KI-Campus
Die Lernplattform
für Künstliche Intelligenz

Deutsches
Forschungszentrum
für Künstliche
Intelligenz GmbH

TU berlin

GEFÖRDERT VOM
Bundesministerium
für Bildung
und Forschung

Transfer Learning

Salar Mohtaj | DFKI

Transfer learning

- What is transfer learning
- Motivation for transfer learning
- Different approaches for transfer learning
- BERT

Transfer learning

- What is transfer learning
- Motivation for transfer learning
- Different approaches for transfer learning
- BERT

What is transfer learning

- Humans have an ability to transfer knowledge across tasks
- What we acquire as knowledge while learning about a task, we utilize in the same way to solve related tasks
  - Knowledge about how to change setting in Windows → how to change setting in the other OSs
- In many cases we transfer and leverage our knowledge from what we have learnt in the past!

https://jeanvitor.com

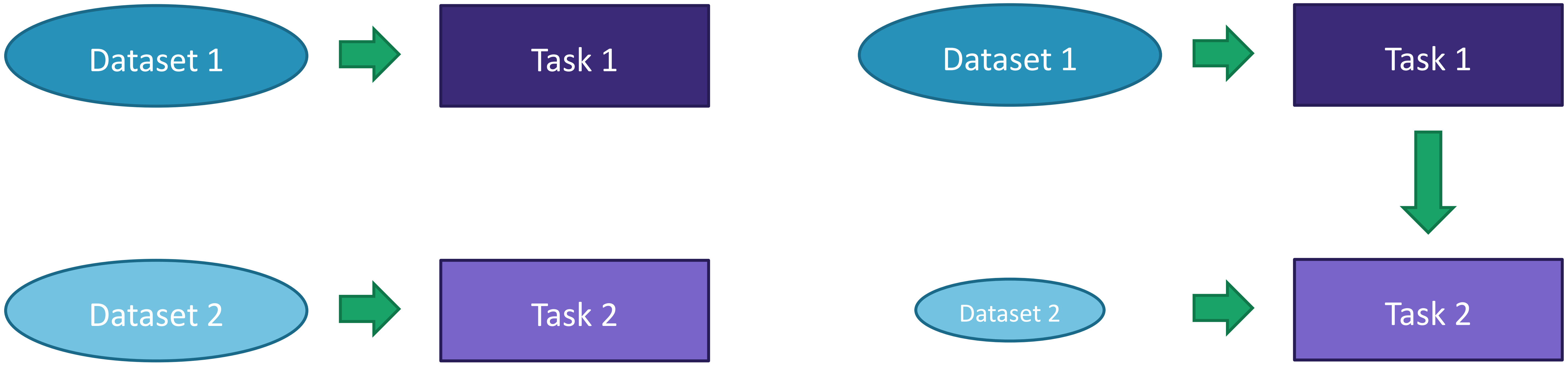
What is transfer learning

- Transfer learning is a learning procedure in which representations learned on a source task are **transmitted** to improve learning on the target task
- In transfer learning, the knowledge of an already trained **machine learning** model is applied to a different problem
- Transfer learning is a machine learning technique where a model trained on one task is re-purposed on a second related task
- With transfer learning, we basically try to exploit what has been learned in one task to improve generalization in another.
- We transfer the weights that a network has learned at a task (task A) to a new task (task B)

What is transfer learning

- Suppose we have a sentiment analysis task for the domain of customer reviews
- We have enough labelled data for this supervised task and train a model for it
- It would work well in different problems related to customer review
- But, as soon as we apply it for the same task in another domain, such as stock market, the performance would decrease
- The idea is to tune the model that is trained on the first domain, to work well also in the second domain

What is transfer learning



What is transfer learning

- As a formal definition:
  - Given a source domain ***Ds***, a corresponding source task ***Ts***, as well as a target domain ***Dt*** and a target task ***Tt***, the objective of transfer learning is to enable us to learn the target conditional probability distribution **P(Yt|Xt)** in ***Dt*** with the information gained from ***Ds*** and where **Ds ≠ Dt** or **Ts ≠ Tt**.

Transfer learning

- What is transfer learning
- Motivation for transfer learning
- Different approaches for transfer learning
- BERT

Motivation for transfer learning

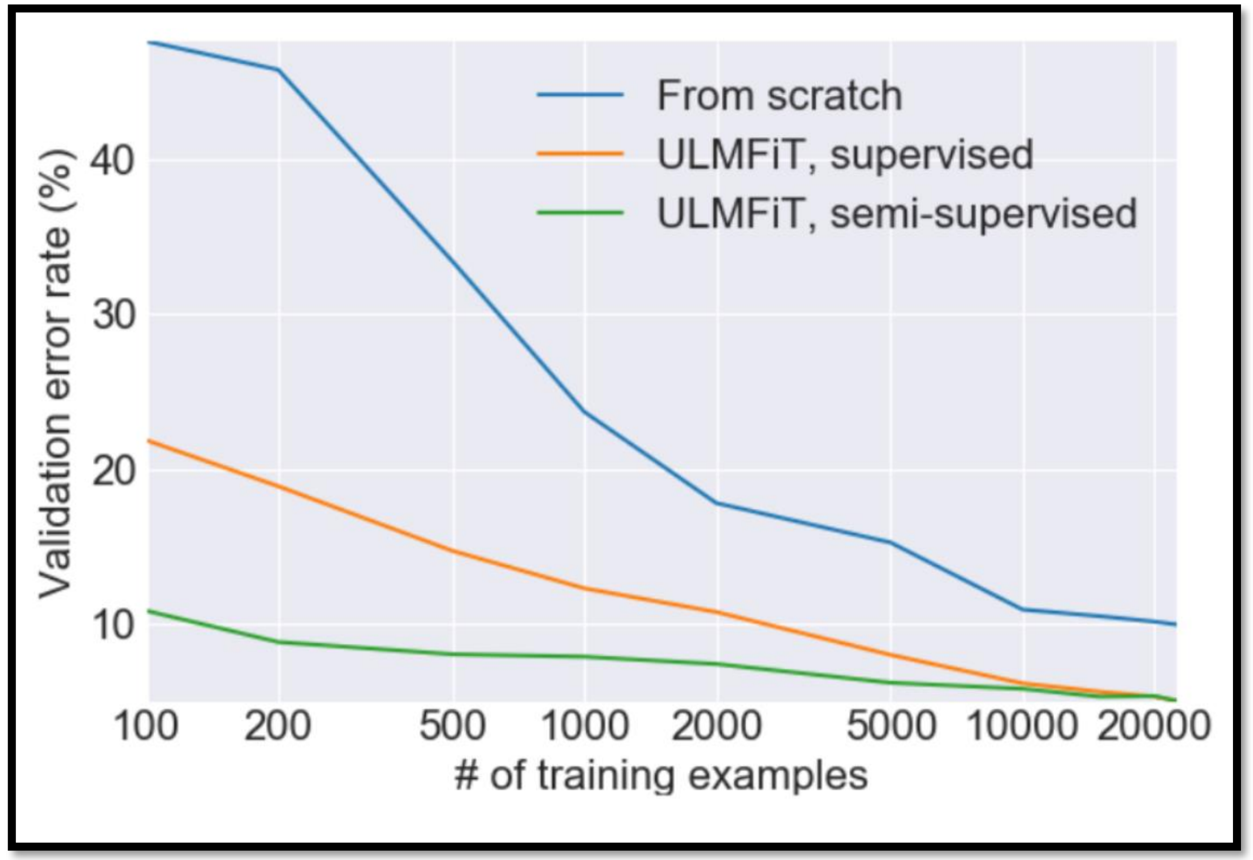
- Traditional learning is isolated and occurs purely based on specific tasks, datasets and training separate isolated models on them
  - No knowledge is retained which can be transferred from one model to another
- In transfer learning, you can leverage knowledge (e.g., features, weights) from previously trained models for training newer models
  - Tackle problems like having less data for the newer task!

Motivation for transfer learning

- Transfer learning has several benefits, but the main advantages are:
- Saving training time
- Better performance
- Not needing a lot of data

Motivation for transfer learning

- One of the main benefits of pretraining is that it reduces the need for annotated data
  - Transfer learning based models could achieve similar performance compared to a simple model with 10x fewer examples (Howard and Ruder, 2018).



Howard, Jeremy, and Sebastian Ruder. "Universal language model fine-tuning for text classification." *arXiv preprint arXiv:1801.06146* (2018)

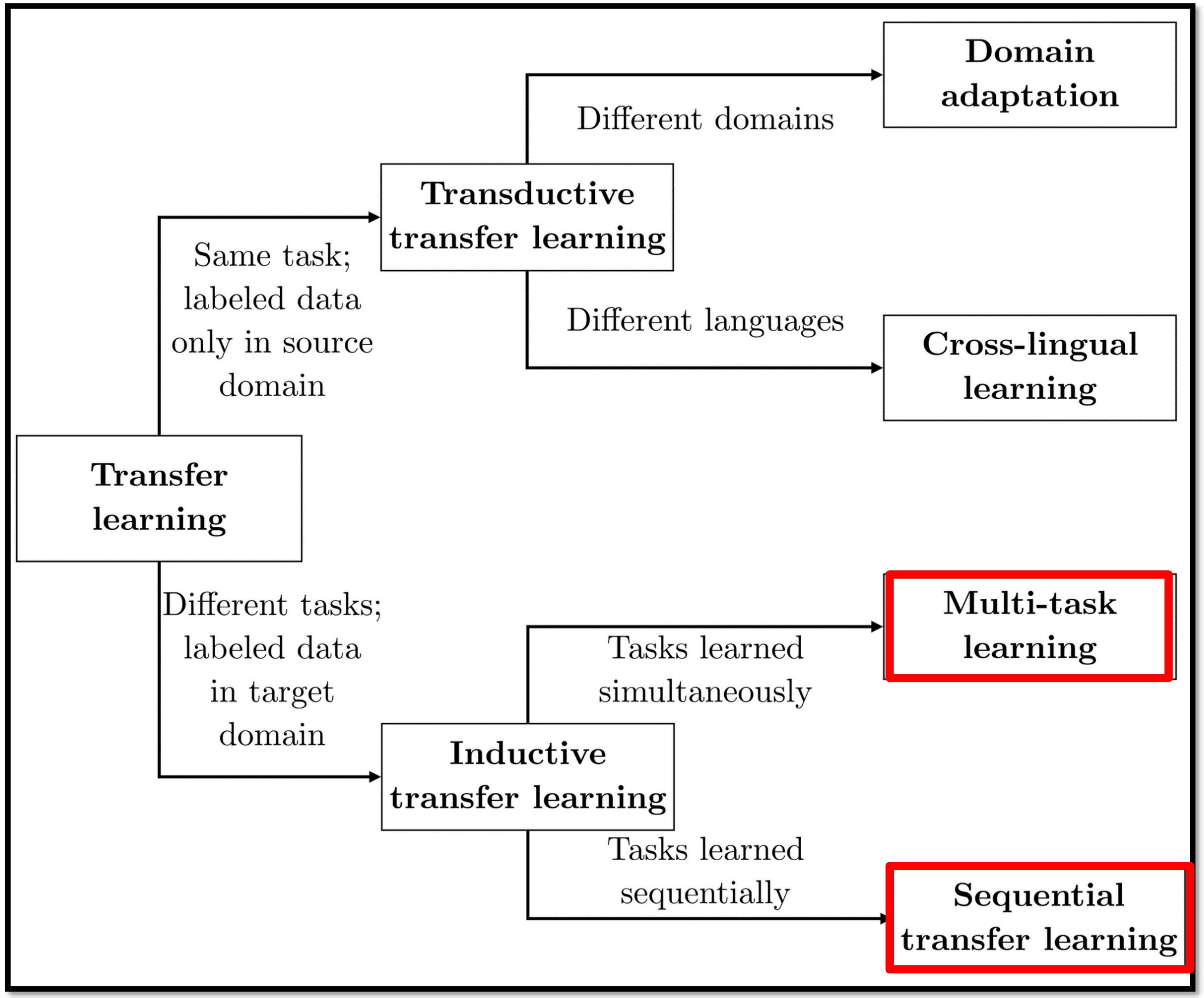
Motivation for transfer learning

- Most of the labeled text datasets are not big enough to train deep neural networks because these networks have a huge number of parameters and training such networks on small datasets will cause overfitting

Transfer learning

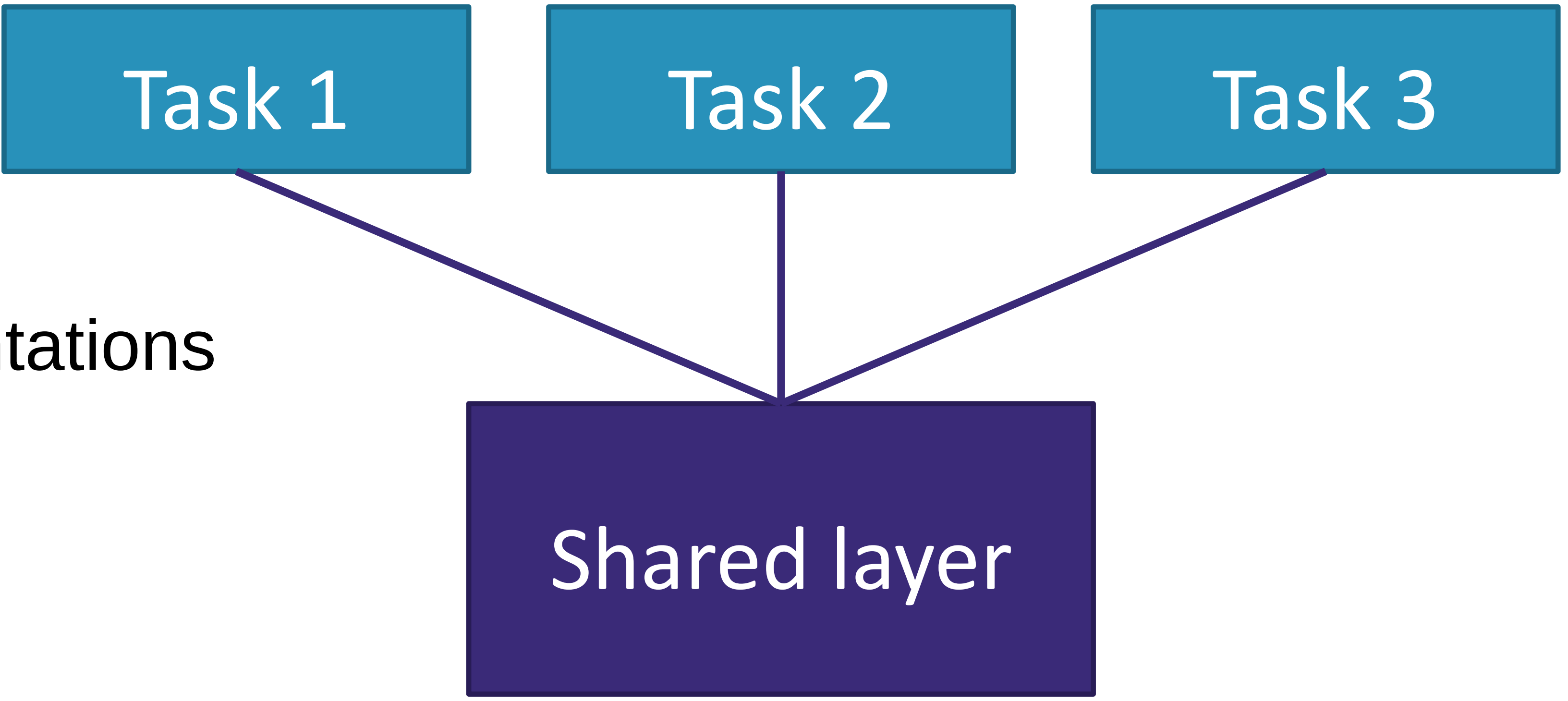
- What is transfer learning
- Motivation for transfer learning
- Different approaches for transfer learning
- BERT

Different approaches for transfer learning


Domain adaptation
Different domains
Transductive transfer learning
Same task; labeled data only in source domain
Different languages
Cross-lingual learning
Transfer learning
Multi-task learning
Different tasks; labeled data in target domain
Tasks learned simultaneously
Inductive transfer learning
Tasks learned sequentially
Sequential transfer learning

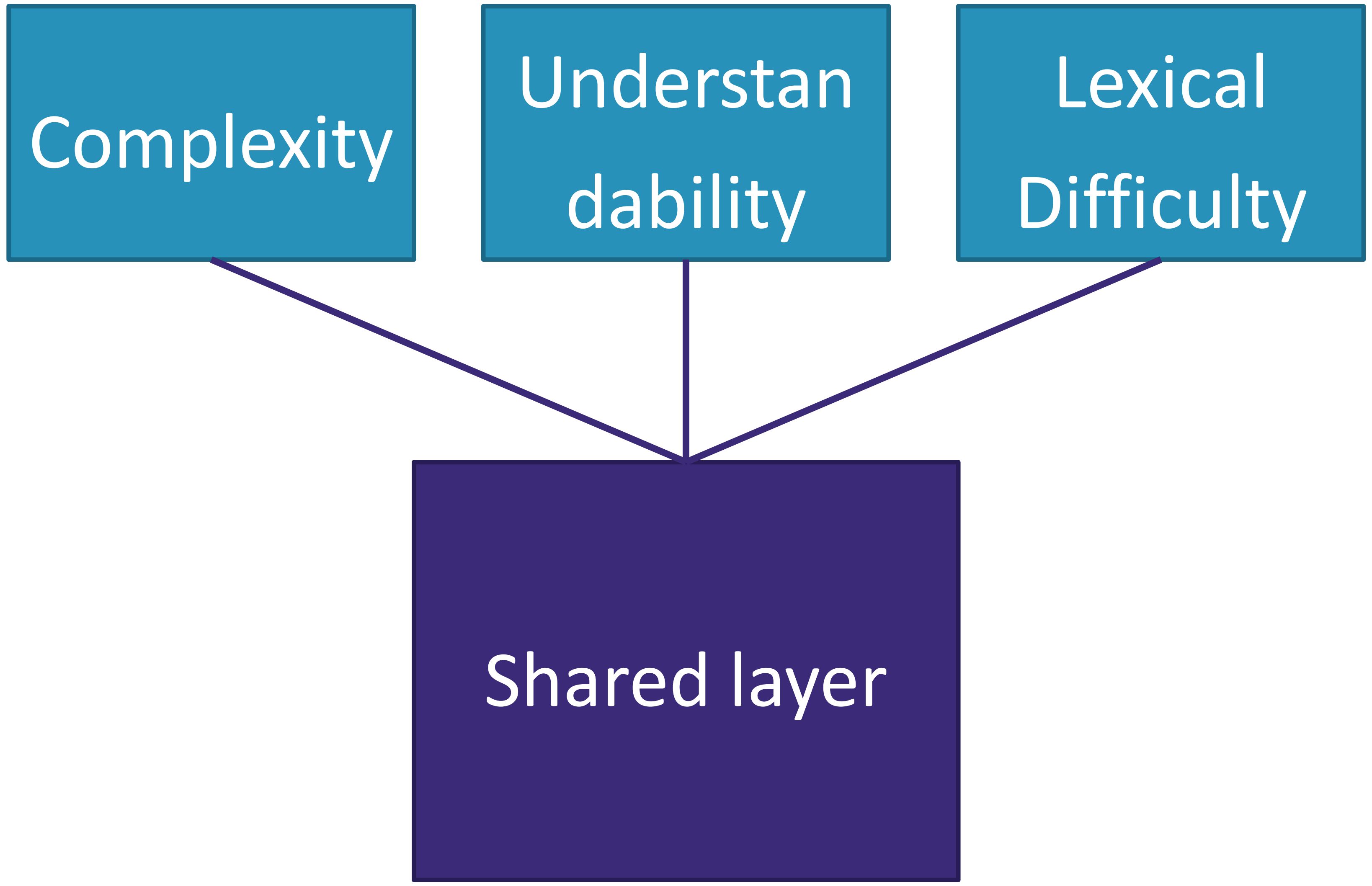
Different approaches for transfer learning

- Multi-task learning
  - Multi-task learning (MTL) is a subfield of machine learning in which multiple tasks are simultaneously learned by a shared model
- Such approaches offer advantages like:
  - Improved data efficiency
  - Reduced overfitting through shared representations
  - fast learning



Different approaches for transfer learning

- Subjective assessment of German text complexity



Different approaches for transfer learning

- Different approaches for sequential transfer learning
  - Training a model to reuse it
  - Fine-tuning a pre-trained model
  - Feature extraction

Different approaches for transfer learning

- Training a model to reuse it
  - Imagine you want to solve the task of sentiment analysis for a specific domain
  - But you don’t have enough data to train a deep neural network
  - One way around this is to find a related task B with an abundance of data
  - Train a model on task B and use the model as a starting point for solving the task of sentiment analysis for the target domain
  - The trained model can be fine-tuned on the task B

Different approaches for transfer learning

- Fine-tuning a pre-trained model
  - There are a lot of pre-trained models for NLP which are trained for different tasks
  - One popular approach for transfer learning would be using one of these pre-trained models
  - The model can fine-tuned based on a small data which is available for the target task

Different approaches for transfer learning

- Feature extraction
  - Pre-trained models may be used as feature extraction models
  - Here, the output of the model is used as input to a new classifier model
  - Here there is no fine-tuning of the weights

Different approaches for transfer learning

- Using a pre-trained model
  - Empirically, language modelling works better than other pretraining tasks
- Language modelling is a very difficult task, even for humans
- To have any chance at solving this task, a model is required to learn about syntax, semantics, as well as certain facts about the world
- Given enough data, a large number of parameters, and enough compute, a model can do a reasonable job

https://ruder.io/state-of-transfer-learning-in-nlp/

Different approaches for transfer learning

- Advantages of language modelling is that
  - It does not require any human annotation
  - Many languages have enough text available to learn reasonable models
  - Language model is enable to learn both sentence and word representations

Different approaches for transfer learning

- Using a pre-trained model
  - **BERT**
  - GPT-3
  - ELMo
  - XLNet
  - ALBERT
  - ULMFiT
  - RoBERTa

Different approaches for transfer learning

- Important questions on using pre-trained models:
- What to transfer
  - Which part of the knowledge can be transferred from the source to the target
  - Which portion of knowledge is source-specific and what is common between the source and the target
- When to transfer
- How to transfer

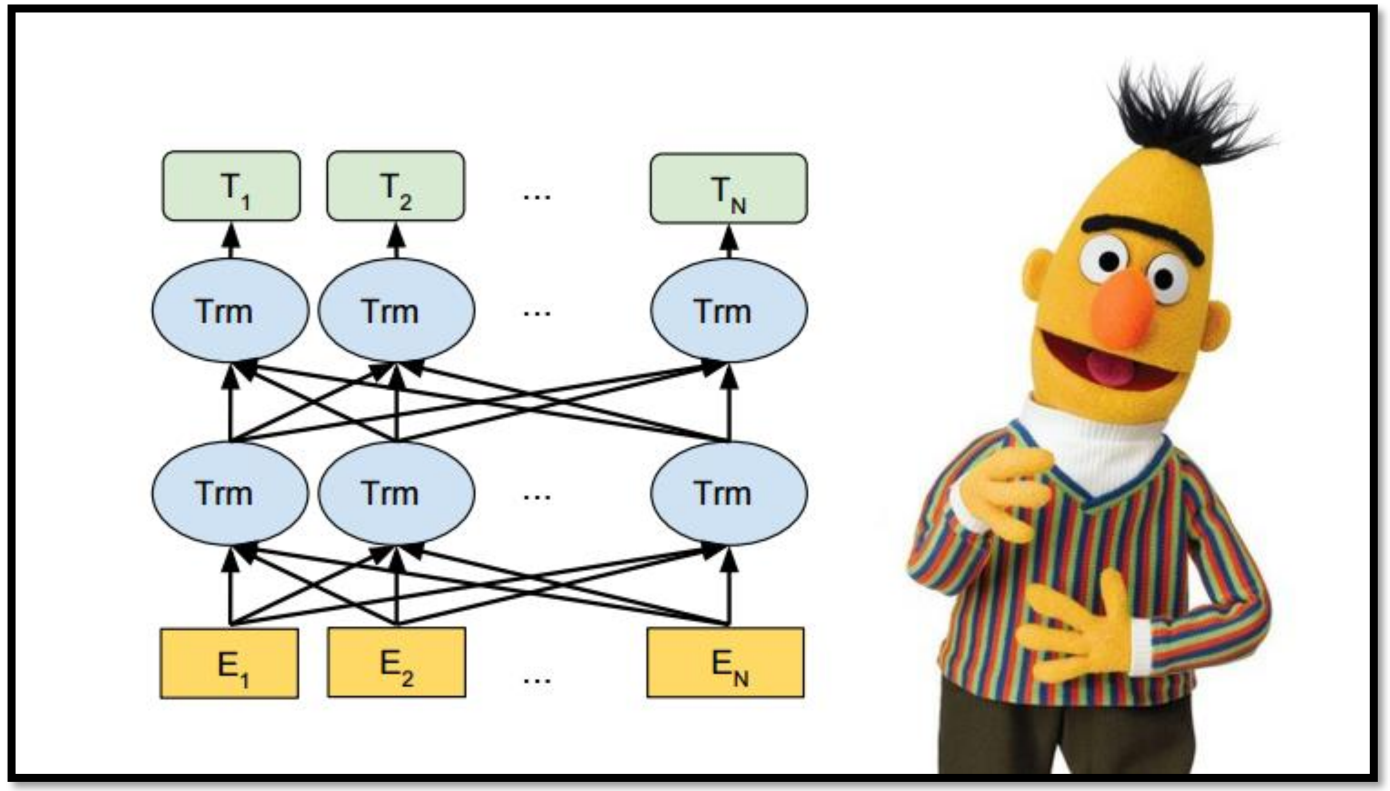
https://ruder.io/state-of-transfer-learning-in-nlp/

Transfer learning

- What is transfer learning
- Motivation for transfer learning
- Different approaches for transfer learning
- BERT

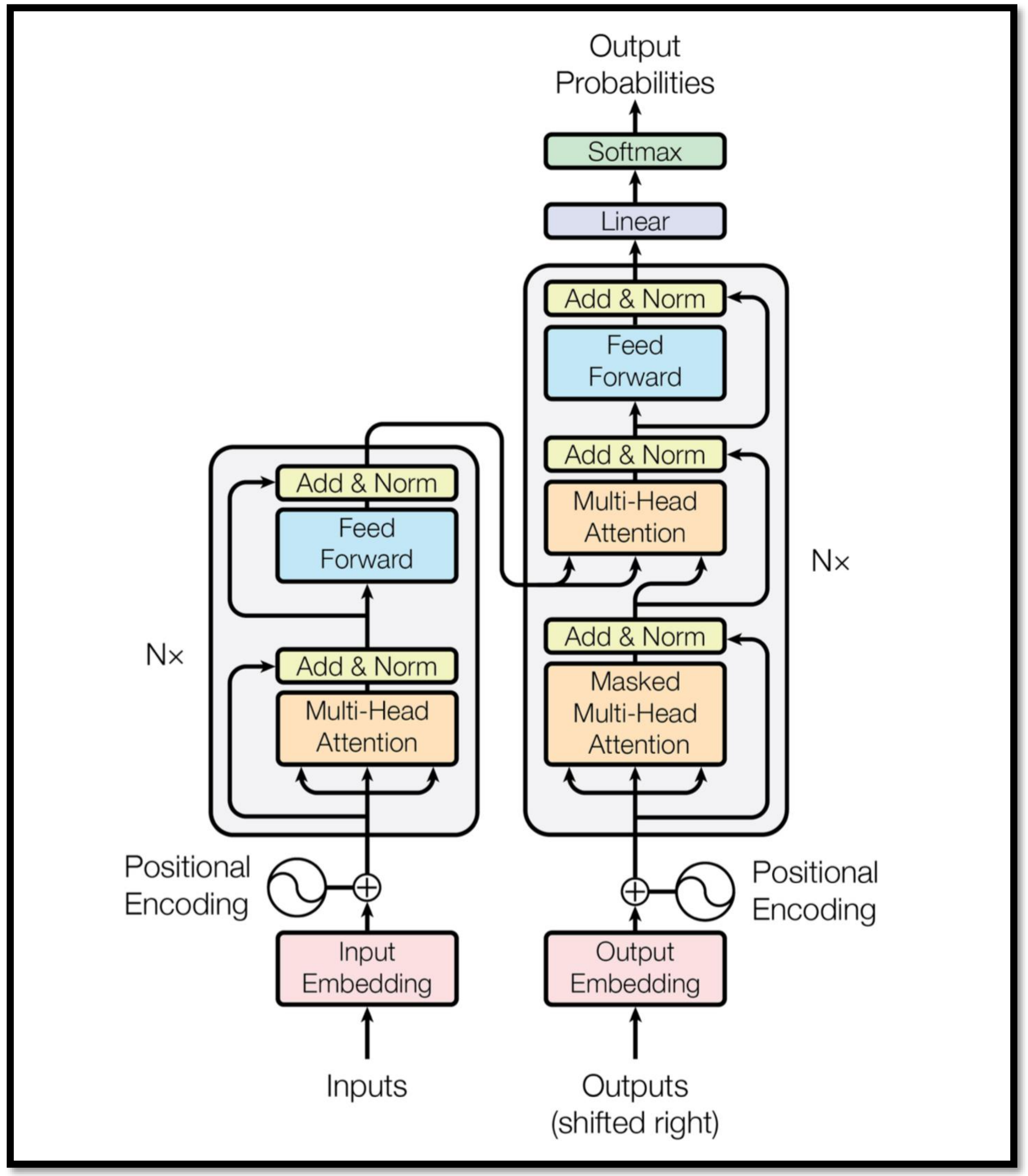
BERT

- BERT (Bidirectional Encoder Representations from Transformers)
  - A language model which is bidirectionally trained
  - Have a deeper sense of language context and flow compared to the single-direction language models



BERT

- The transformer
  - It a new architecture that uses the attention-mechanism
  - Like LSTM, Transformer is an seq2seq architecture but it differs because it does not imply any Recurrent Networks (GRU, LSTM, etc.)
  - The architecture which is only with attention-mechanisms without any RNN (Recurrent Neural Networks) improved the results in many NLP tasks includes translation task



Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *arXiv preprint arXiv:1706.03762* (2017).

BERT

- Instead of predicting the next word in a sequence, BERT makes use of a novel technique called masked LM
    - From each input sequence 15% of the tokens are processed as follows:
        - with 0.8 probability the token is replaced by [MASK]
        - with 0.1 probability the token is replaced by another random token
        - with 0.1 probability the token is unchanged

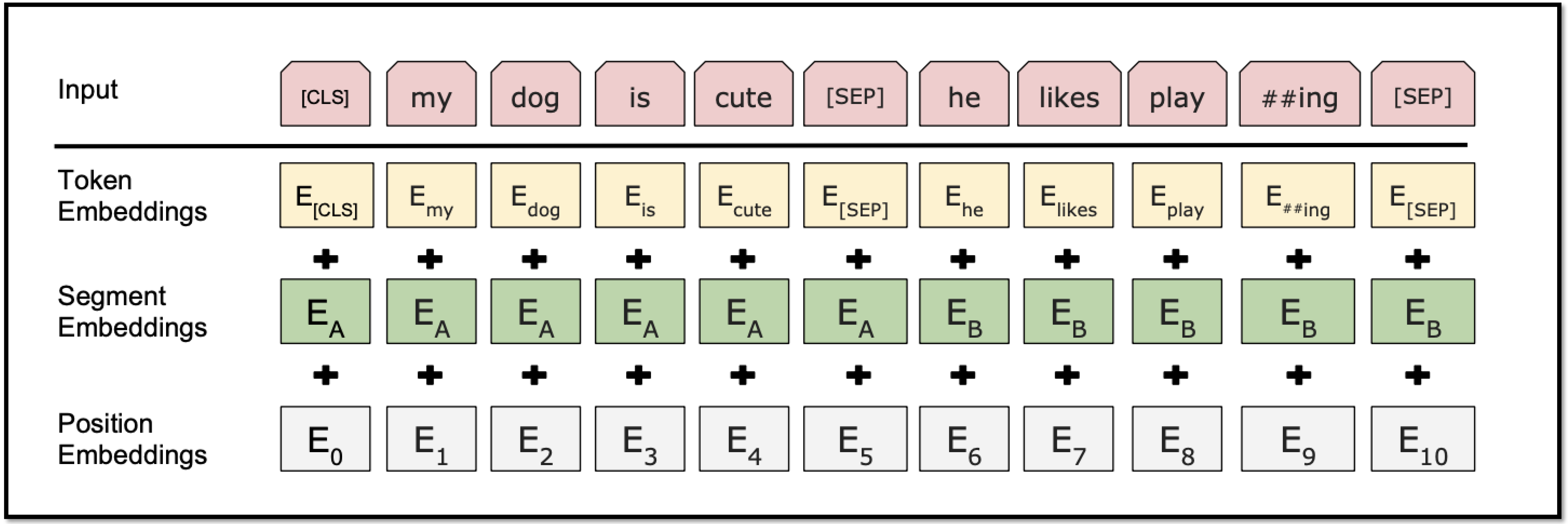
BERT

- The input is composed of two sentences
- These two sentences A and B are separated with the special token [SEP]
- 50% of the time B is the actual next sentence and 50% of the time is a random sentence

Input= [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
Label=IsNext
Input=[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]
Label=NotNext

Input= [CLS] the man went to [MASK] store [SEP] he bought a gallon [MASK] milk [SEP]
Label=IsNext
Input=[CLS] the man [MASK] to the store [SEP] penguin [MASK] are flight ##less birds [SEP]
Label=NotNext

BERT

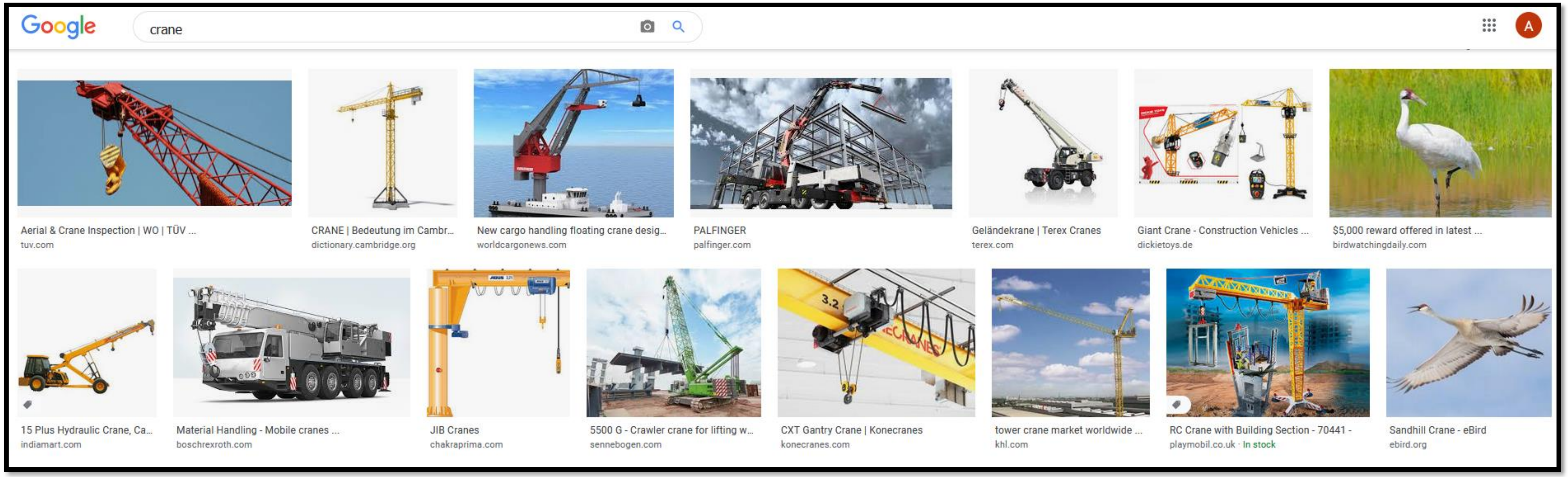


Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}

BERT

- Context based embedding
  - Unlike context-free models which generate a single vector (word embedding) representation for each token
  - Context based embeddings like BERT generate word vectors based on the context

Crane



Aerial & Crane Inspection | WO | TÜV ...
tuv.com

CRANE | Bedeutung im Cambr...
dictionary.cambridge.org

New cargo handling floating crane desig...
worldcargonews.com

PALFINGER
palfinger.com

Geländekrane | Terex Cranes
terex.com

Giant Crane - Construction Vehicles ...
dickietoys.de

$5,000 reward offered in latest ...
birdwatchingdaily.com

15 Plus Hydraulic Crane, Ca...
indiamart.com

Material Handling - Mobile cranes ...
boschrexroth.com

JIB Cranes
chakraprima.com

5500 G - Crawler crane for lifting w...
sennebogen.com

CXT Gantry Crane | Konecranes
konecranes.com

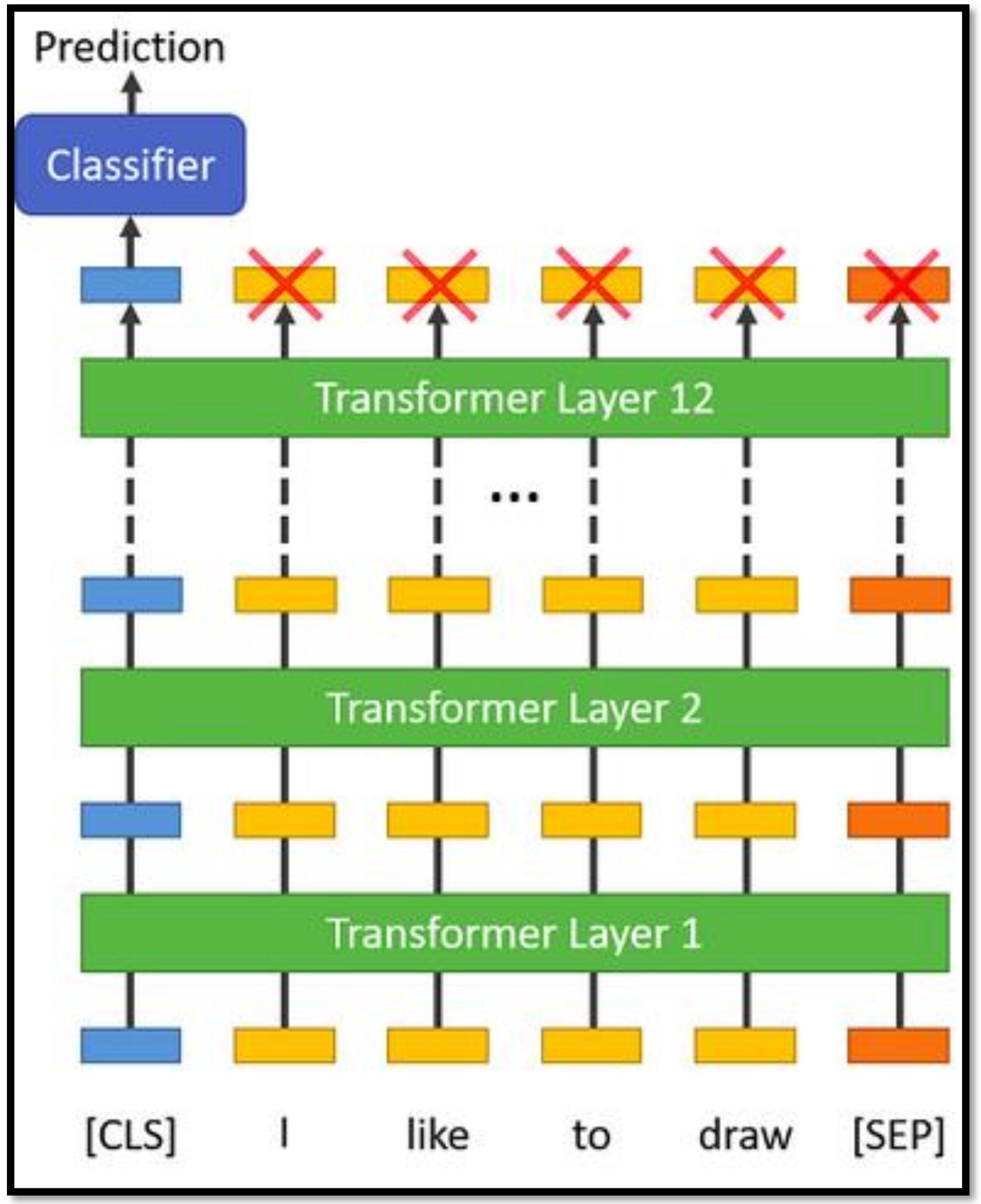
tower crane market worldwide ...
khl.com

RC Crane with Building Section - 70441 -
playmobil.co.uk · In stock

Sandhill Crane - eBird
ebird.org

BERT

- BERT-Base
  - 12-layer
  - 768-hidden-nodes
  - 12-attention-heads
  - 110M parameters
- BERT-Large
  - 24-layer
  - 1024-hidden-nodes
  - 16-attention-heads
  - 340M parameters



BERT

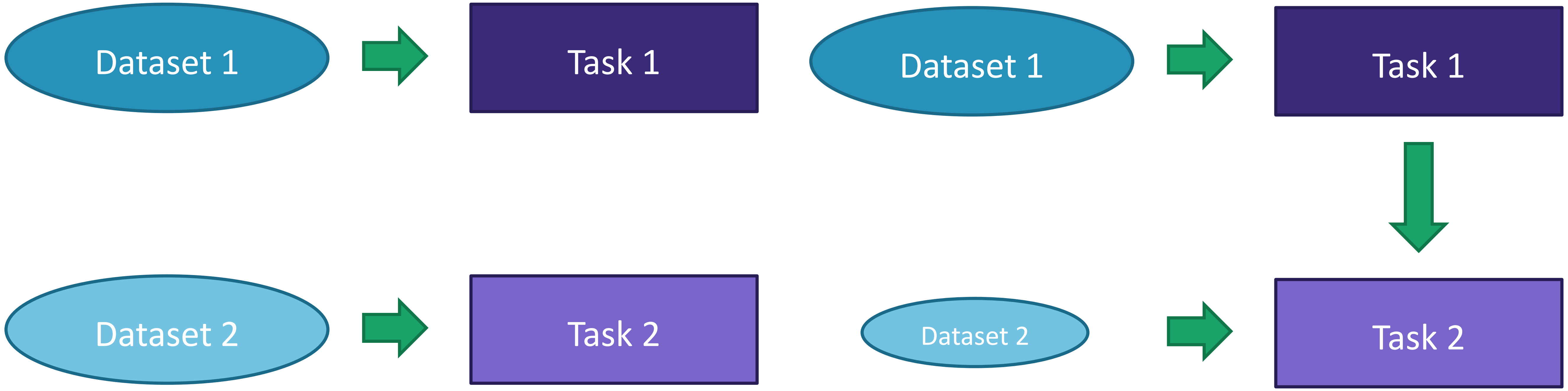
- Updating the weights
  - Not tune (feature extraction)
    - Pretrained representations can be used as features in a downstream model
  - Tune (fine-tuning)
    - The pretrained weights are used as initialization for parameters of the downstream model
    - The whole pretrained architecture is then trained during the adaptation phase

https://ruder.io/state-of-transfer-learning-in-nlp/

Summary

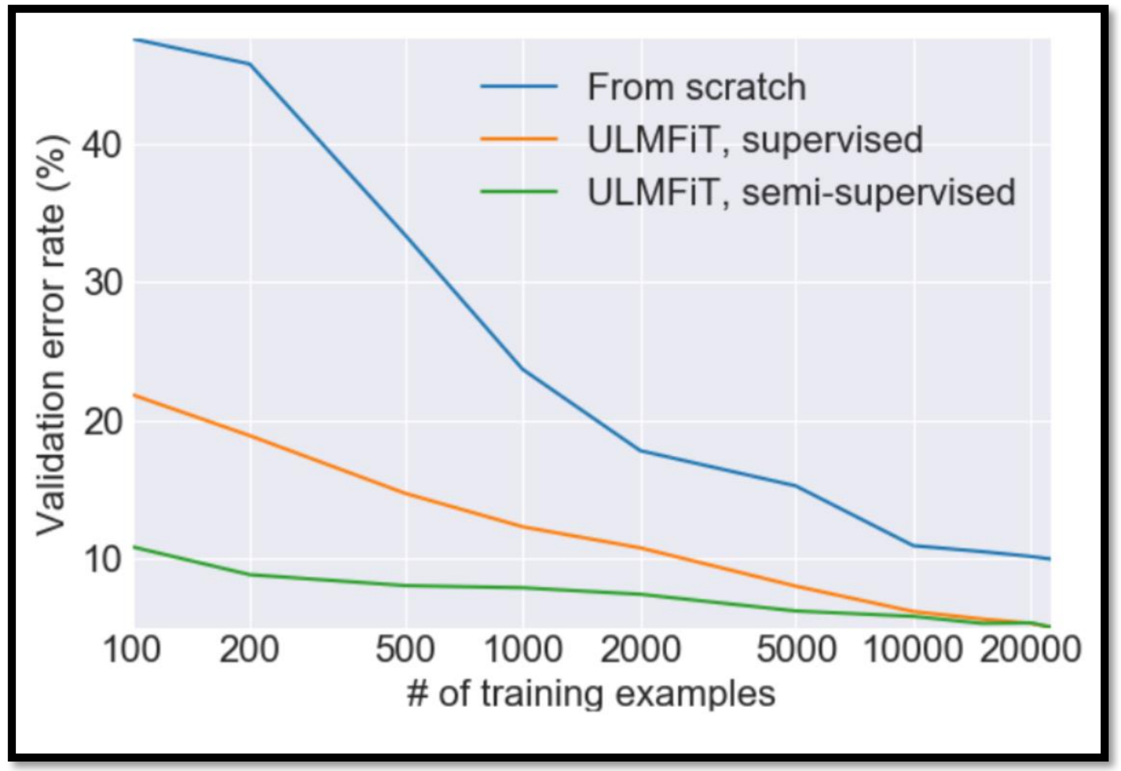
- Transfer learning is a learning procedure in which representations learned on a source task are **transmitted** to improve learning on the target task

https://jeanvitor.com

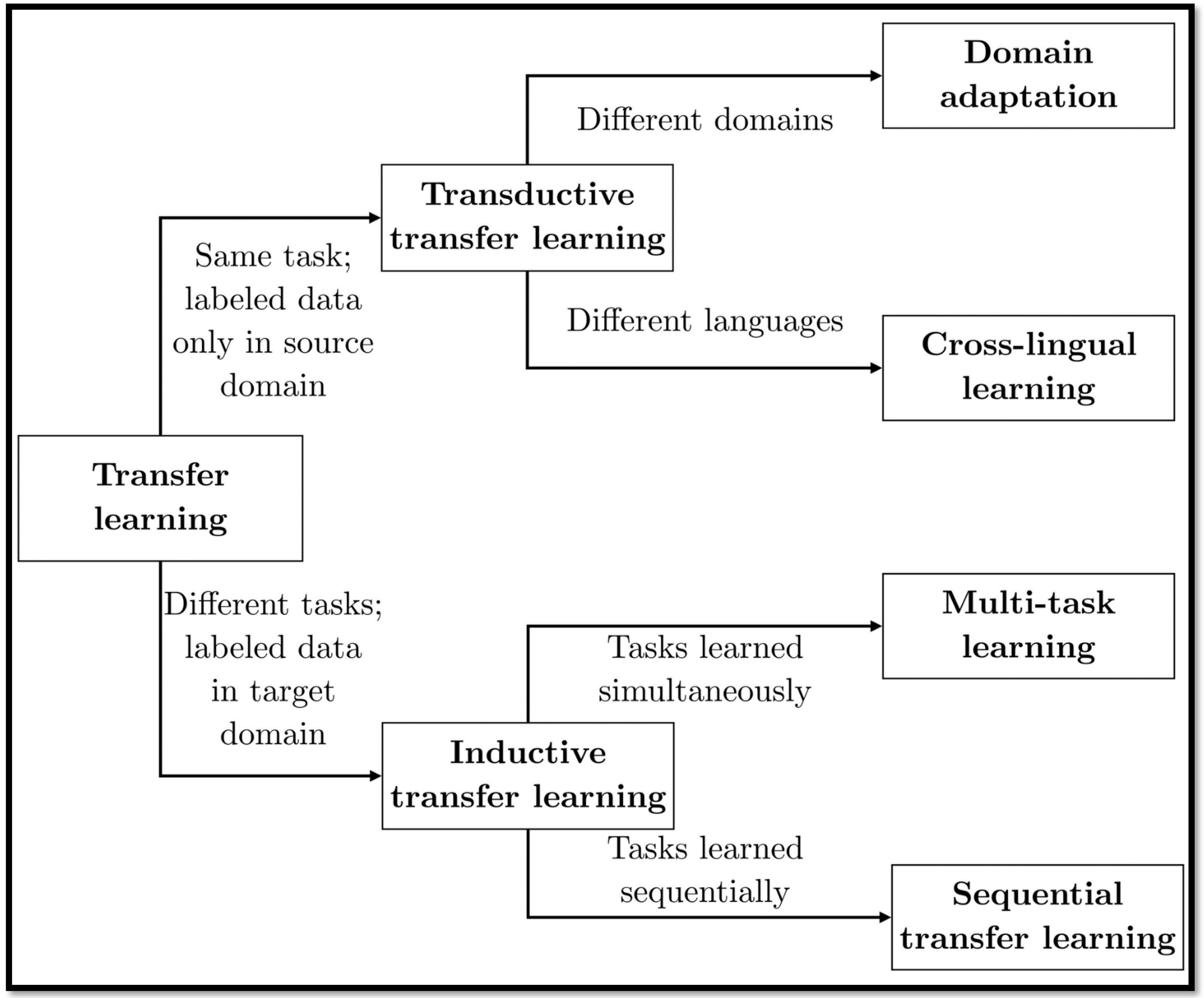


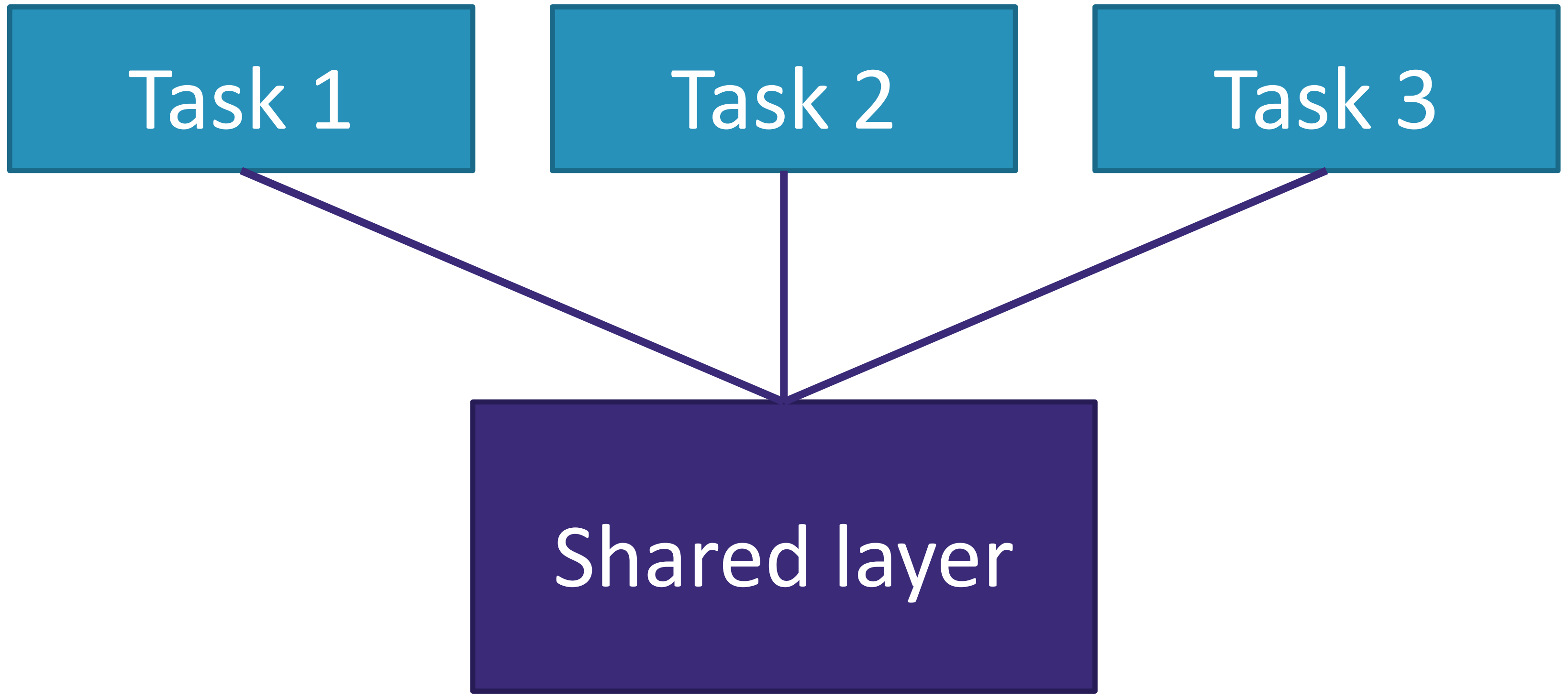
Summary

- Transfer learning has several benefits, but the main advantages are:
  - Saving training time
  - Better performance
  - Not needing a lot of data



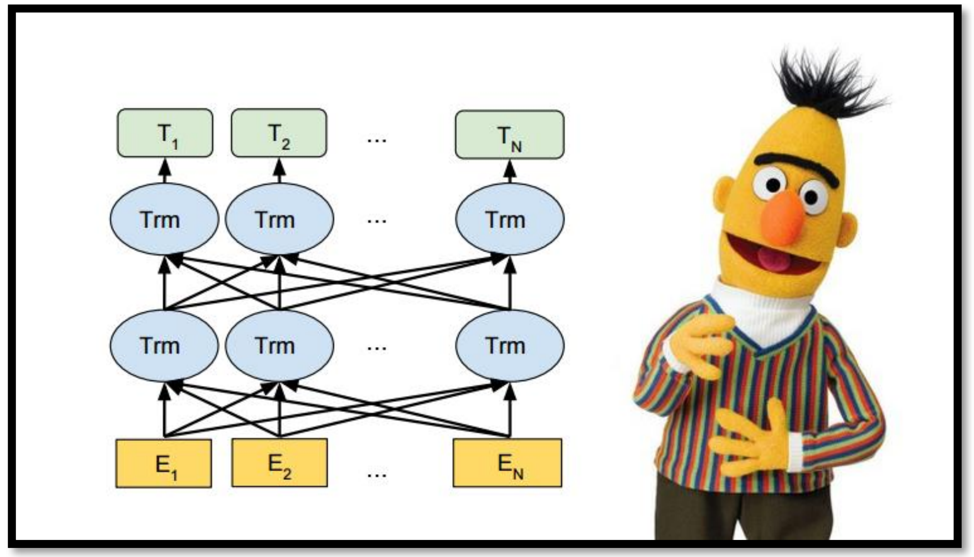
Summary

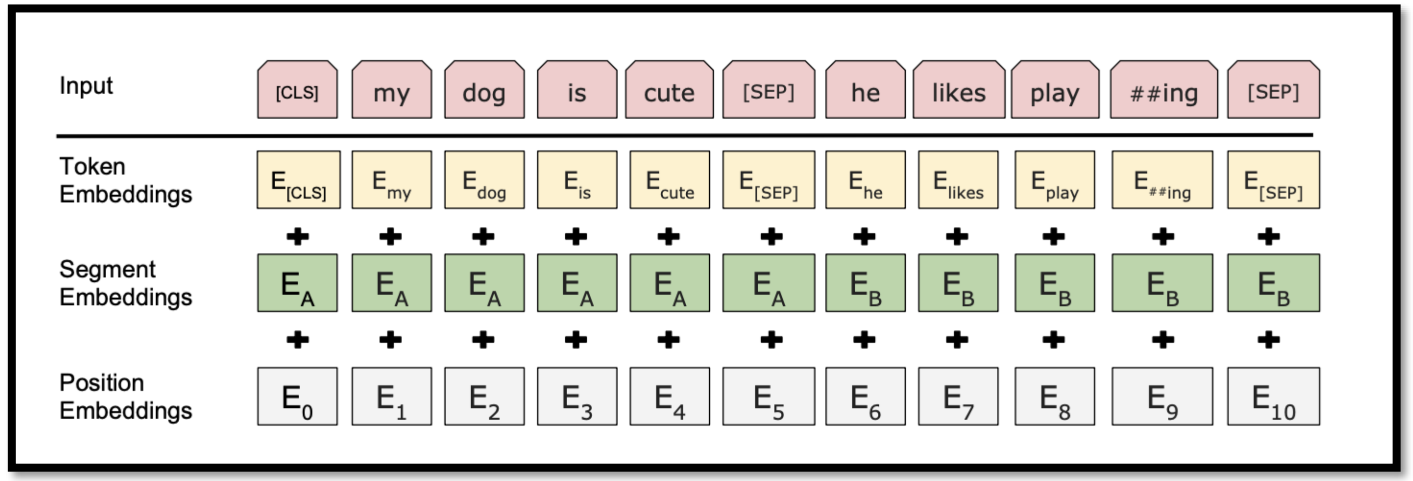

Domain adaptation
Different domains
Transductive transfer learning
Same task; labeled data only in source domain
Different languages
Cross-lingual learning
Transfer learning
Different tasks; labeled data in target domain
Multi-task learning
Tasks learned simultaneously
Inductive transfer learning
Tasks learned sequentially
Sequential transfer learning


Task 1
Task 2
Task 3
Shared layer

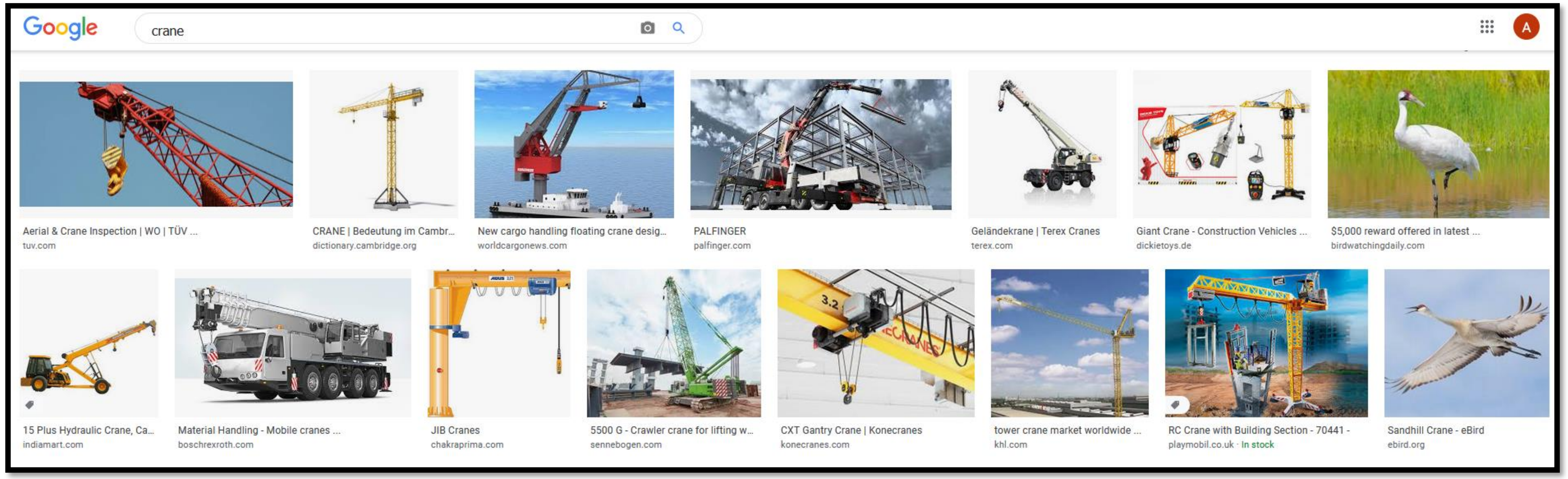
Summary

- Different approaches for sequential transfer learning
  - Training a model to reuse it
  - Fine-tuning a pre-trained model
  - Feature extraction





Input	[CLS]	my	dog	is	cute	[SEP]	he	likes	play	##ing	[SEP]
Token Embeddings	$E_{[CLS]}$	E_{my}	E_{dog}	E_{is}	E_{cute}	$E_{[SEP]}$	E_{he}	E_{likes}	E_{play}	$E_{\#\#ing}$	$E_{[SEP]}$
	+	+	+	+	+	+	+	+	+	+	+
Segment Embeddings	E_A	E_A	E_A	E_A	E_A	E_A	E_B	E_B	E_B	E_B	E_B
	+	+	+	+	+	+	+	+	+	+	+
Position Embeddings	E_0	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}



Aerial & Crane Inspection | WO | TÜV ...
tuv.com

CRANE | Bedeutung im Cambr...
dictionary.cambridge.org

New cargo handling floating crane desig...
worldcargonews.com

PALFINGER
palfinger.com

Geländekrane | Terex Cranes
terex.com

Giant Crane - Construction Vehicles ...
dickietoys.de

$5,000 reward offered in latest ...
birdwatchingdaily.com

15 Plus Hydraulic Crane, Ca...
indiamart.com

Material Handling - Mobile cranes ...
boschrexroth.com

JIB Cranes
chakraprima.com

5500 G - Crawler crane for lifting w...
sennebogen.com

CXT Gantry Crane | Konecranes
konecranes.com

tower crane market worldwide ...
khl.com

RC Crane with Building Section - 70441 -
playmobil.co.uk · In stock

Sandhill Crane - eBird
ebird.org