Google AI
Brain Team

Learning with Imperfect Supervision

Mostafa Dehghani

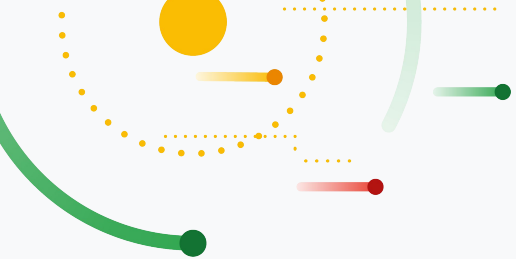
Research Scientist at Google Brain

dehghani@google.com

Stephan Gouws, Oriol Vinyals, Jakob Uszkoreit, Łukasz Kaiser, Samira Abnar, Jaap Kamps, Hosein Azarbonyad, Maarten de Rijke, Arash Mehrjou, Bernhard Schölkopf, Aliaksei Severyn, Sascha Rothe, Hamed Zamani, W. Bruce Croft, Pascal Fleury, Maarten Marx.

Google Brain · University of Amsterdam · DeepMind · Google AI · Max-Planck-Gesellschaft · University of Massachusetts Amherst

Google AI



Google AI
Brain Team

"Give orange me give eat orange me eat orange give me eat orange give me you"

"Give orange me give eat orange me eat orange give me eat orange give me you"

Nim Chimpsky

Nim Chimpsky (November 19, 1973 – March 10, 2000) was a chimpanzee who was the subject of an extended study of animal language acquisition at Columbia University, led by Herbert S. Terrace, as a challenge to Chomsky's thesis that full-fledged language use was innate only to humans. ***This quote is the Nim's longest recorded sentence.***

The Fantastic Human Being

- Human effortlessly learn about new concepts and solve complex problems from limited, noisy or inconsistent observations and routinely draw successful generalization on them.

- **Poverty of the Stimulus**

> *"My own suspicion is that a central part of what we call “learning” is actually better understood as the growth of cognitive structures along an internally directed course under the triggering and partially shaping effect of the environment."*
>
> Noam Chomsky

Chomsky, Noam. "Rules and representations." Behavioral and brain sciences 3.1 (1980): 1-15.

Data Hungry Models

- The performance of today's successful learning models is often strongly correlated with the amount of available labeled data.

The more data you have,

the more accurate your model will be!

Sun, Chen, et al. "Revisiting unreasonable effectiveness of data in deep learning era." *Proceedings of the IEEE international conference on computer vision*. 2017.

Dealing with Data Scarcity in Practice

- **Using distant or heuristic supervision**
  - A heuristic labeling rule or function which can be relying on an external source of knowledge.
- **Using incidental signals**
  - Signals that exist in the data and the environment independently of the tasks and they are co-related to the target tasks.

Dealing with Data Scarcity in Practice

- **Providing supervision by specifying constraints**
  - Supervising by setting constraints that should hold over the output space.
- **Applying bootstrapping, self-supervised feature learning, and data augmentation**
  - Make statistically efficient reuse of available data.
- **Using transfer learning**
  - generalizing knowledge across domains/tasks.

Dealing with Data Scarcity in Practice

- **Using active learning and response-based supervision**
  - Designing models that learn from the feedback that it receives by interacting with an environment
- **Introducing a form of structured prior knowledge**
  - Using the property of the data to learn more about the data

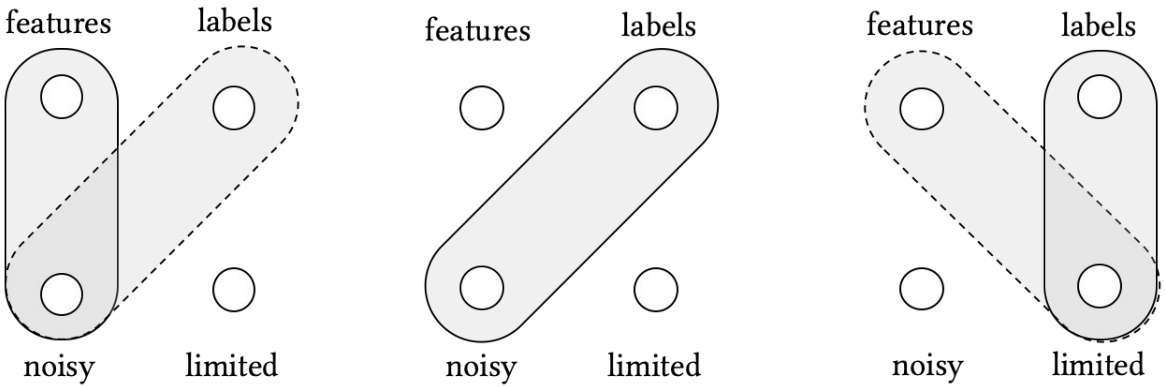
Dealing with Data Scarcity in Practice

- **Indirect supervision**
  - For instance, a companion binary task is defined for which obtaining training data is easier
- **Zero/one/few-shot learning**
  - Learning knowledge that can be extended to new tasks by observing just a few examples

Dealing with Data Scarcity in Practice

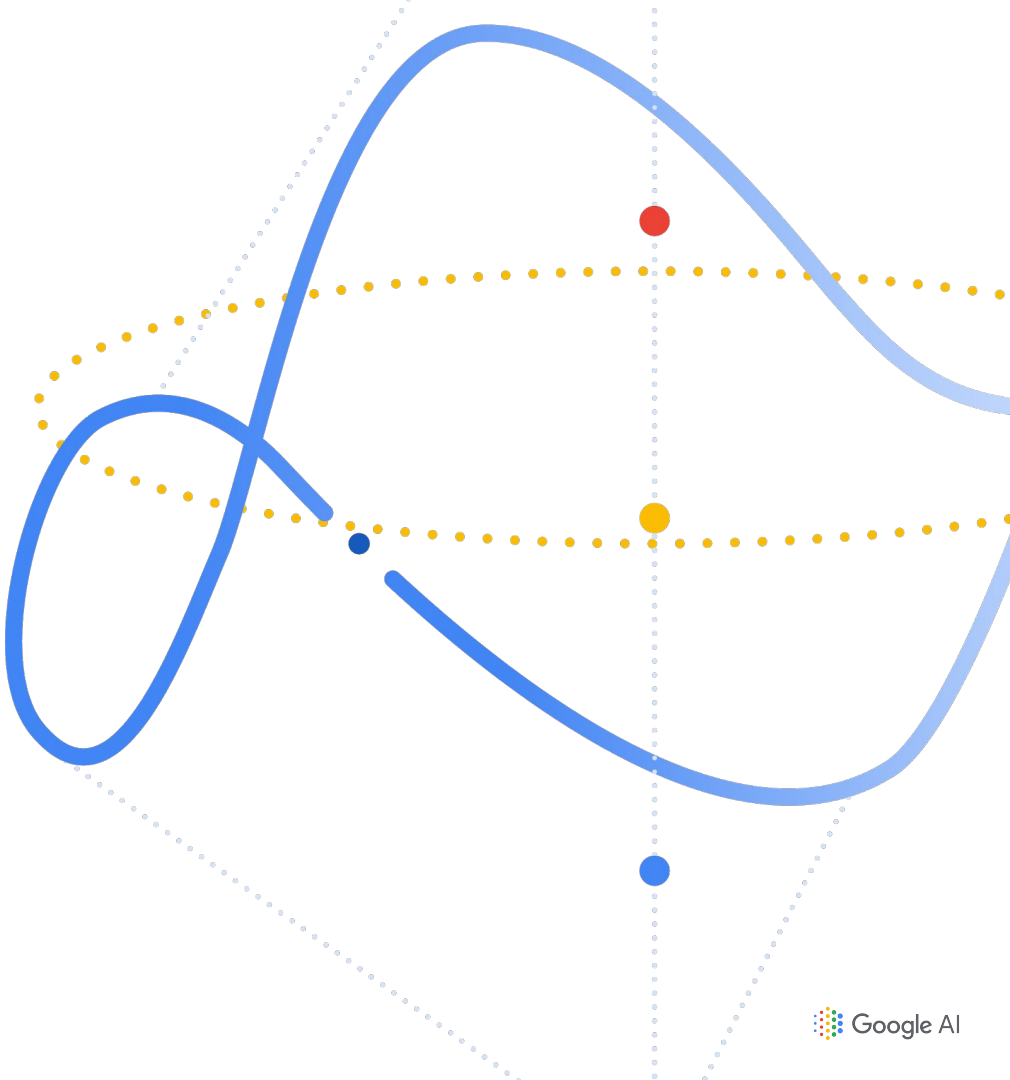
- **Exploiting noisy and inaccurate labels**
  - Learning from inaccurate, incomplete, and inexact supervision.
    - biased or weak classifiers, crowd-sourced data
- **Injecting inductive biases into algorithms**
  - Encoding modeling assumption as inductive biases to generalize better on unobserved data

Imperfection in the Supervision Signal



Google AI
Brain Team

Knowledge Matters: Structure of the Data as Prior Knowledge



Google AI

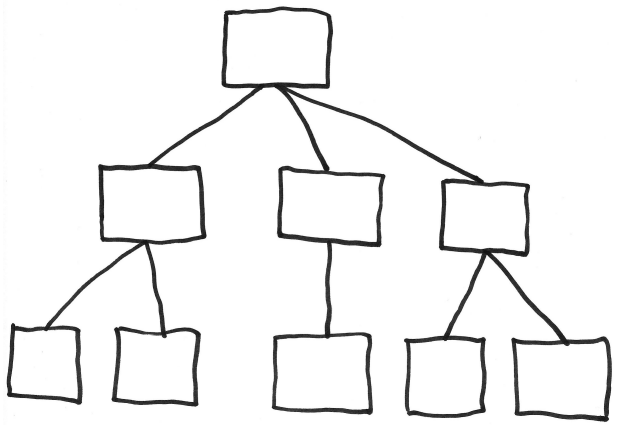
We understand the world in structural terms...

- We approach new problems armed with extensive **prior experience and knowledge**
    - When we learn:we either **fit** the **new knowledge** into our **existing structured** representations.or we **adjust** the **existing structure** to better accommodate our new and the old observation.

- When **building intelligent machines**, taking the structure of the data into account facilitates **modeling complex information**.

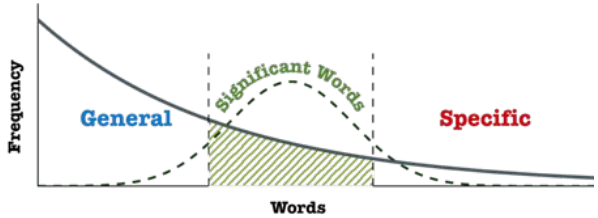
Hierarchies

- Hierarchical structures:
  - Model different levels of associations and abstract away fine-grained differences.



Neither General, nor Specific, but Significant

- Significant Words Language Models
  - Capturing, only and all, the significant features, by removing general and specific features:
    - General features: Not discriminative
    - Specific features: Not inclusive



Dehghani, M. (2016). Significant words representations of entities.In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '16

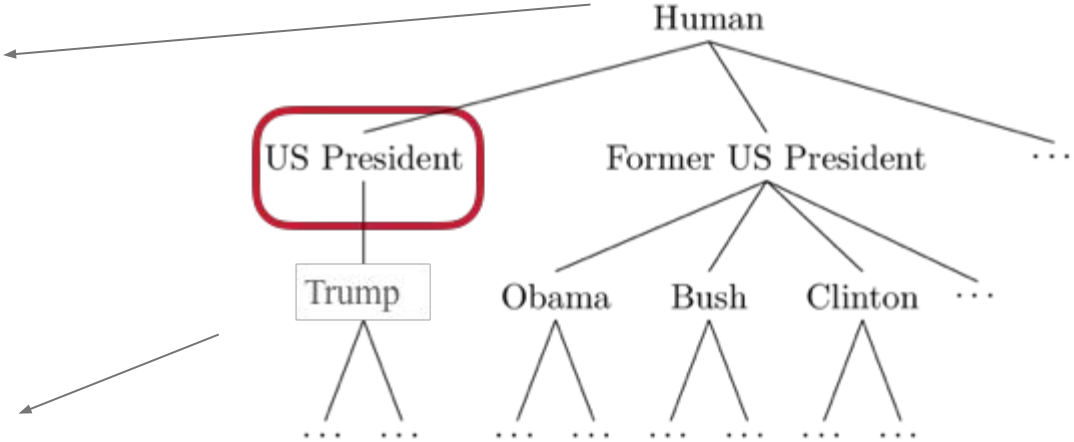
Example

propertyless common observations

being able to eat

having two legs

Tweet per second rate



unreliable rare observations

Pseudo Relevance Feedback in the Retrieval

- **Ranking** task: given a Query and a set of documents, retrieve and rank relevant documents.
  - **PRF**: using the top-ranked documents in the initial retrieved results for the feedback and improve the ranking.
  - **Noisy** data:
    - top ranked are not always relevant
    - many noisy terms, even in relevant documents

Dehghani, M., Azarbonyad, H., Kamps, J., Hiemstra, D., andMarx, M. (2016c). Luhn revisited: Significant words language models. InProceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM ’16.

Pseudo Relevance Feedback in the Retrieval

- Topic 374 of the TREC Robust04 test collection: ``**Nobel prize winners**'

Standard-LM	
prize	5.55e-02
nobel	3.36e-02
physics	2.35e-02
science	2.18e-02
...	
time	1.68e-02
...	
palestiniar	1.34e-02
year	1.34e-02
...	

General-LM	
new	3.70e-03
cent	2.98e-03
two	2.97e-03
dollars	2.76e-03
people	2.71e-03
...	
time	2.47e-03
...	
year	2.16e-03
...	

SMM [45]	
prize	6.07e-02
nobel	4.37e-02
awards	3.43e-02
chemistry	3.23e-02
physics	2.82e-02
palestiniar	2.18e-02
cesium	2.09e-02
arafat	1.94e-02
university	1.92e-02
...	

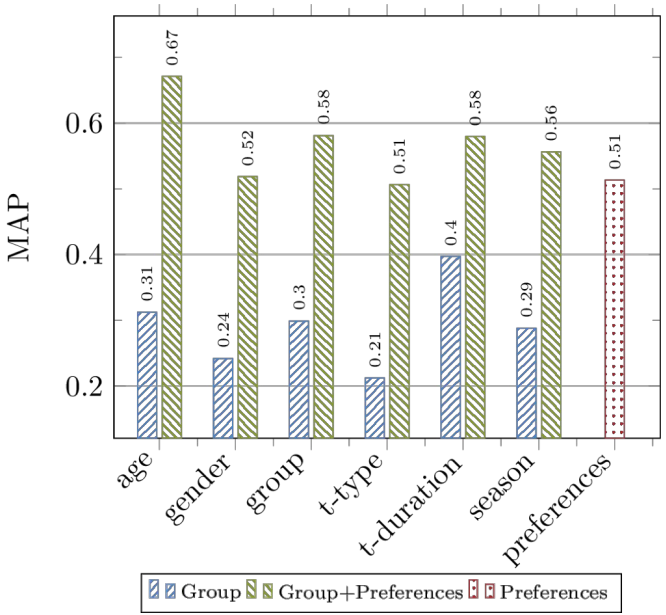
Specific-LM	
insulin	2.25e-02
palestinian	2.15e-02
dehmelt	1.81e-02
oscillations	1.79e-02
waxman	1.69e-02
marcus	1.69e-02
attack	1.61e-02
...	
arafat	1.29e-02
...	

SWLM	
prize	6.02e-02
nobel	4.53e-02
science	2.68e-02
award	2.43e-02
physics	1.94e-02
winner	1.90e-02
won	1.80e-02
peace	1.80e-02
discovery	1.71e-02
...	

Dehghani, M., Azarbonyad, H., Kamps, J., Hiemstra, D., andMarx, M. (2016c). Luhn revisited: Significant words language models. InProceedings of the 25th ACM International on Conference on Information and Knowledge Management, CIKM '16.

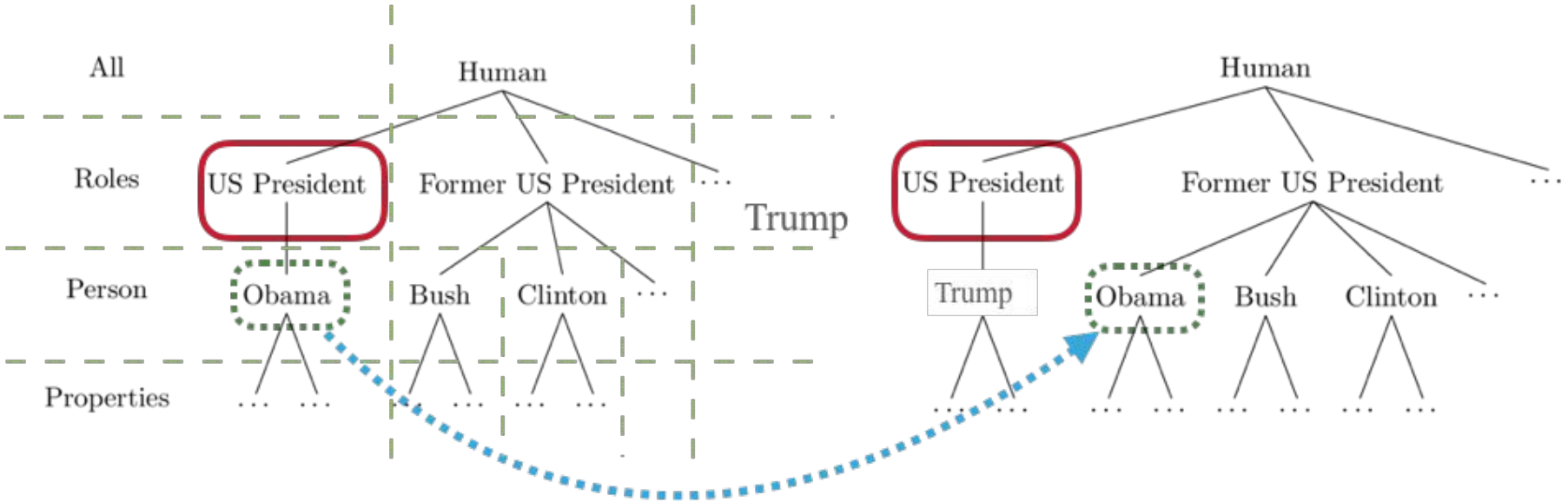
Group Profiling for Recommendations

- Cold Start problem
  - Modeling a **group** of people and use these models based on memberships as initial profile



Dehghani, M., Azarbonyad, H., Kamps, J., and Marx, M. (2016d).Generalized group profiling for content customization. InCHIIR'16, CHIIR '16.
Dehghani, M., Azarbonyad, H., Kamps, J., and Marx, M. (2016f).Significant words language models for contextual suggestion.Proceedings National Institute for Standards and Technology. NIST Special Publication: SP, 500

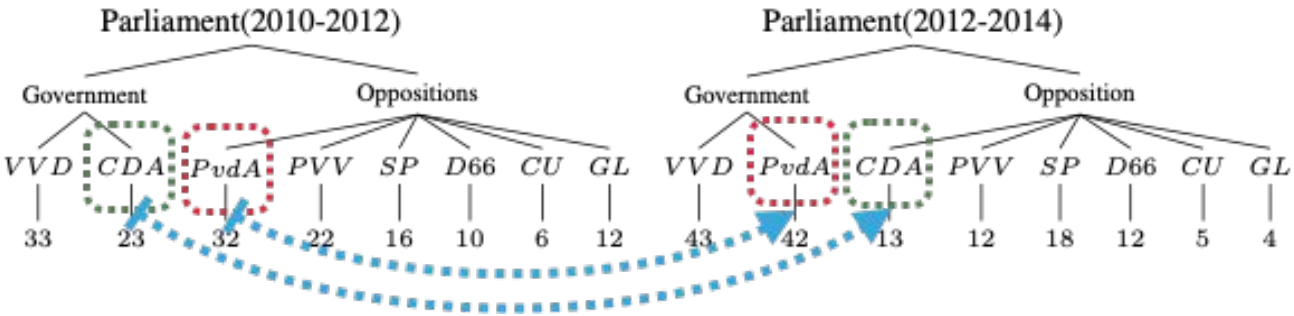
Transferability over time



Dehghani, M., Azarbonyad, H., Kamps, J., and Marx, M. (2016e).On horizontal and vertical separation in hierarchical text classification. In The proceedings of ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR'16), ICTIR'16

Dehghani, M., Azarbonyad, H., Kamps, J., and Marx, M. (2016g).Two-way parsimonious classification models for evolving hierar-chies. InProceedings of Conference and Labs of the Evaluation Forum,CLEF '16

Transferability over time

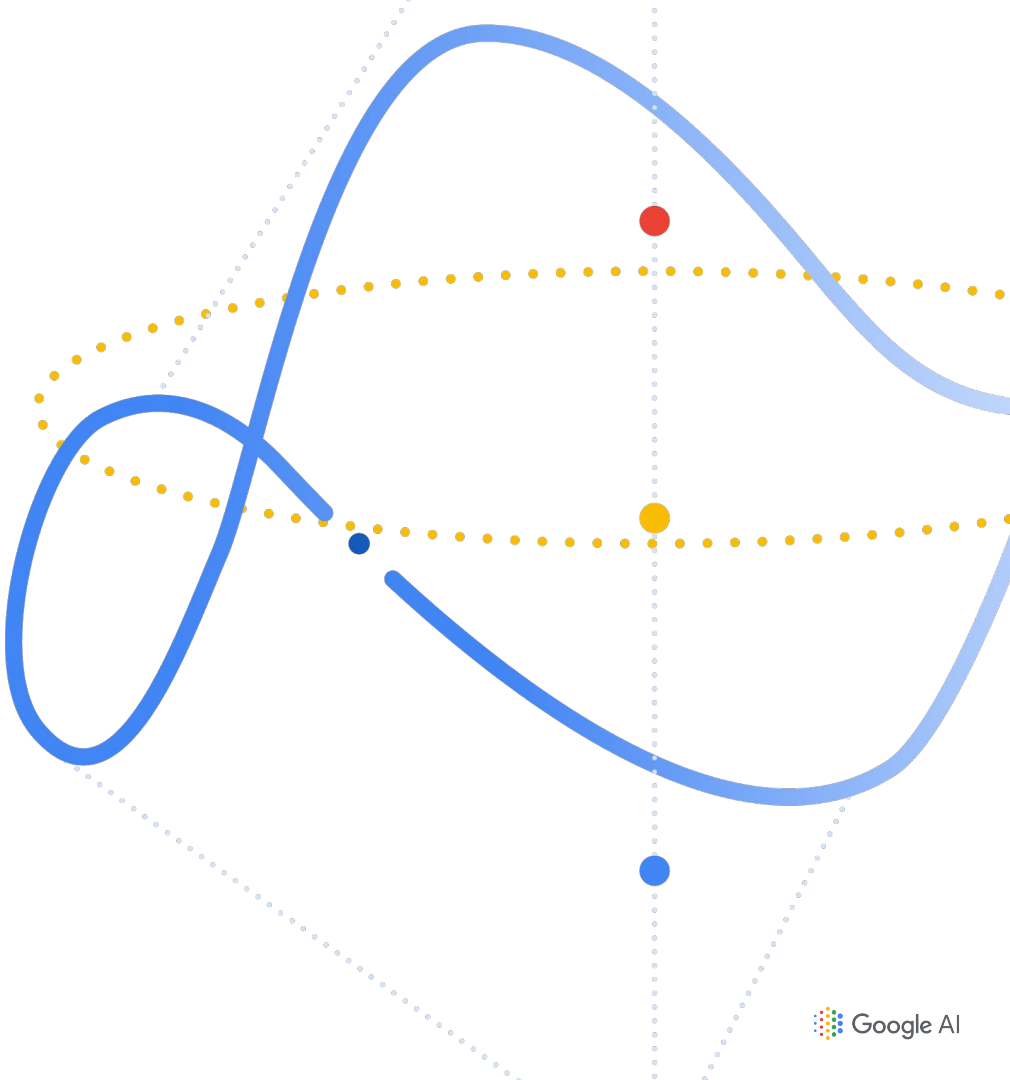


Dehghani, M., Azarbonyad, H., Kamps, J., and Marx, M. (2016e).On horizontal and vertical separation in hierarchical text classification. In The proceedings of ACM SIGIR International Conference on the Theory of Information Retrieval (ICTIR'16), ICTIR'16

Dehghani, M., Azarbonyad, H., Kamps, J., and Marx, M. (2016g).Two-way parsimonious classification models for evolving hierar-chies. InProceedings of Conference and Labs of the Evaluation Forum,CLEF '16

Google AI
Brain Team

Learning with Weak Labels



Google AI

Success of ML?

- Most of the successes are on **stable benchmark** tasks where **standard large-enough datasets exist** to train neural networks.
- What happens if we stray slightly from these standard benchmark tasks toward the **realm of real-world applications?**
    - **No labeled data!**

Supervising Learning algorithms Programmatically

- What if human can supervise machine learning systems, **by labeling training data programmatically** instead of **labeling by hand**?
  - Using **heuristic based** methods as a **weak annotator** to generate **pseudo-labels** for a large set of unlabeled instances.
  - How to **generalize beyond the imperfection** of the weak annotator?

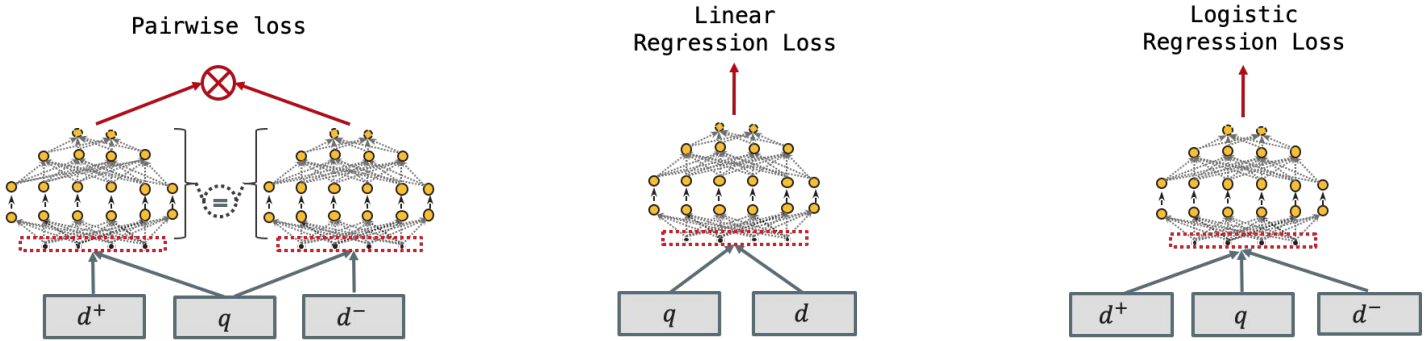
Preserving Privacy

- Building models that can learn from noisy signals can benefit preserving privacy where some noise is **intentionally added** to the training signal to preserve privacy.
    - Usually, adding noise is an important step to guarantee a certain level of differential privacy.

Dehghani, M., Azarbonyad, H., Kamps, J., and de Rijke, M.(2017a). Share your model instead of your data: Privacy preserving mimic learning for ranking. In SIGIR Workshop on Neural Information Retrieval (Neu-IR)

Training a Neural Ranker with BM25's output!

- Labels: based on the BM25 score
  - 6 million queries
  - Different input representations and objective functions



Key Ingredients -1

- The first is the proper input representation. Providing the network with raw data and letting the network to learn the features that matter, gives the network a chance of learning how to ignore imperfection in the training data.
    - Consider use embedding instead of feature engineering

Dehghani, M., Zamani, H., Severyn, A., Kamps, J., and Croft,W. B. (2017g). Neural ranking models with weak supervision.InProceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’17

Key Ingredients -2

- The second ingredient is to target the right goal and define a proper objective function. In the case of having weakly annotated training data, by targeting some explicit labels from the data, we may end up with a model that learned to express the data very well, but is incapable of going beyond it.
    - This is especially the case with deep neural networks where there are many parameters and it is easy to learn a model that overfits the data.

Dehghani, M., Zamani, H., Severyn, A., Kamps, J., and Croft,W. B. (2017g). Neural ranking models with weak supervision.InProceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’17

Key Ingredients -3

- The third ingredient is providing the network with a considerable amount of diverse training examples.
    - Thanks to weak supervision, we can generate as much training data as we need with almost no cost.
    - Diversity: hard and easy examples → let the model learns at the edge of its ability

---

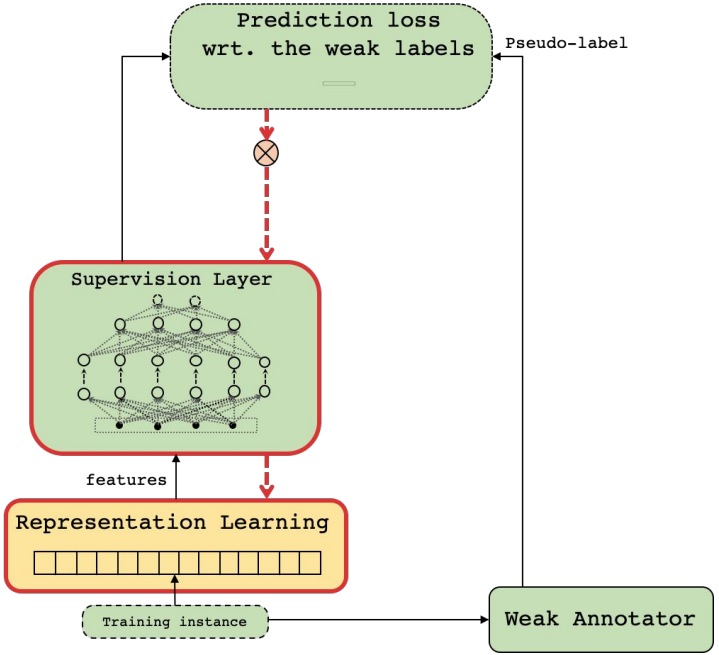
Dehghani, M., Azarbonyad, H., Kamps, J., and de Rijke, M.(2017a). Share your model instead of your data: Privacy preserving mimic learning for ranking. In SIGIR Workshop on Neural Information Retrieval (Neu-IR)

Meta-Learning the Label's Quality

- The third ingredient is providing the network with a considerable amount of diverse training examples.
    - Thanks to weak supervision, we can generate as much training data as we need with almost no cost.
    - Diversity: hard and easy examples → let the model learns at the edge of its ability

Dehghani, M., Mehrjou, A., Gouws, S., Kamps, J., and Schölkopf,B. (2018). Fidelity-weighted learning. In International Conference on Learning Representations (ICLR)
Dehghani, M., Severyn, A., Rothe, S., and Kamps, J. (2017f).Learning to learn from weak supervision by full supervision.InNIPS2017 workshop on Meta-Learning (MetaLearn 2017)
Dehghani, M., Severyn, A., Rothe, S., and Kamps, J. (2017e).Avoiding your teacher ’s mistakes: Training neural networks with controlled weak supervision.arXiv preprint arXiv:1711.00313

Weak annotation

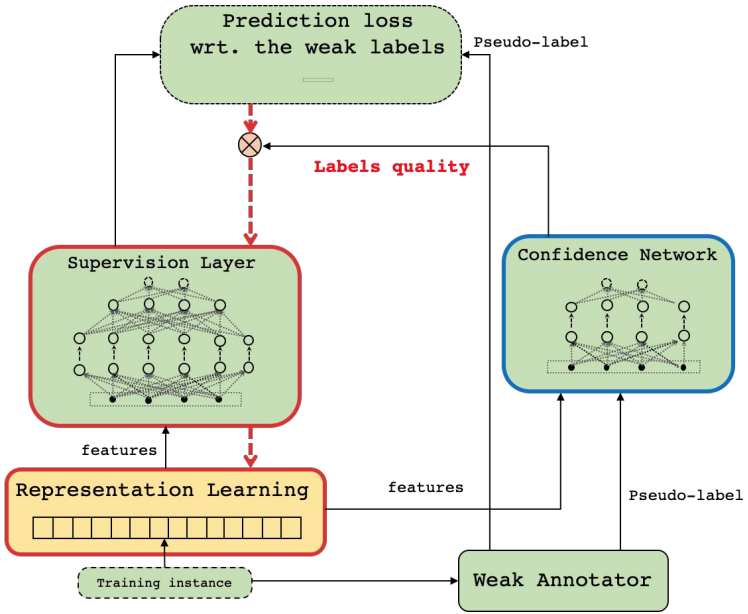


Label's Quality

> “All labels are equal, but some labels are more equal than others.”

Inspired by George Orwell, Animal Farm, 1945

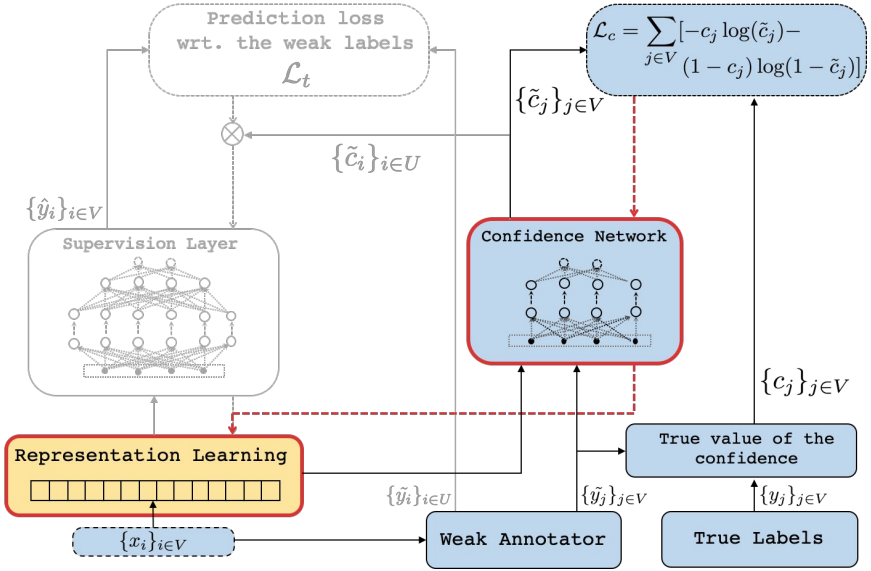
Meta-Learning the Label's Quality

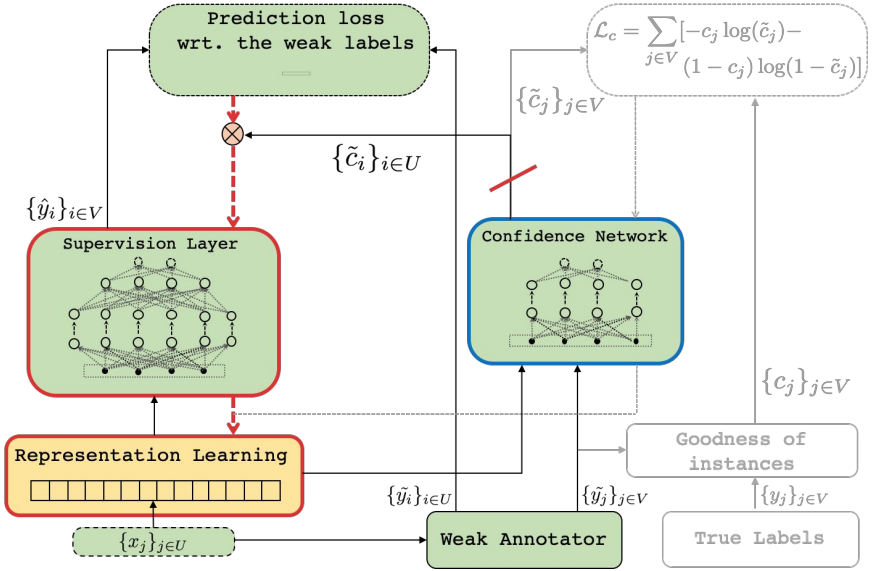


Dehghani, M., Severyn, A., Rothe, S., and Kamps, J. (2017f).Learning to learn from weak supervision by full supervision.InNIPS2017 workshop on Meta-Learning (MetaLearn 2017)
Dehghani, M., Severyn, A., Rothe, S., and Kamps, J. (2017e).Avoiding your teacher 's mistakes: Training neural networks with controlled weak supervision.arXiv preprint arXiv:1711.00313

Google AI
Brain Team

Learning with Controlled Weak Supervision

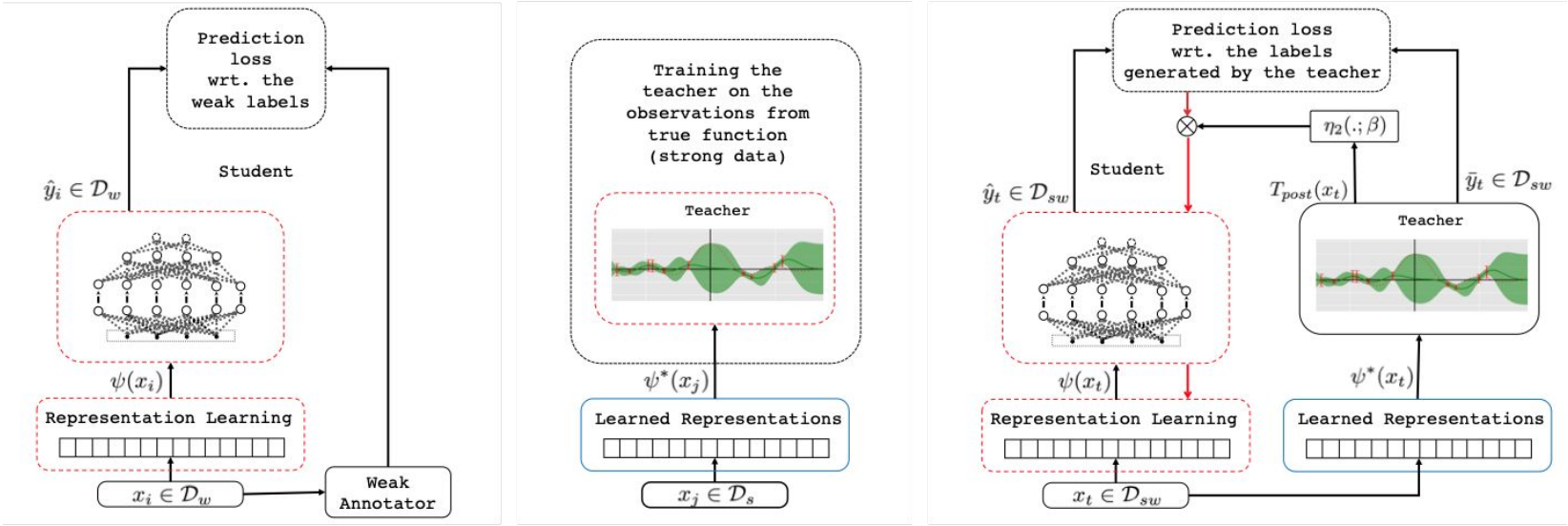




Dehghani, M., Severyn, A., Rothe, S., and Kamps, J. (2017f).Learning to learn from weak supervision by full supervision.InNIPS2017 workshop on Meta-Learning (MetaLearn 2017)

Dehghani, M., Severyn, A., Rothe, S., and Kamps, J. (2017e).Avoiding your teacher 's mistakes: Training neural networks with controlled weak supervision.arXiv preprint arXiv:1711.00313

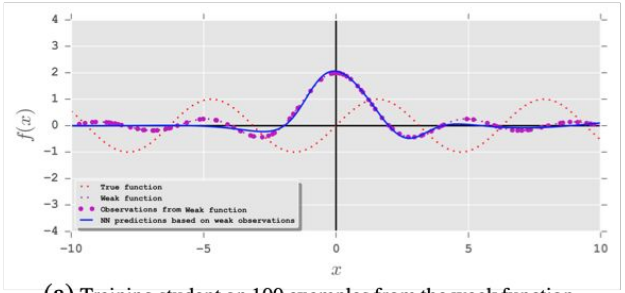
Fidelity Weighted Learning



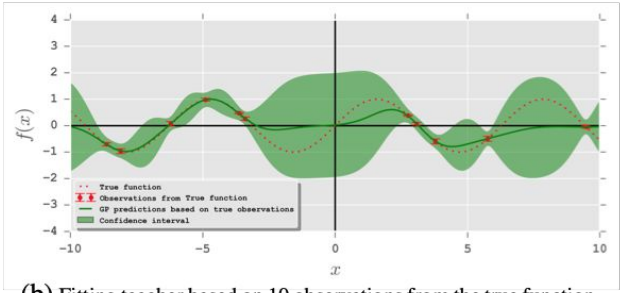
(a) Step 1 (b) Step 2 (c) Step 3

Dehghani, M., Mehrjou, A., Gouws, S., Kamps, J., and Schölkopf,B. (2018). Fidelity-weighted learning. In International Conference on Learning Representations (ICLR)

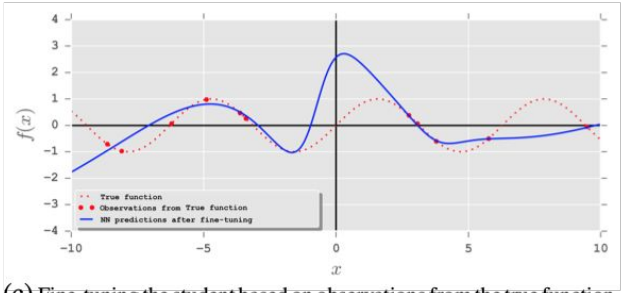
Fidelity Weighted Learning



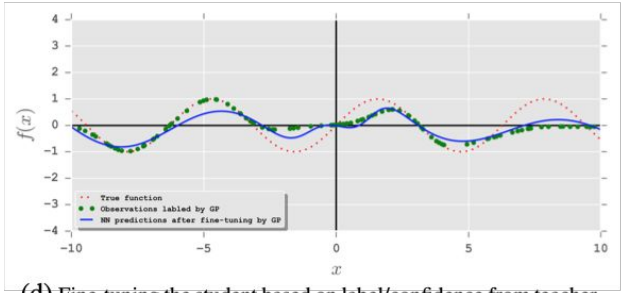
(a) Training student on 100 examples from the weak function.



(b) Fitting teacher based on 10 observations from the true function.



(c) Fine-tuning the student based on observations from the true function.

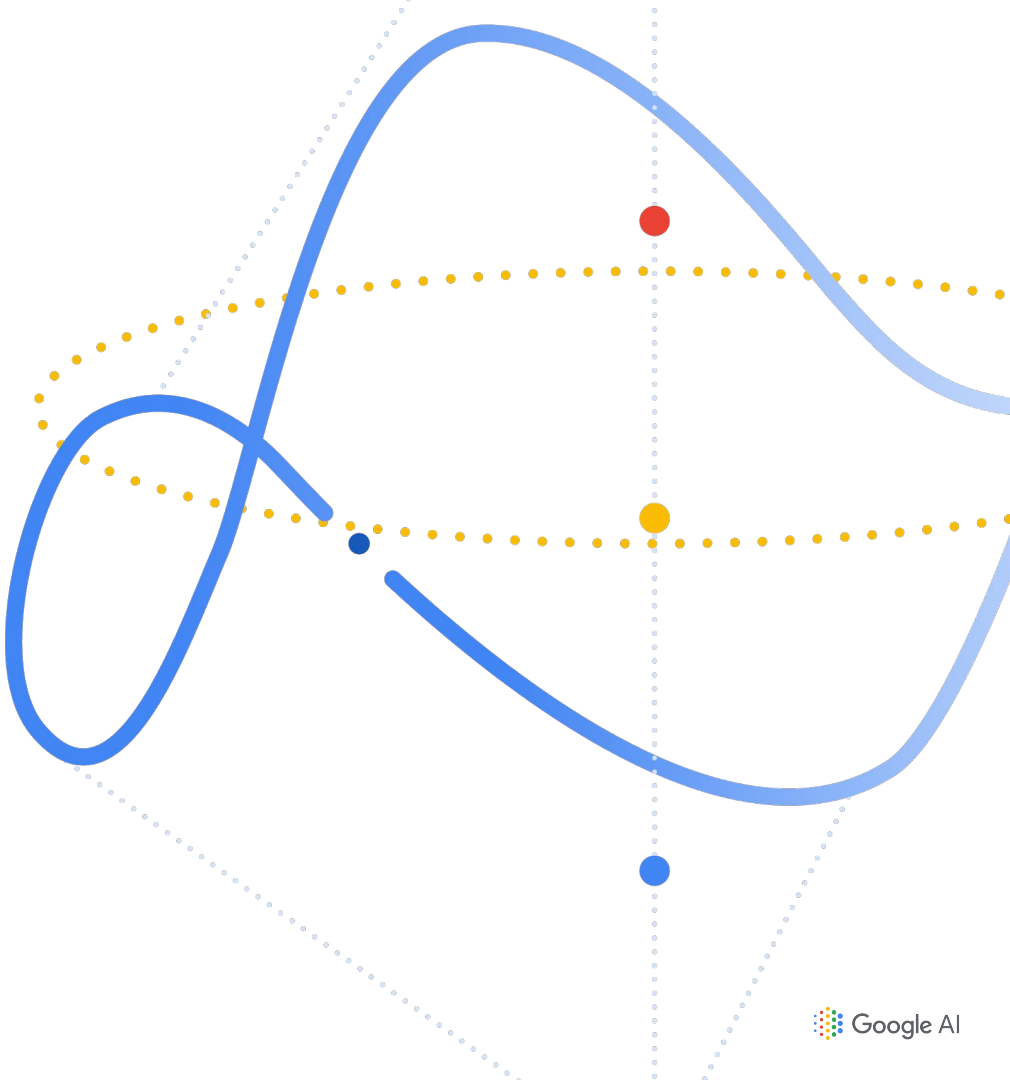


(d) Fine-tuning the student based on label/confidence from teacher.

Dehghani, M., Mehrjou, A., Gouws, S., Kamps, J., and Schölkopf,B. (2018). Fidelity-weighted learning. In International Conference on Learning Representations (ICLR)

Google AI
Brain Team

Injecting Inductive Biases for Data efficiency



Google AI

Inductive Biases

- **Inductive biases**:
  - Factors that lead a learner to favor one hypothesis over another that are **independent of the observed data**.
  - Great ways for encoding modeling assumptions

Mitchell, Tom M. The need for biases in learning generalizations. New Jersey: Department of Computer Science, Laboratory for Computer Science Research, Rutgers Univ., 1980.

Example: Invariance / Equivariance

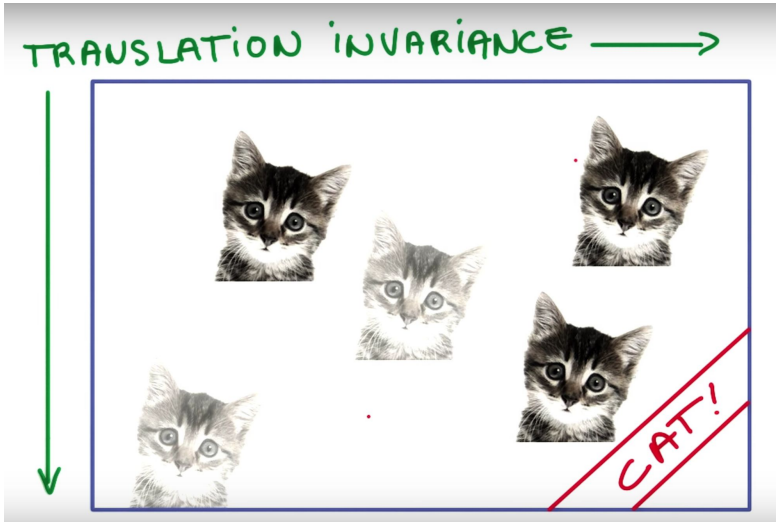


Image source: https://www.cc.gatech.edu/~san37/post/dlhc-cnn/

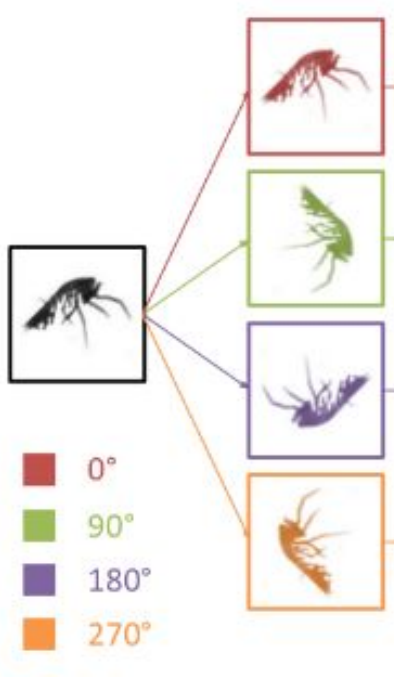
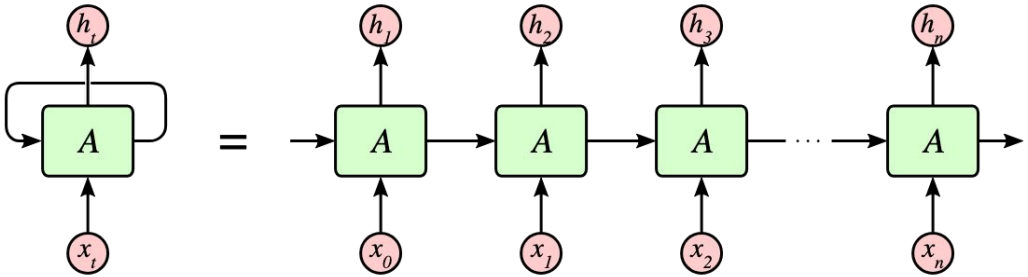


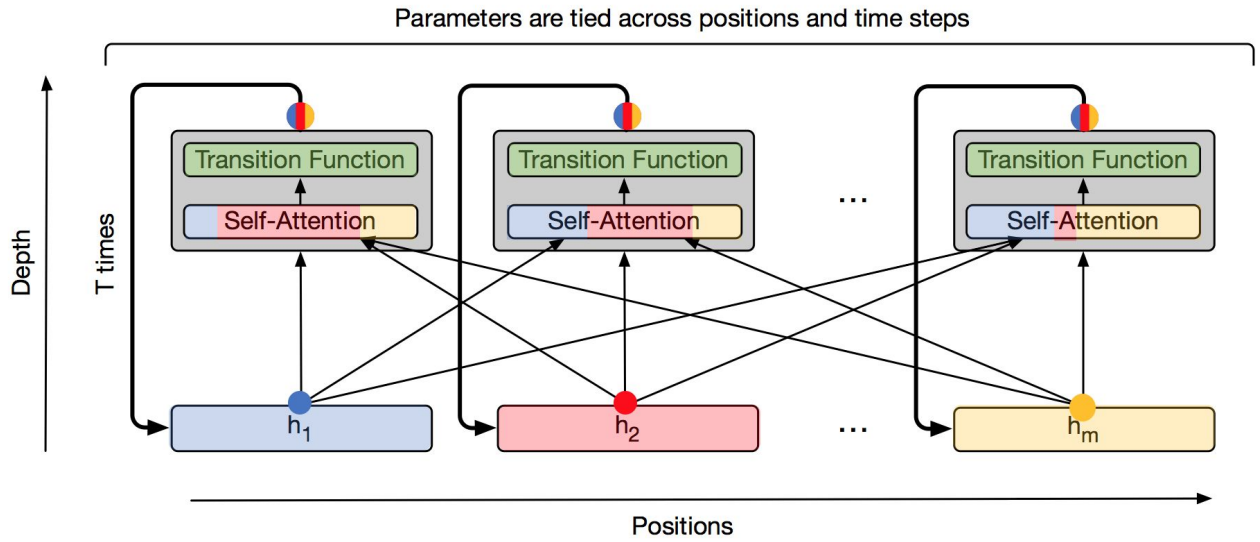
Image source: https://bigsnarf.wordpress.com/2017/01/27/cnn-image-rotationinvariance/

Recurrent Inductive Bias for Sequence Modeling

- **Sequential Processing:**
  - "**Re-occurrence**" of referring back to all previous internal states.
    - In each step, we can run the same function that process the input taking the previous inputs into account.



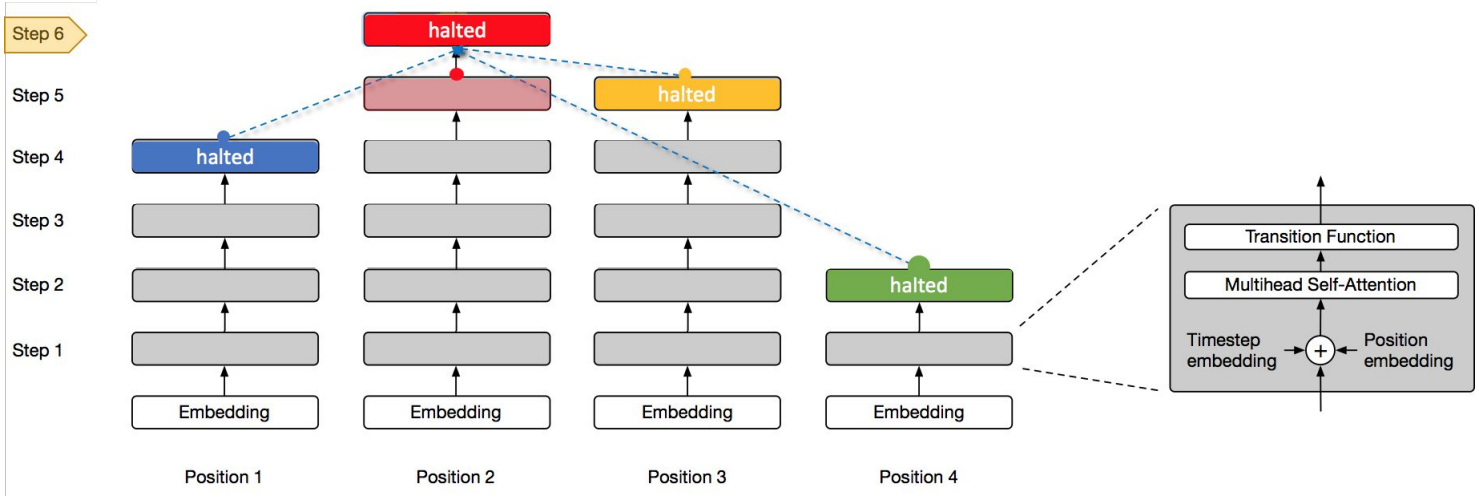
Universal Transformer: Recurrence in Depth



Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser,Ł. (2019b). Universal transformers. In International Conference on Learning Representations (ICLR)
Dehghani, M., Azarbonyad, H., Kamps, J., and de Rijke, M.(2019a). Learning to transform, combine, and reason in open-domain question answering. InProceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM'19

Google AI
Brain Team

Dynamic Halting

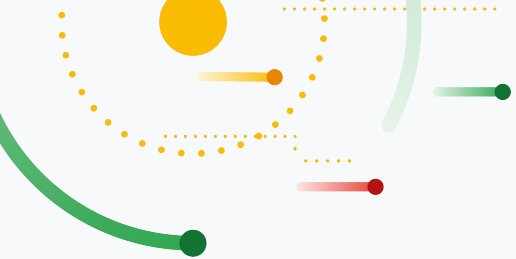


Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser,Ł. (2019b). Universal transformers. In International Conference on Learning Representations (ICLR)

Universal Transformer: Recurrence in Depth

- **Weight sharing**: Following intuitions behind weight sharing found in CNNs and RNNs
    - Strikes an effective balance between inductive bias and model expressivity
- **Conditional computation**: Equipping the Universal Transformer with the ability to halt or continue computation.
    - More computations for more complex inputs

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Kaiser,Ł. (2019b). Universal transformers. In International Conference on Learning Representations (ICLR)



Google AI
Brain Team

Thank you!