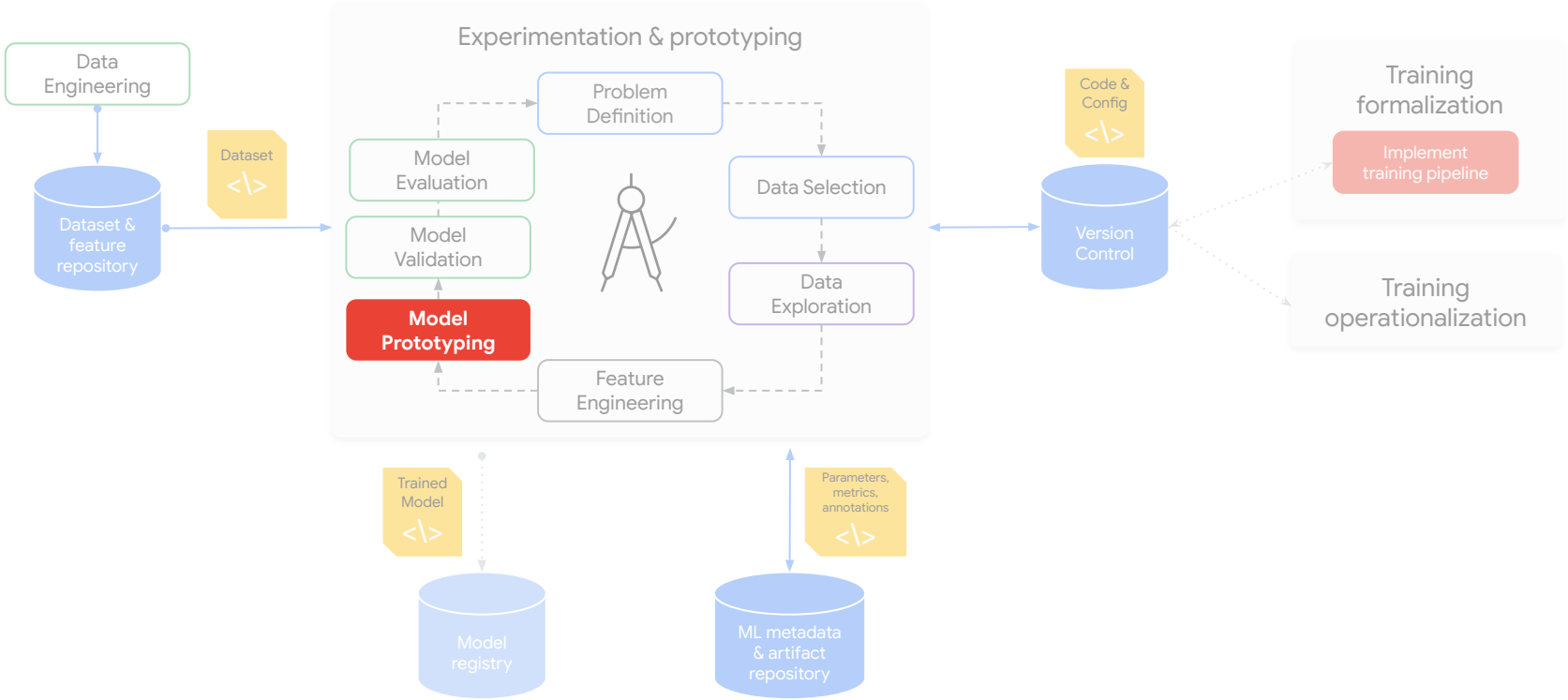
Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

ML Development: Model Prototyping

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

MLOps: ML Development



Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

The MLOps **Personas**

ML Engineer

ML Researcher

Data Scientist

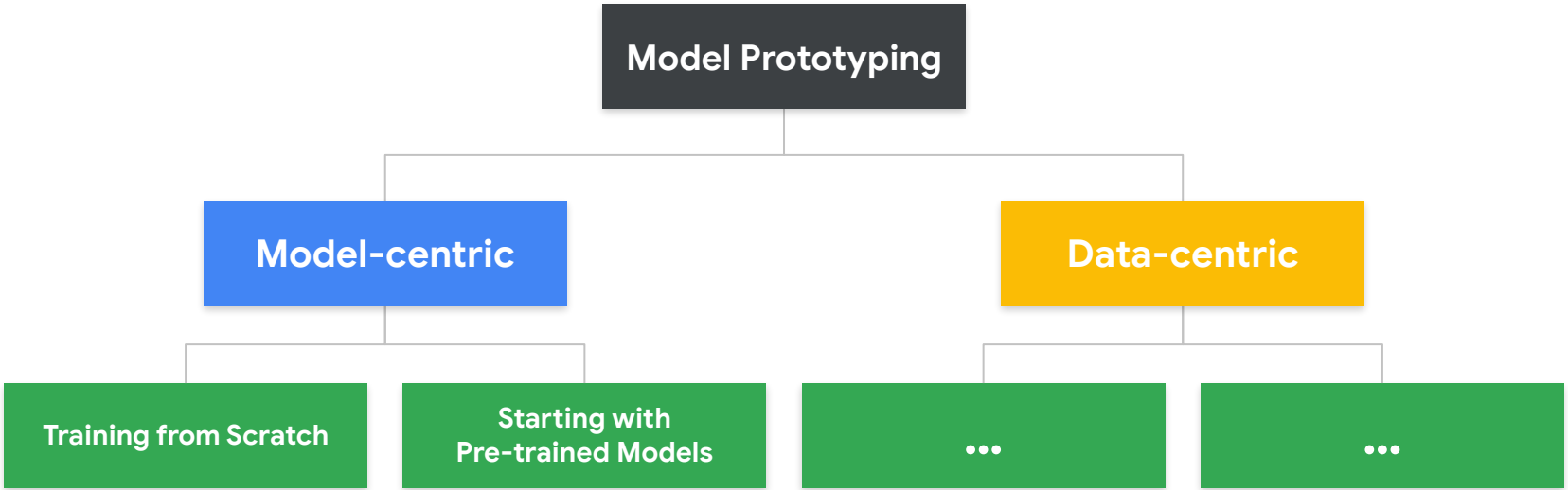
Data Engineer

Software Engineer

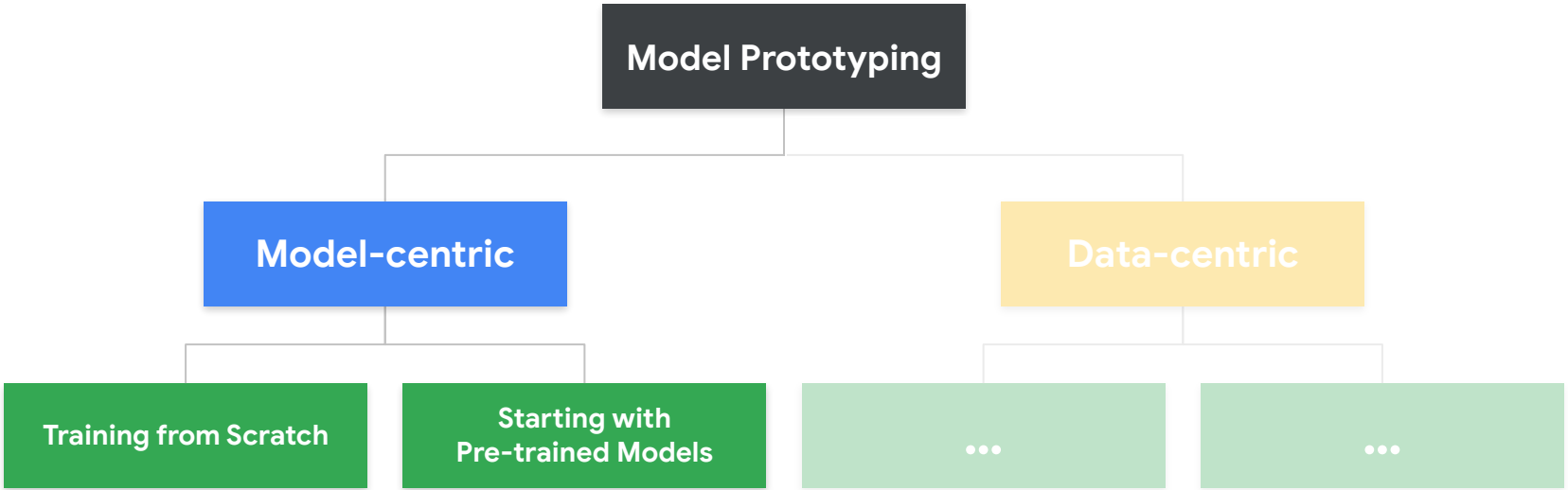
DevOps

Business Analyst

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

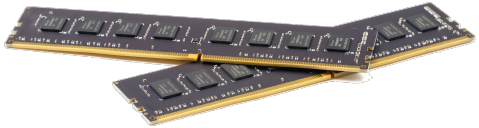


Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

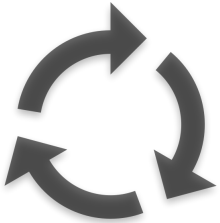


Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

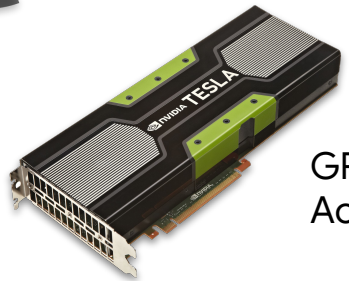
Training Pipeline: **Need Compute Resources**



Memory



Compute



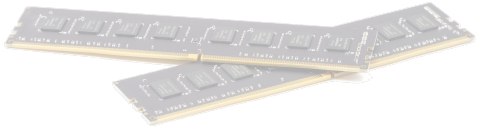
GPU and
Accelerators

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

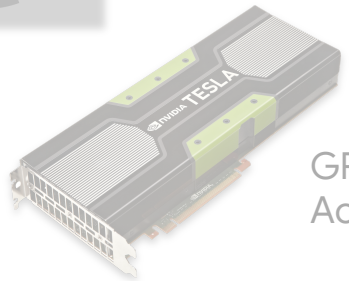
Training Pipeline: **Need Compute Resources**

Computationally Intensive
Repeated Many Times (Epochs)

Compute



Memory

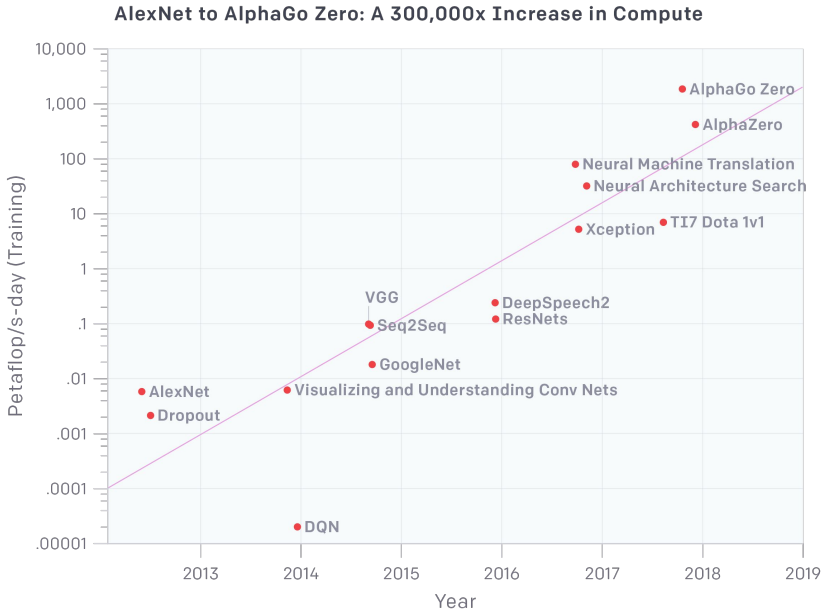


GPU and
Accelerators

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Compute Requirements are **Growing**

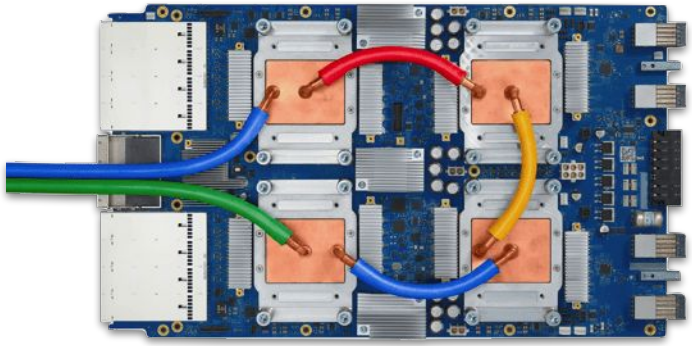
"… **since 2012 the amount of compute used in the largest AI training runs has been increasing exponentially with a 3.5 month-doubling time** (by comparison, Moore's Law had an 18-month doubling period). Since 2012, this metric has **grown by more than 300,000x (an 18-month [Moore's Law] doubling period would yield only a 12x increase)**. Improvements in compute have been a key component of AI progress, so as long as this trend continues, it's worth preparing for the implications of systems far outside today's capabilities."

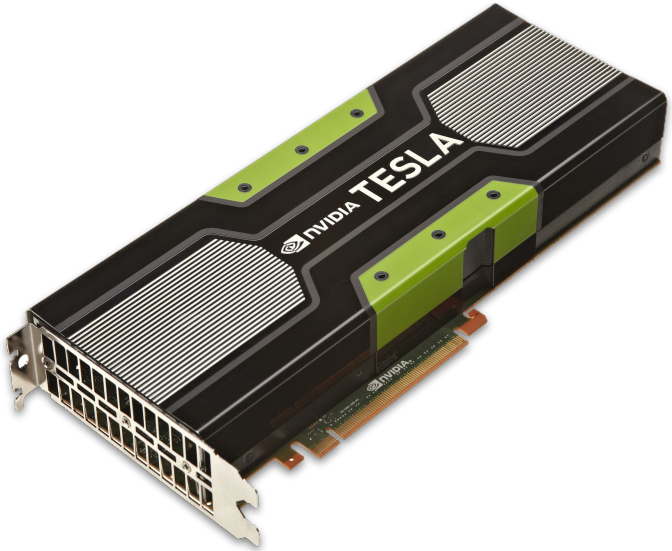


Source: https://blog.openai.com/ai-and-compute/

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

TPUs/GPUs





Elliot Turner
@eturner303

Holy crap: It costs $245,000 to train the XLNet model (the one that's beating BERT on NLP tasks..512 TPU v3 chips * 2.5 days * $8 a TPU) - arxiv.org/abs/1906.08237

XLNet: Generalized Autoregressive Pretraining for Language Understanding

Zhilin Yang*[1], Zihang Dai*[12], Yiming Yang[1], Jaime Carbonell[1], Ruslan Salakhutdinov[1], Quoc V. Le[2]
[1]Carnegie Mellon University, [2]Google Brain
{zhiliny,dzihang,yiming,jgc,rsalakhu}@cs.cmu.edu, qvl@google.com

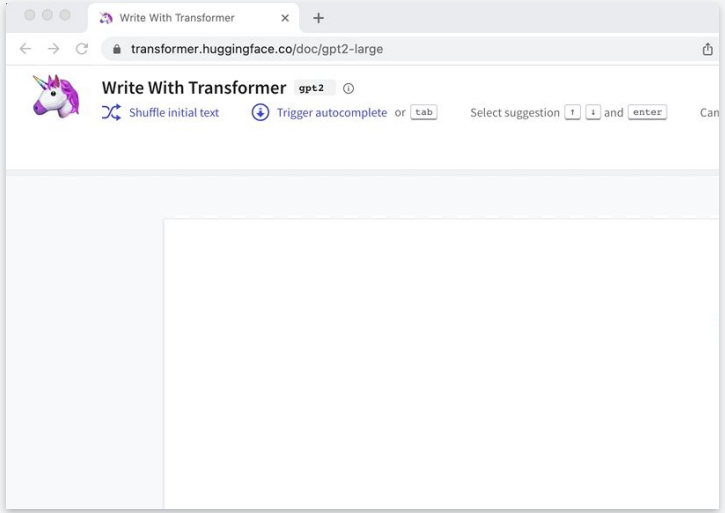
Abstract

With the capability of modeling bidirectional contexts, denoising autoencoding based pretraining like BERT achieves better performance than pretraining approaches based on autoregressive language modeling. However, relying on corrupting the input with masks, BERT neglects dependency between the masked positions and suffers from a pretrain-finetune discrepancy. In light of these pros and cons, we propose XLNet, a generalized autoregressive pretraining method that (1) enables learning bidirectional contexts by maximizing the expected likelihood over all permutations of the factorization order and (2) overcomes the limitations of BERT thanks to its autoregressive formulation. Furthermore, XLNet integrates ideas from Transformer-XL, the state-of-the-art autoregressive model, into pretraining. Empirically, XLNet outperforms BERT on 20 tasks, often by a large margin, and achieves state-of-the-art results on 18 tasks including question answering, natural language inference, sentiment analysis, and document ranking.[1].

Presenter HERE


GPT-2

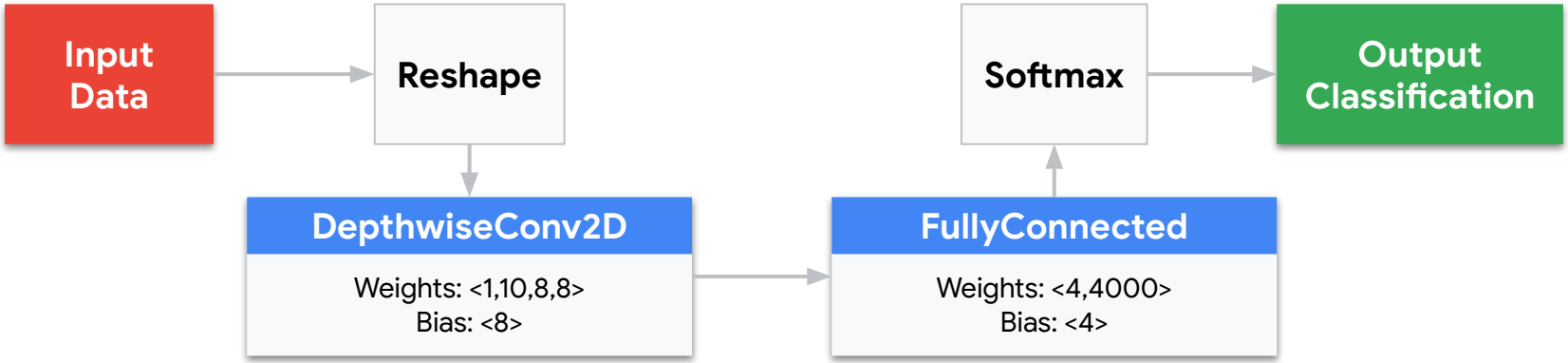
US $256/hr


Write With Transformer
transformer.huggingface.co/doc/gpt2-large
Write With Transformer
gpt2
Shuffle initial text
Trigger autocomplete
or
tab
Select suggestion
and
enter

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

TinyML Also Takes Long

(okay, not that long...)

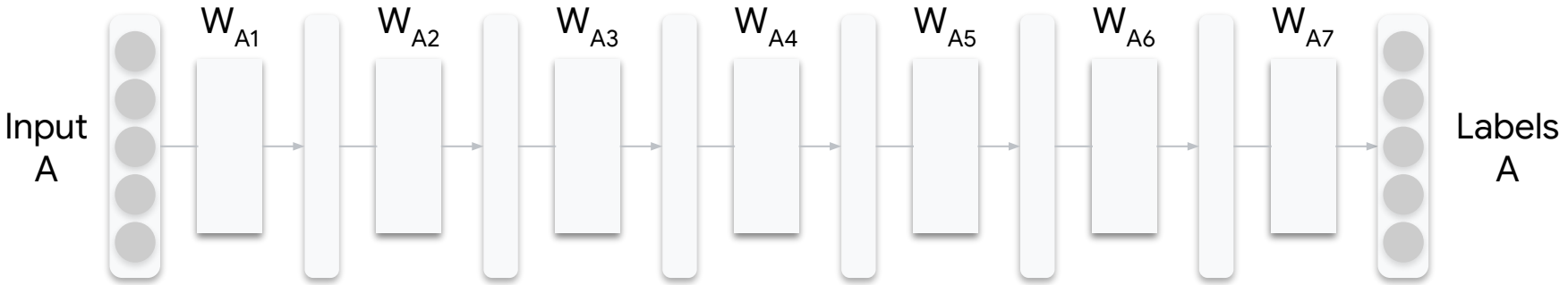


2 hours for a 5-word model

Pre-trained Models

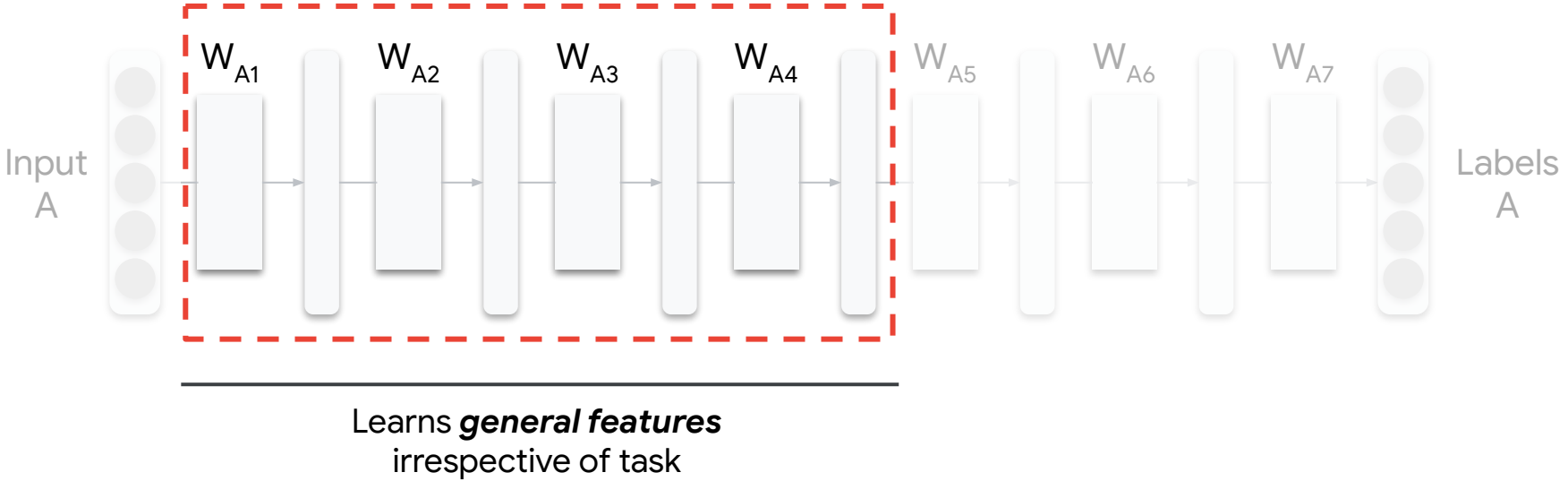
Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

End Result of Training



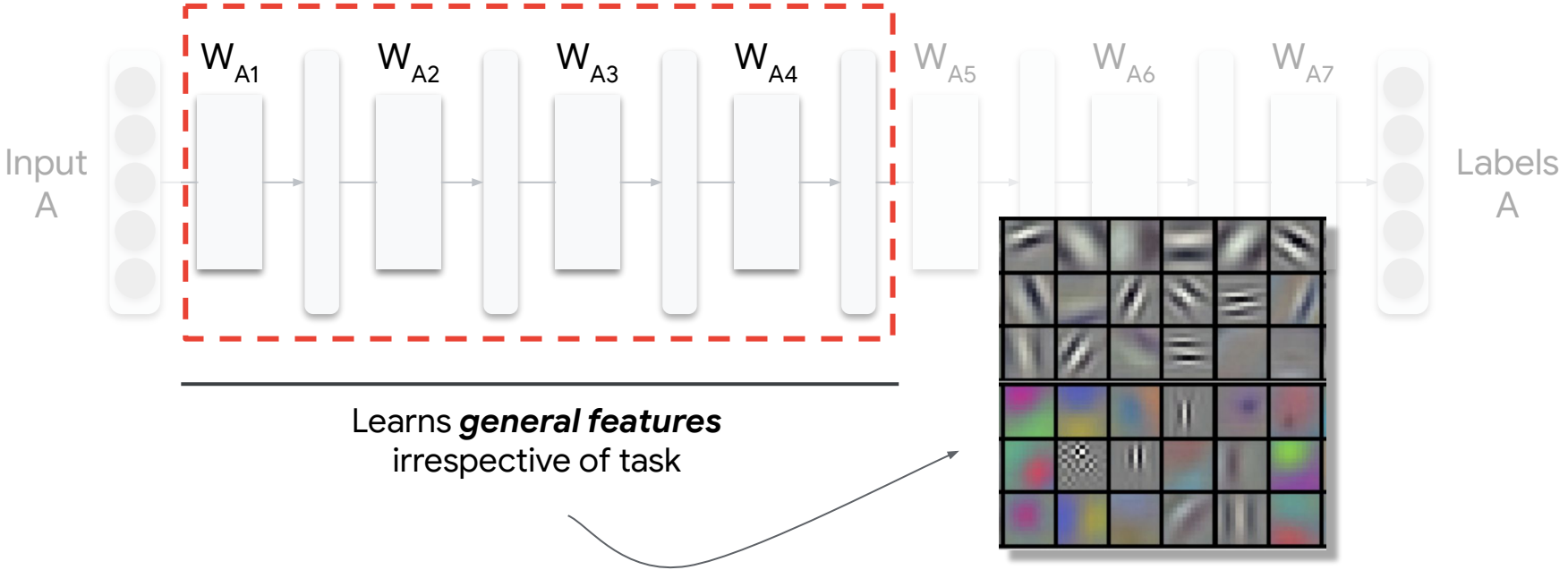
Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

End Result of Training



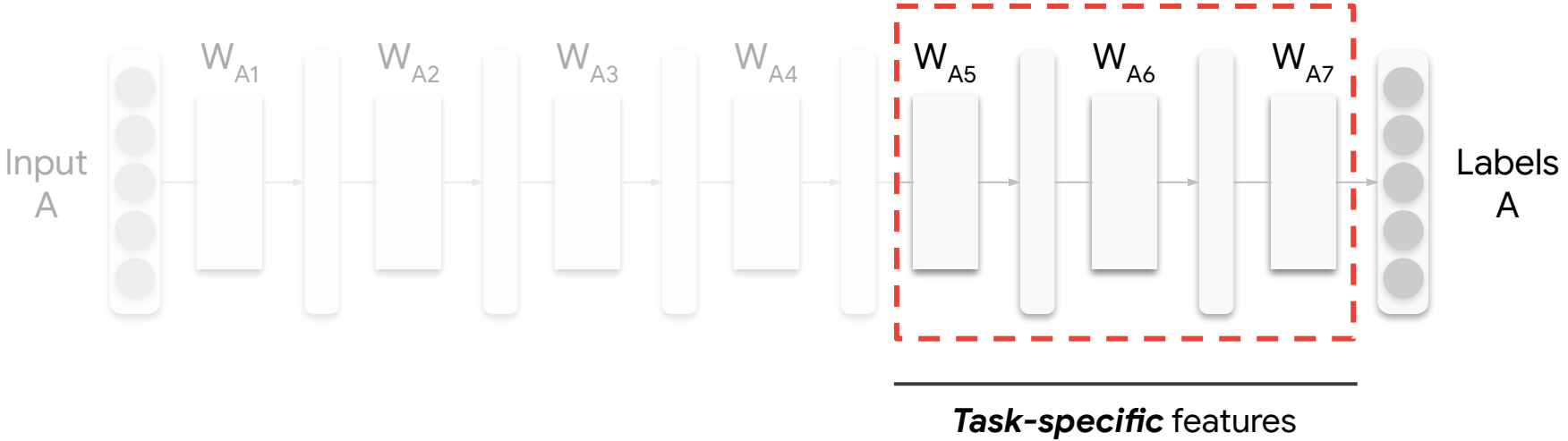
Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

End Result of Training



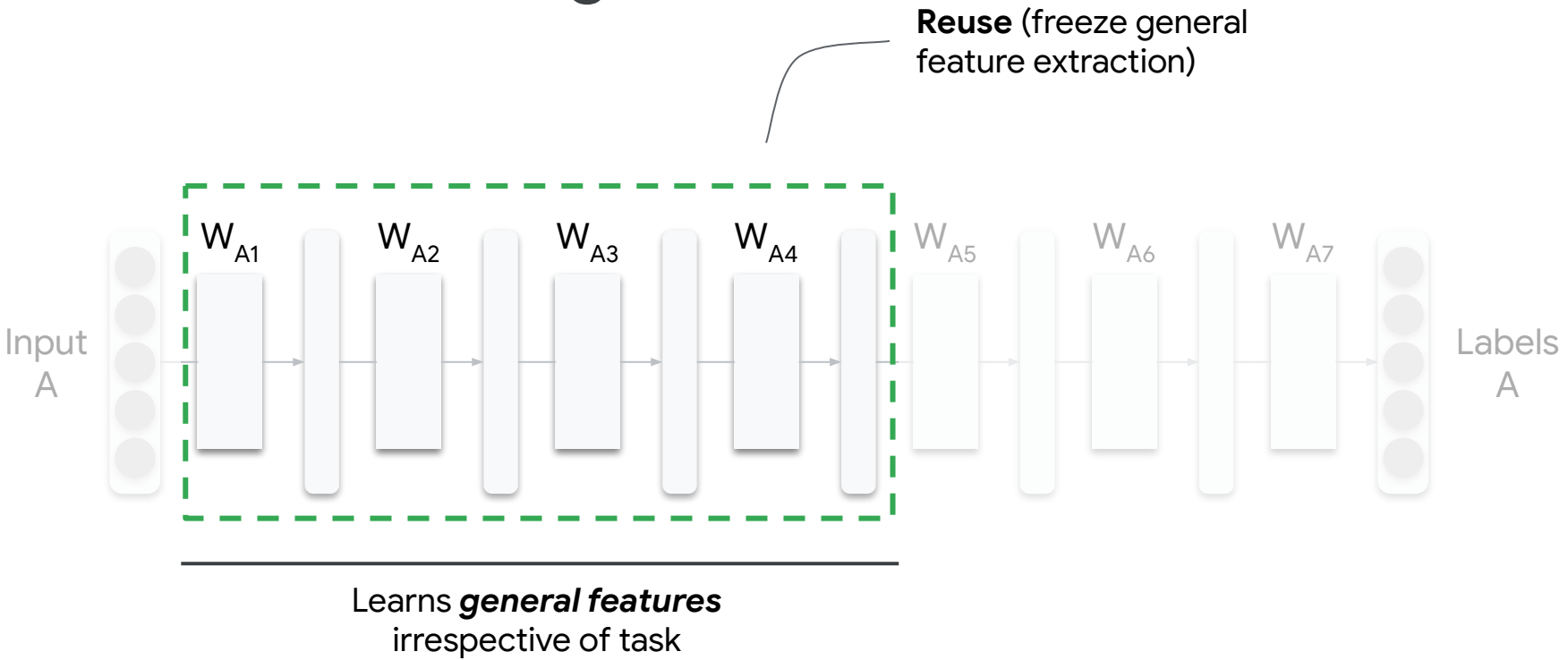
Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

End Result of Training



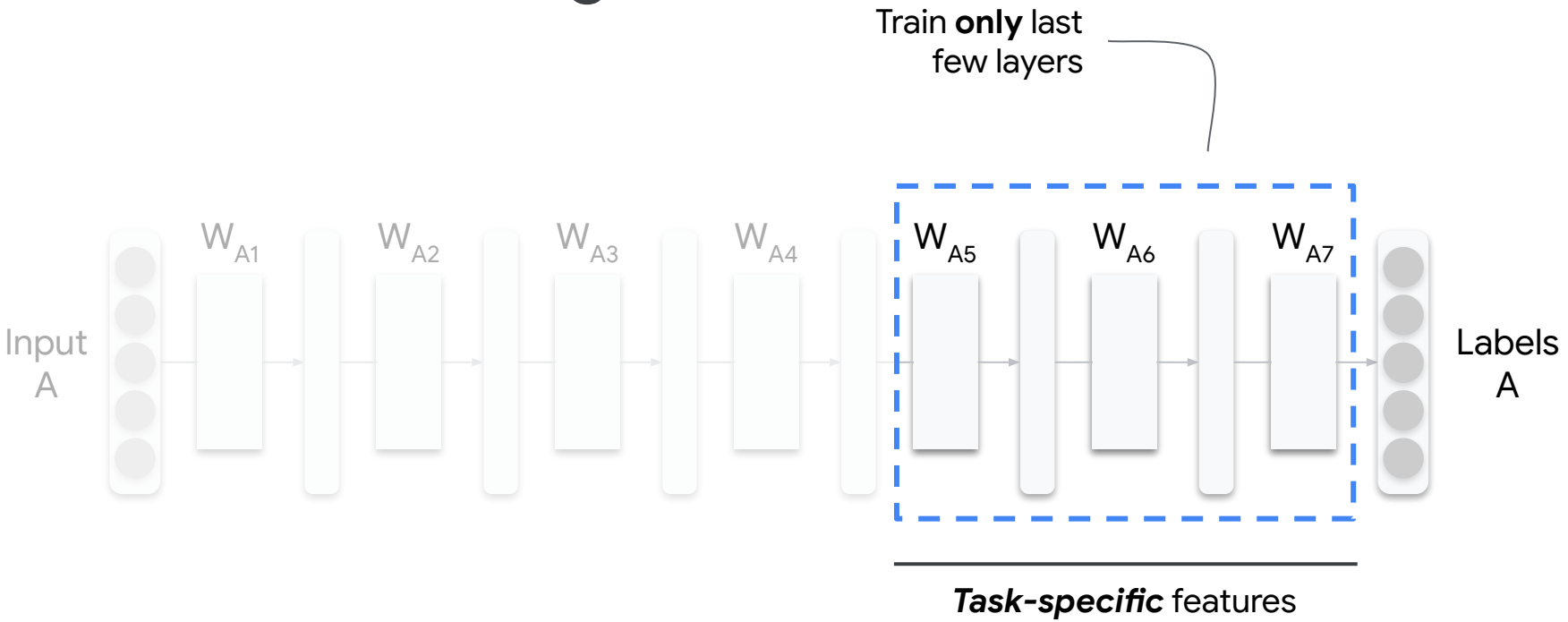
Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Transfer Learning



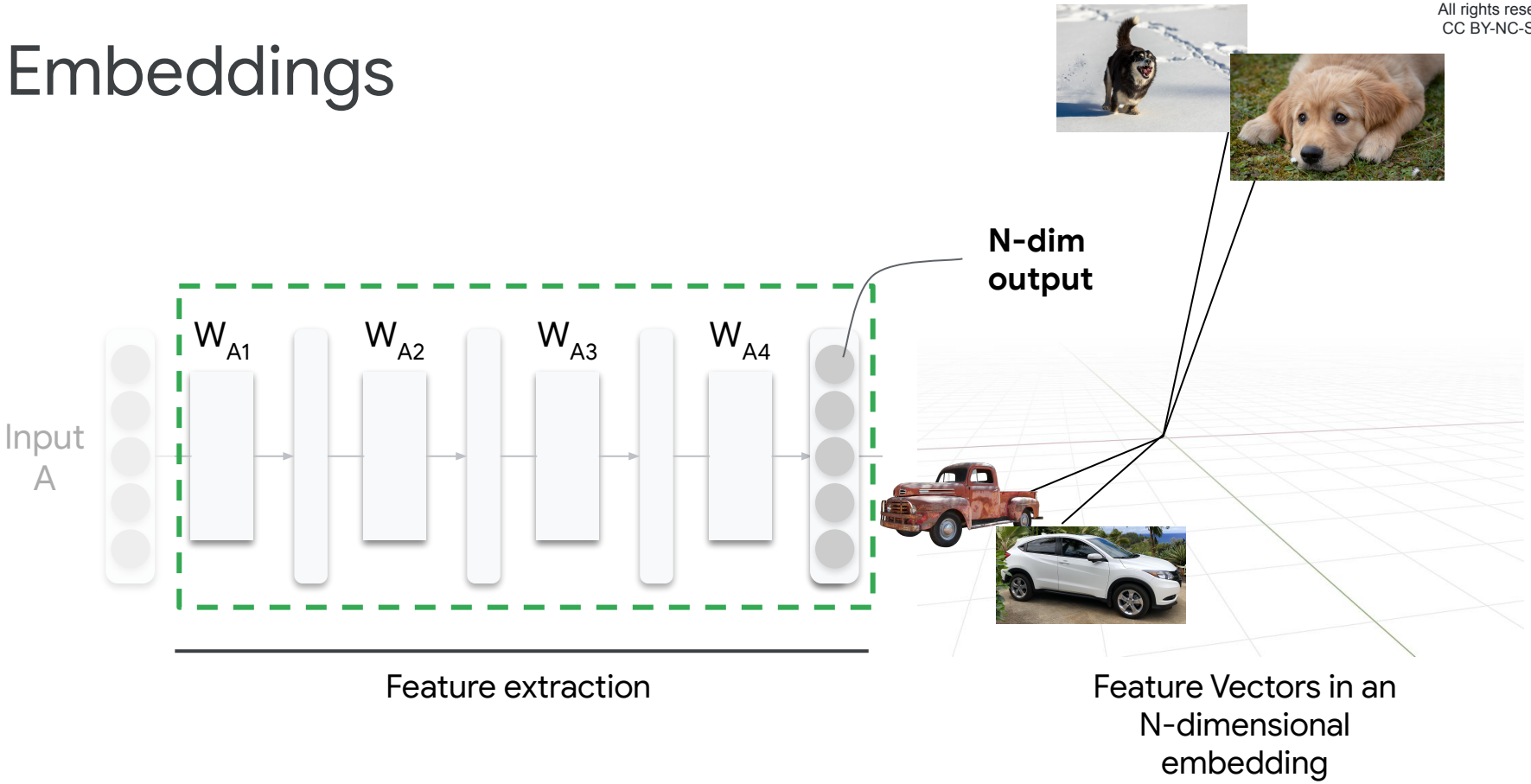
Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Transfer Learning



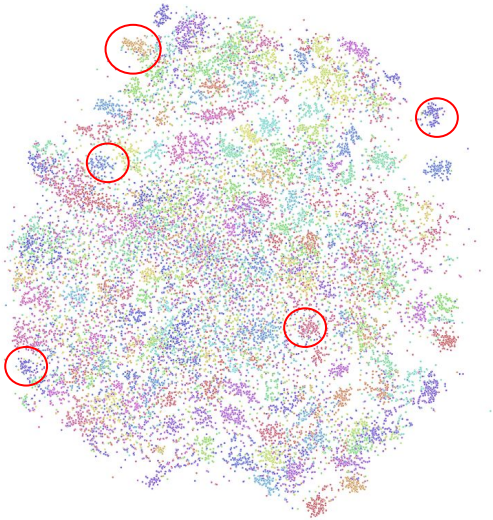
Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Embeddings



Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

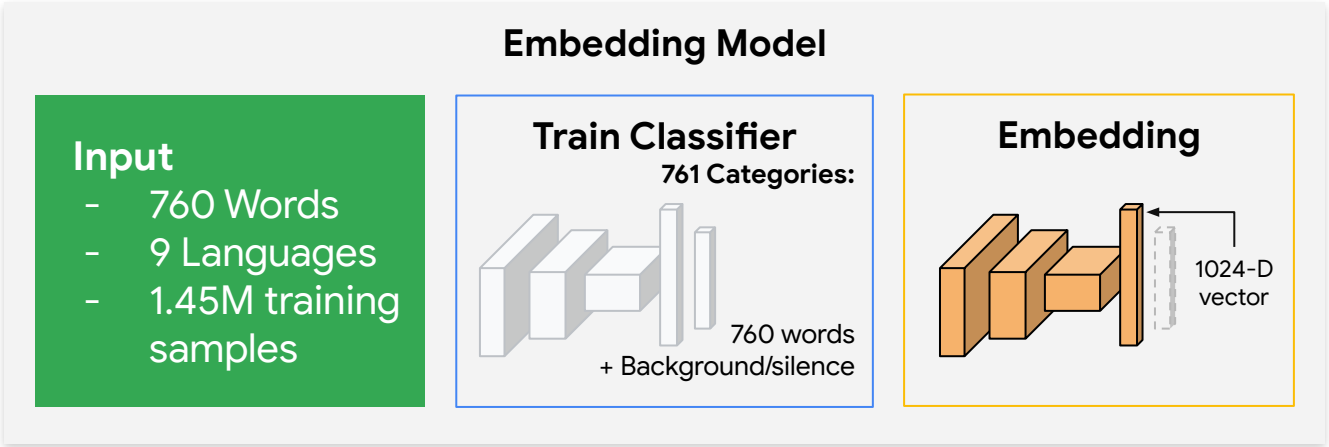
t-SNE view of 165 novel words (not used to train the classifier)

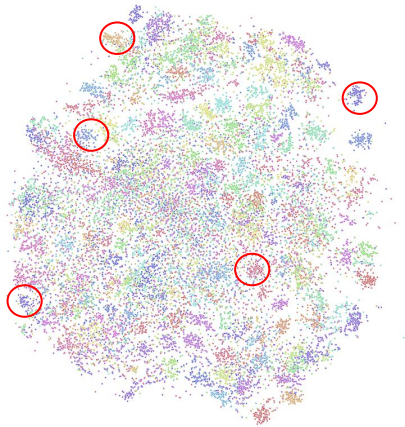


Each cluster is a different word (each point is a different recording from different speakers)

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

KWS Embeddings





- Minimize **intra**-class distance
- Maximize **inter**-class distance

Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0

Few Shot Keyword Spotting in Any Language

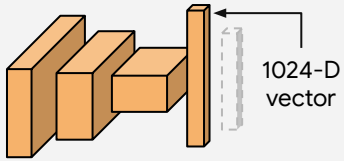
Embedding Model

Input
- 760 Words
- 9 Languages
- 1.45M training samples

Train Classifier
761 Categories:

760 words
+ Background/silence

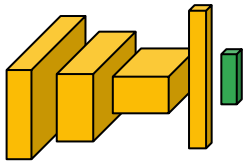
Embedding



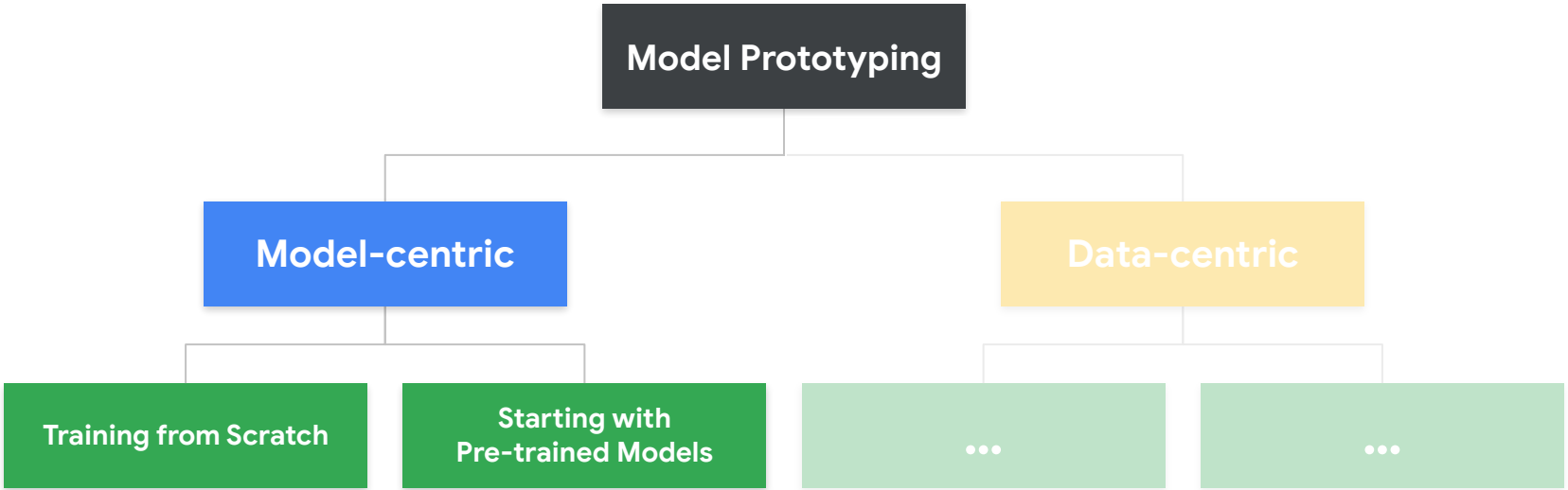
Input: 5 words in any language

Finetune KWS Model

3 Categories
Target Keyword
Unknown Word
Background/silence



Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0



Copyright (c) 2022 TinyMLedu.
All rights reserved.
CC BY-NC-SA 4.0