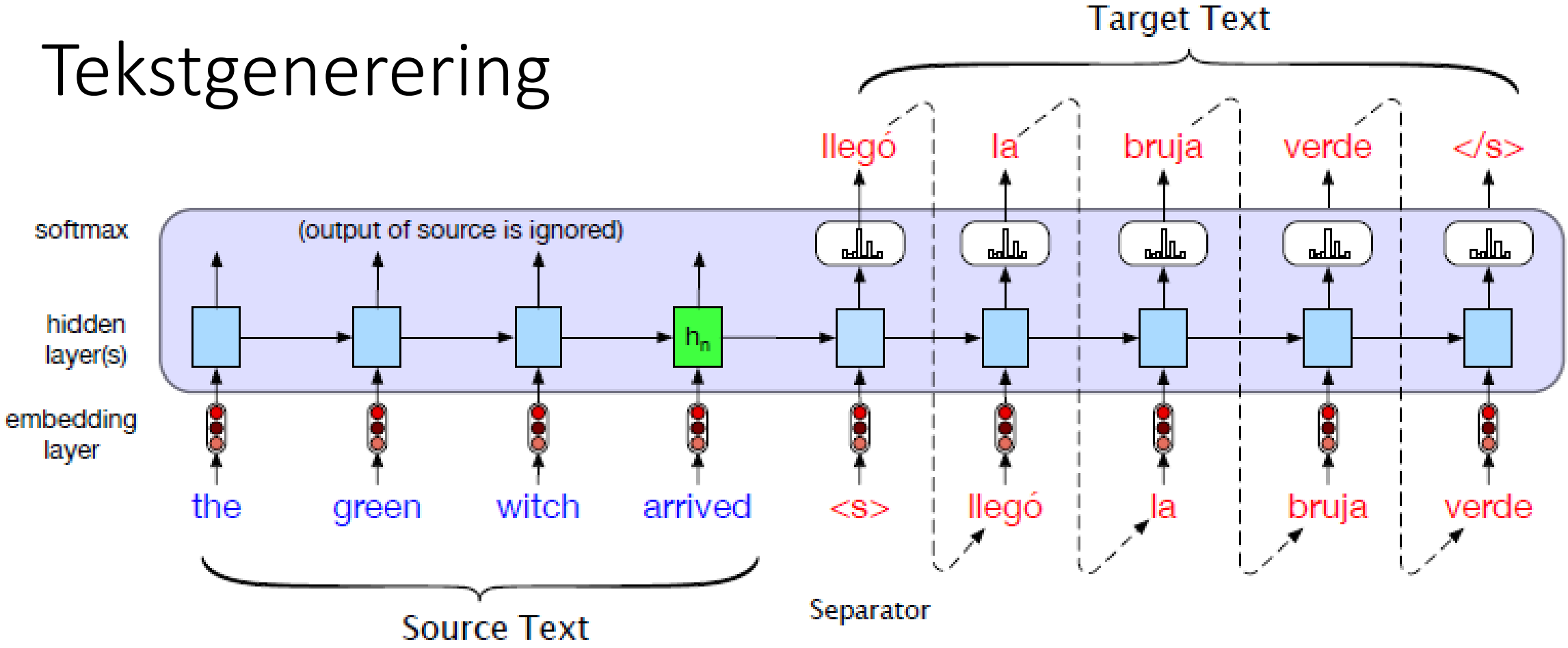
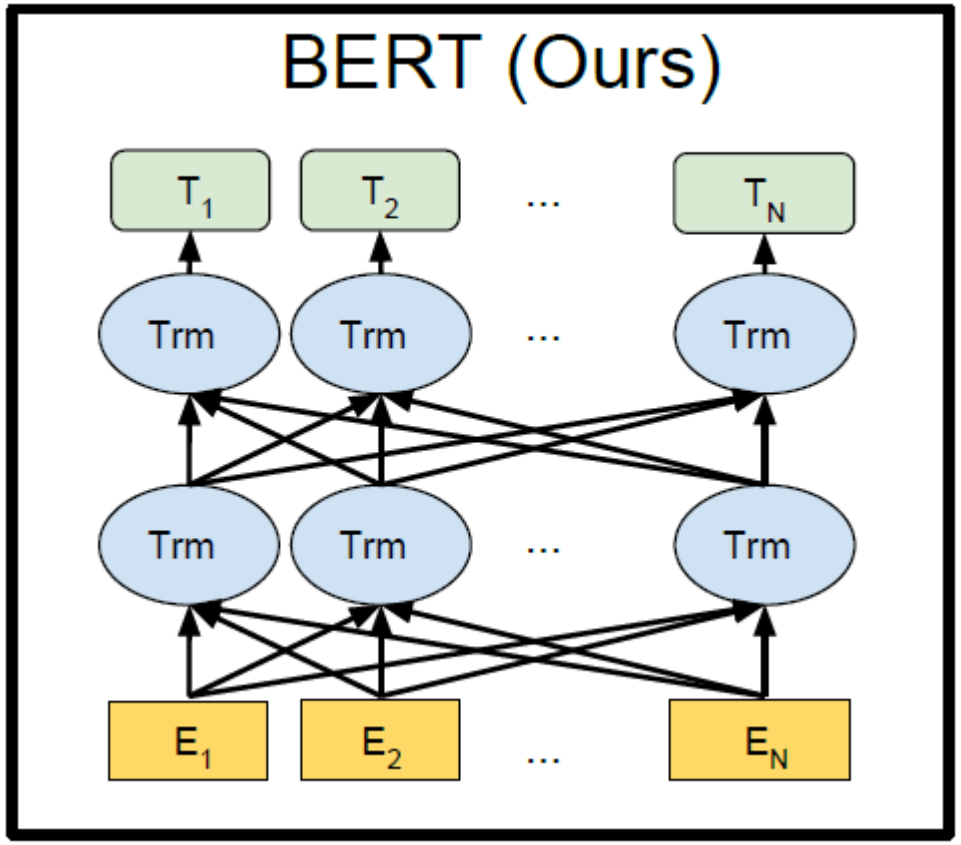



**Figure 10.4** Translating a single sentence (inference time) in the basic RNN version of encoder-decoder approach to machine translation. Source and target sentences are concatenated with a separator token in between, and the decoder uses context information from the encoder's last hidden state.

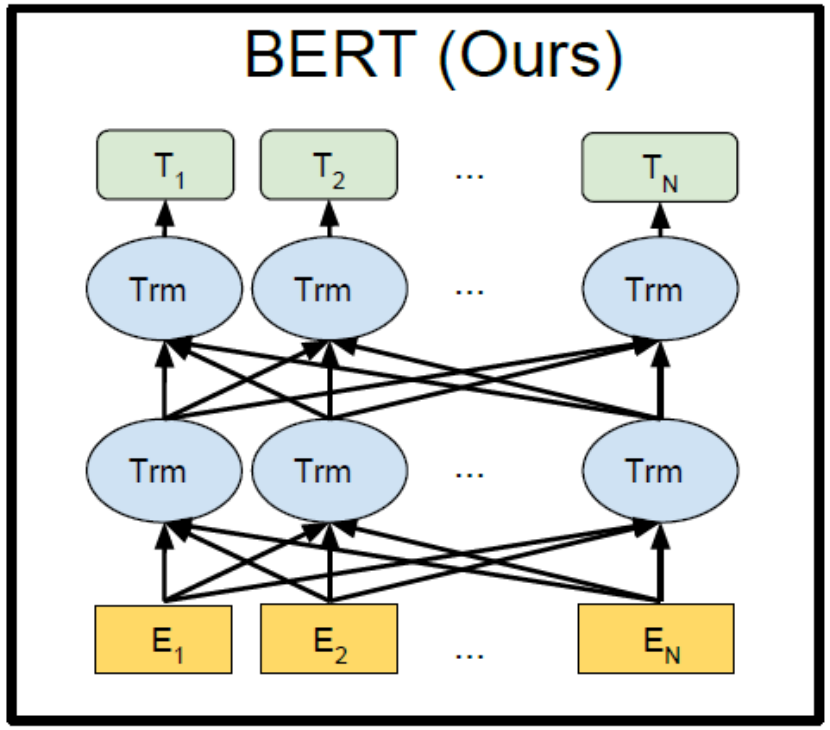
# Language transformers (2017)

- Representasjon av hele setninger
- En kontekstavhengig representasjon av hvert ord, e.g., *rett*
- Trenes på store mengder data

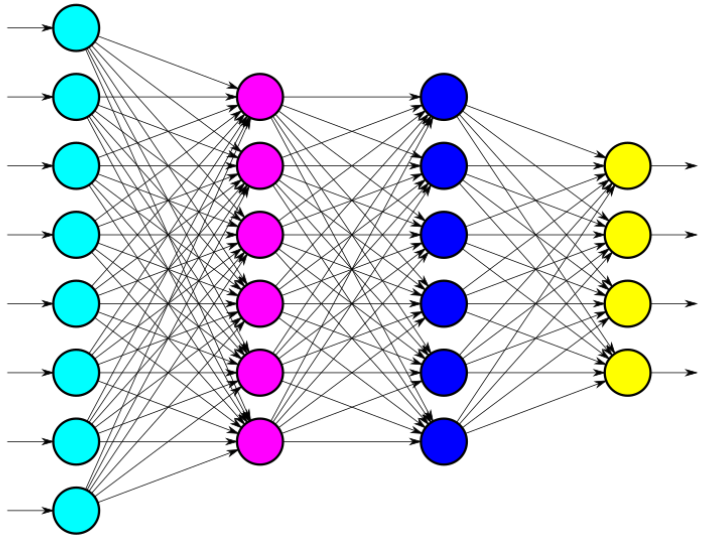



J. Devlin et. al. BERT, https://arxiv.org/pdf/1810.04805.pdf

# Tilpasning ("fine-tuning")




Input tekst

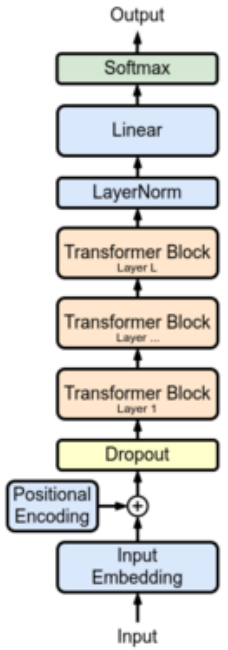


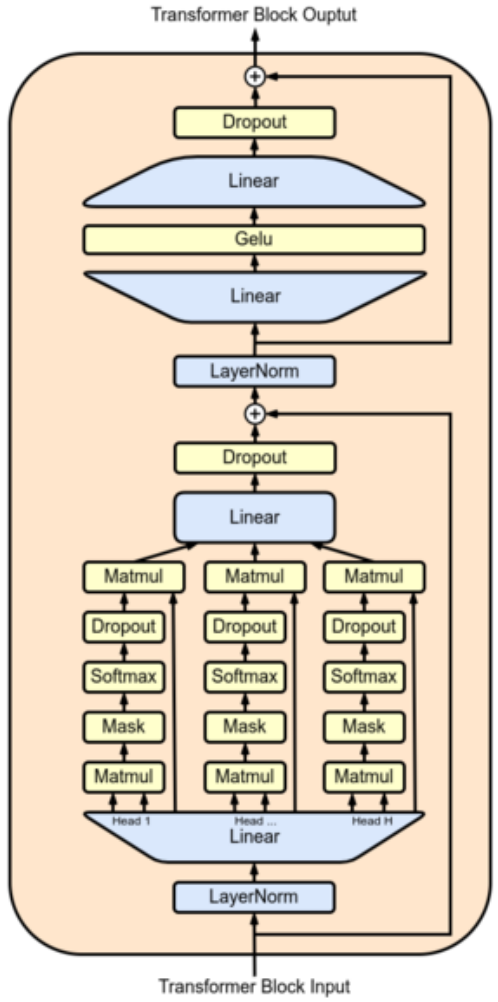
Output

- En fortrenet språkmodell kan gi input til en oppgavespesifikk modell
  - mye mindre mengder data
  - vektene i språkmodellen kan tilpasses ("fine-tune") til oppgaven.

# Generative Pre-trained Transformer

- Trent på å predikere fortsettelser
- GPT 2
- GPT 3
  - 570 GB tekst:
    - 400 000 000 000 ord
    - Web, bøker, wikipedia
  - 175 000 000 000 parametere (vekter)
- GPT 3.5
- GPT 4




https://en.wikipedia.org/wiki/GPT-2

# ChatGPT

1. GPT 3.5/4
2. Tilpasning ("fine-tuning") til spm.-svar-oppgaver
3. Forsterkende ("reinforcement") læring:
    1. Innsamling av brukeratferd
    2. Tilpasning av modellene
4. Filtrert for upassende innhold

# Begrensninger

- "Kunnskapene" er summen av alle tekster den er trent på
- Upresis:
  - Hallusinerer
  - Vet ikke hva den (ikke) vet
- Etiske utfordringer:
  - Formidler holdninger og meninger
  - Skjevhet ("bias")
  - Karbonavtrykk
  - Kontrolleres av store selskaper
    - ikke fritt tilgjengelig