Meta-Learning in Computer Vision and NLP

Connections to Transfer Learning and Multitasking

Nitish Shirish Keskar

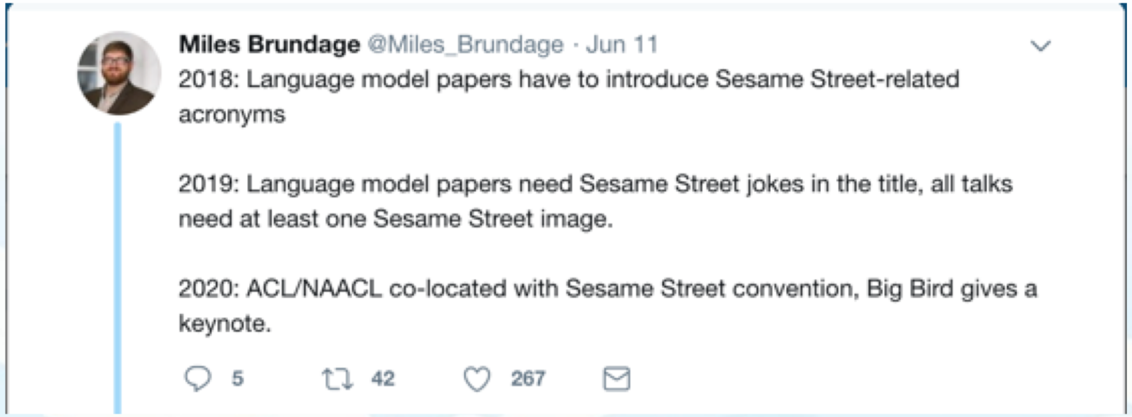
keskarnitish.github.io, @StrongDuality

Outline

- My (constrained) definition of meta-learning:
  - Efficient adaptation
  - Efficient assimilation, and
  - Efficient zero-shot learning.
- References from Language and Vision.

Why Talk About NLP?

- Celebrate the similarities and differences between the two modalities
- Underscores the generality of some of the common ingredients
- Differences:
  - Sequential v/s non-sequential data
  - Input and output spaces; e.g., image class vs free-form natural language
  - Intensity and type of task bias
- Useful for anyone building multi-modal systems

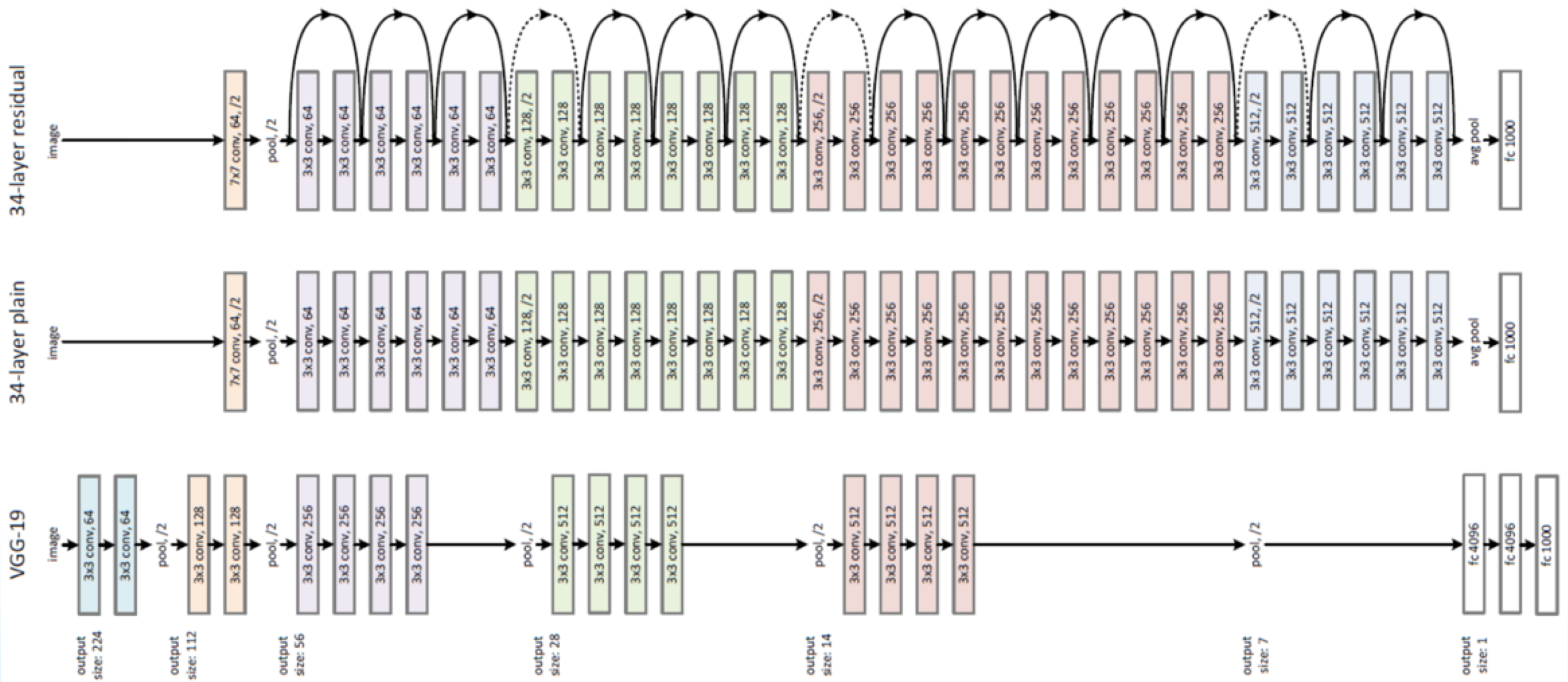


Preliminaries

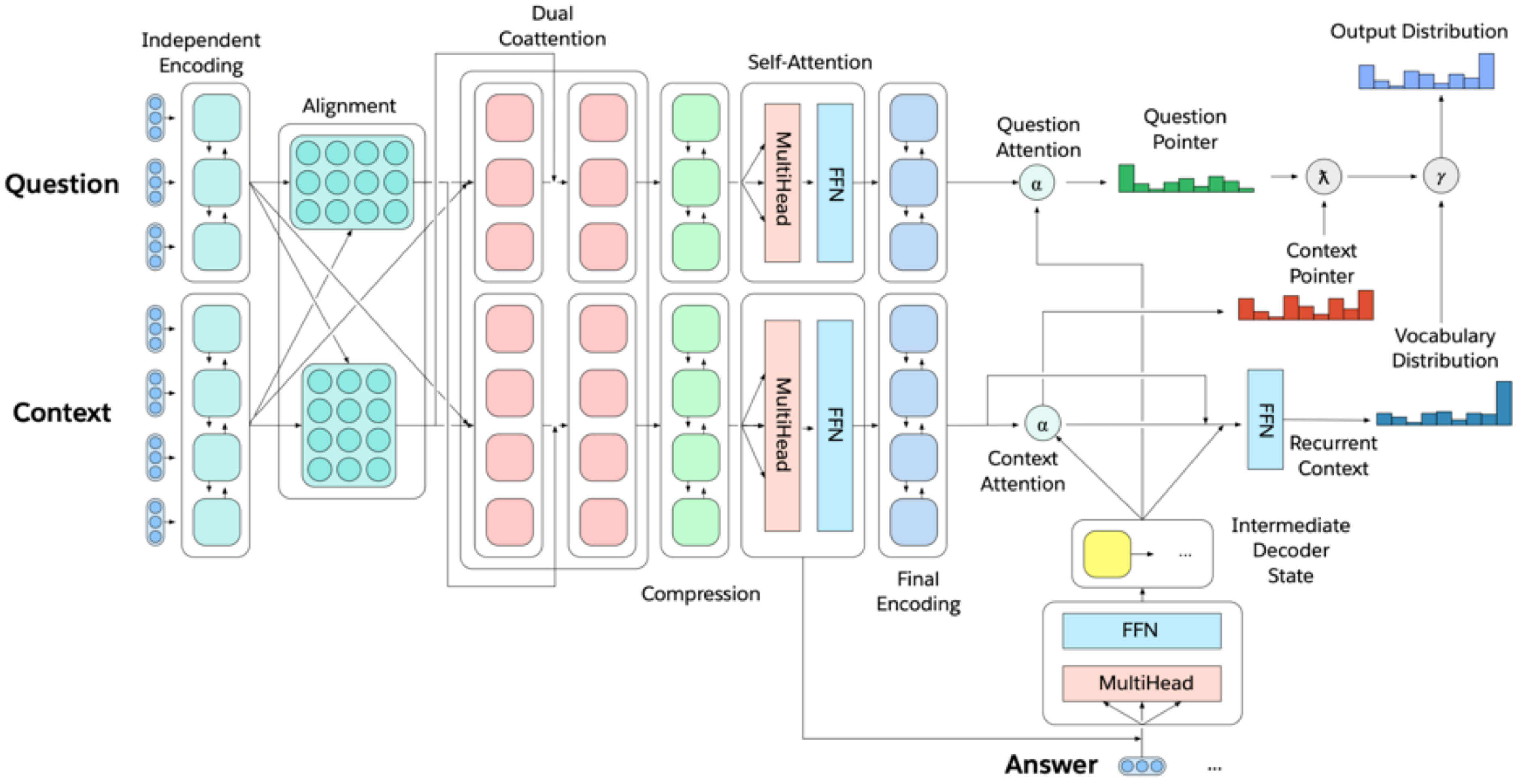
Def.: Meta-Learning

- Efficient adaptation to new tasks
  - A network is available from which it is easy to adapt from
- Efficient assimilation to new tasks
  - A network is available to which new tasks are added
- Efficient zero-shot learning of new tasks
  - We have the ability to perform well on tasks without any labeled data

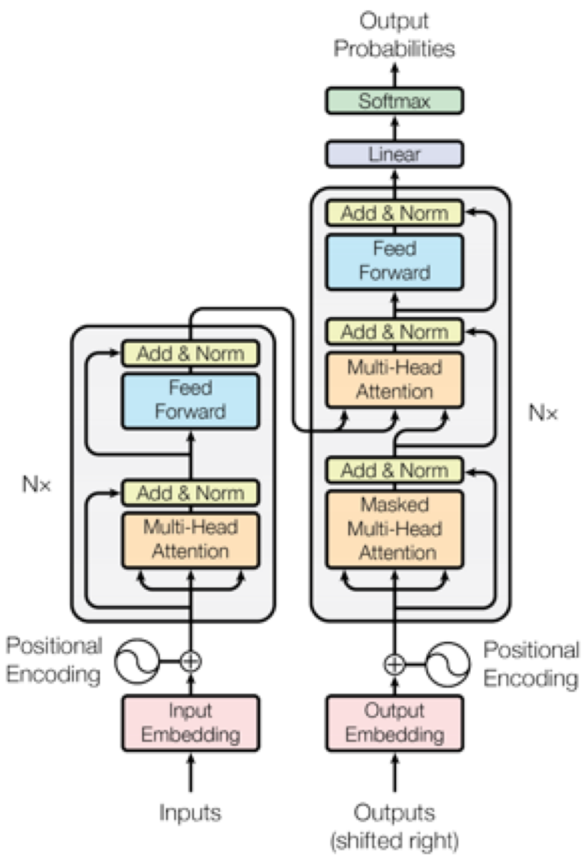
Typical CV Pipeline

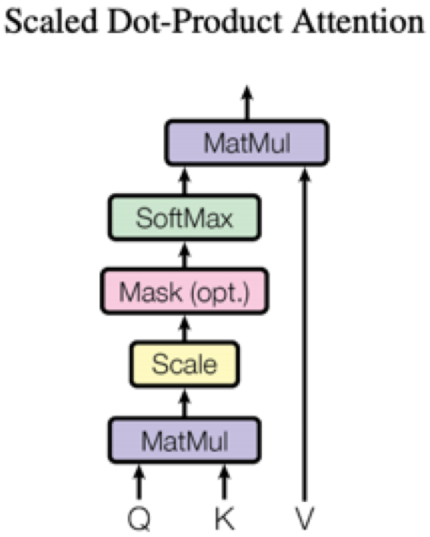


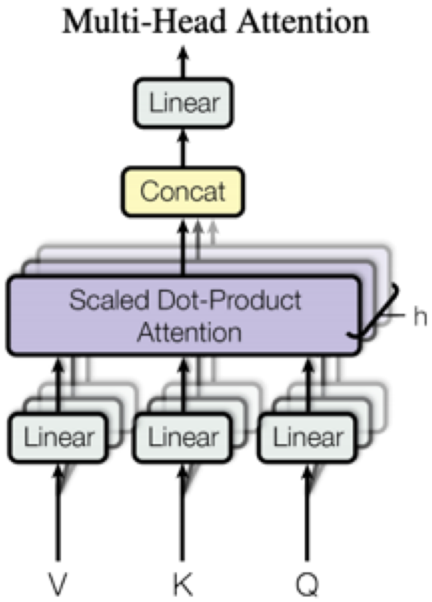
Typical NLP Pipeline



Transformers







$$\mathrm{Attention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Training Process

Single-Task

- Sample a mini-batch
- Compute the gradient of the loss function
- Take a step with your favorite optimizer

Multitasking

- Sample a mini-batch
  - Mix all data or **keep separate**?
  - Mini-batch **filled with one task** or proportion?
  - **Oversample** smaller datasets or not?
- Compute the gradient of the loss function
  - How do you balance multiple task objectives —**equal** or weighted?
- Take a step with your favorite optimizer
  - **One optimizer** for all or separate?

Efficient Adaptation

What We Desire

- Train a network on *task(s)* **[Phase I]** such that it adapts quickly to new domains, or new tasks altogether **[Phase II]**.
- Phase I does not require us to know downstream tasks
- Phase I is *scalable*:
  - Data
  - Compute
- Adaptation is *beneficial*:
  - Learning outcomes better with Phase I than without
- Adaptation is *efficient:*
  - Amount of data needed for Phase II is reduced
  - Computational effort for Phase II is low

Solution 1 – Pre-Training

Pre-Training with a Relevant Task

- Pre-train model on a relevant task on a large amount of data
- Doesn't have to be supervised!
- Scalable
- Beneficial
- Efficient

In Vision

- Train on a large dataset (e.g., ImageNet); transfer representations.
- Either fine-tune bottom layers, or keep fixed.
- Unsupervised — VAEs, GANs
- Two recent results:
  - Exploring the Limits of Weakly Supervised Pretraining
  - Do Better ImageNet Models Transfer Better?

tl;dr

Exploring the Limits of Weakly Supervised Pretraining

Dhruv Mahajan Ross Girshick Vignesh Ramanathan Kaiming He
Manohar Paluri Yixuan Li Ashwin Bharambe Laurens van der Maaten

Facebook

- Hashtag prediction on *billions* of images.
- Transfer (to ImageNet) continues to improve with size of dataset & accuracy on pre-trained task
- Almost as important is the matching of label spaces; label-engineering?
- For pre-training: label noise matters; but not as much as we fear. Emphasis on more data even if little noisy.
- More data needs more capacity; difference can be significant.

tl;dr

Do Better ImageNet Models Transfer Better?

Simon Kornblith*, Jonathon Shlens, and Quoc V. Le
Google Brain
{skornblith,shlens,qvl}@google.com

There is a strong correlation between transferability and accuracy on ImageNet

However, sensitive to the way ImageNet is trained

- Regularizers that improve ImageNet hurt transfer — label smoothing, dropout, auxiliary classifier heads, and scale parameters in BatchNorm.

ImageNet features may not be as general as believed

- On some fine-grained classification tasks, ImageNet fine-tuning is no better than random
- However, architectures that do well on ImageNet do transfer

Fine-tuning continues to be better than feature extraction; especially for domain mismatch

In NLP – Supervised Machine Translation

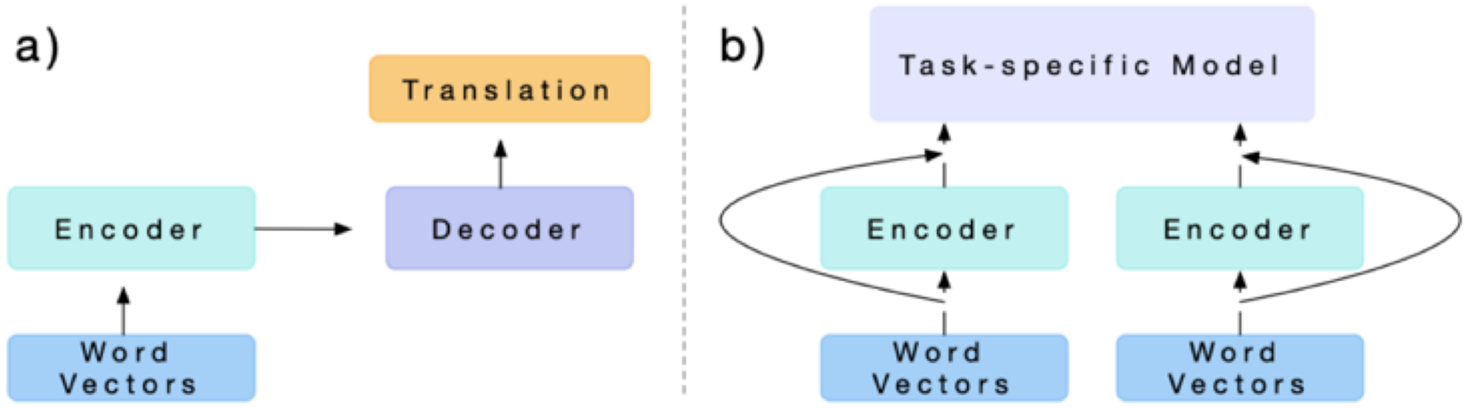
Learned in Translation: Contextualized Word Vectors

Bryan McCann
bmccann@salesforce.com

James Bradbury
james.bradbury@salesforce.com

Caiming Xiong
cxiong@salesforce.com

Richard Socher
rsocher@salesforce.com



In NLP — Unsupervised

Deep contextualized word representations

Matthew E. Peters[†], **Mark Neumann**[†], **Mohit Iyyer**[†], **Matt Gardner**[†],
{matthewp,markn,mohiti,mattg}@allenai.org

Christopher Clark[*], **Kenton Lee**[*], **Luke Zettlemoyer**[†*]
{csquared,kentonl,lsz}@cs.washington.edu

[†]Allen Institute for Artificial Intelligence
[*]Paul G. Allen School of Computer Science & Engineering, University of Washington

Improving Language Understanding by Generative Pre-Training

Alec Radford OpenAI alec@openai.com
Karthik Narasimhan OpenAI karthikn@openai.com
Tim Salimans OpenAI tim@openai.com
Ilya Sutskever OpenAI ilyasu@openai.com

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com

Universal Language Model Fine-tuning for Text Classification

Jeremy Howard[*]
fast.ai
University of San Francisco
j@fast.ai

Sebastian Ruder[*]
Insight Centre, NUI Galway
Aylien Ltd., Dublin
sebastian@ruder.io

Language Modeling

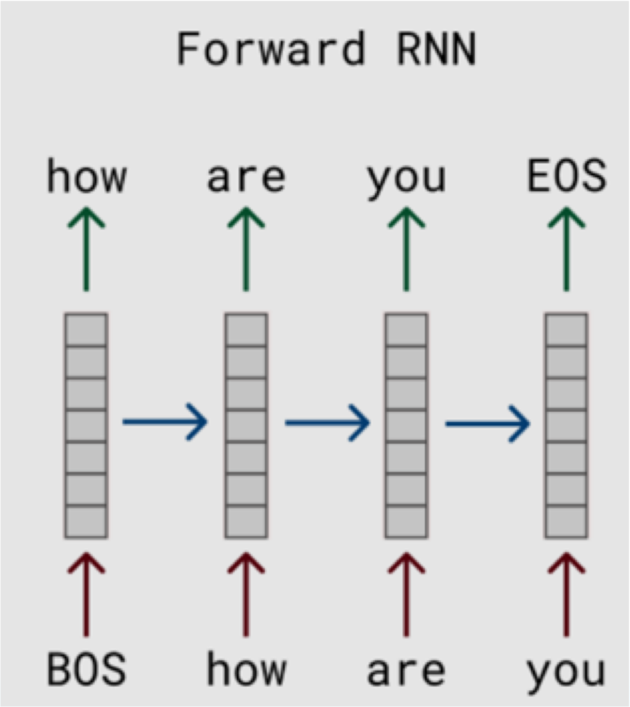
Not an exhaustive list. See (Mansimov, 19) arxiv::1905.12790 for more details.

Causal:

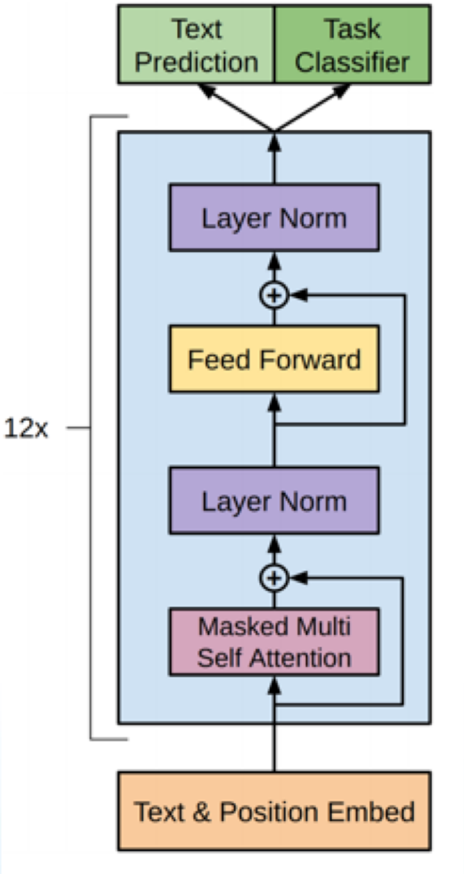
- The quick brown fox jumped over the ?
- Per-token classification problem –
  - Given a sequence length of N ; N prediction problems
  - The -> quick
  - quick -> brown …
  - the -> lazy

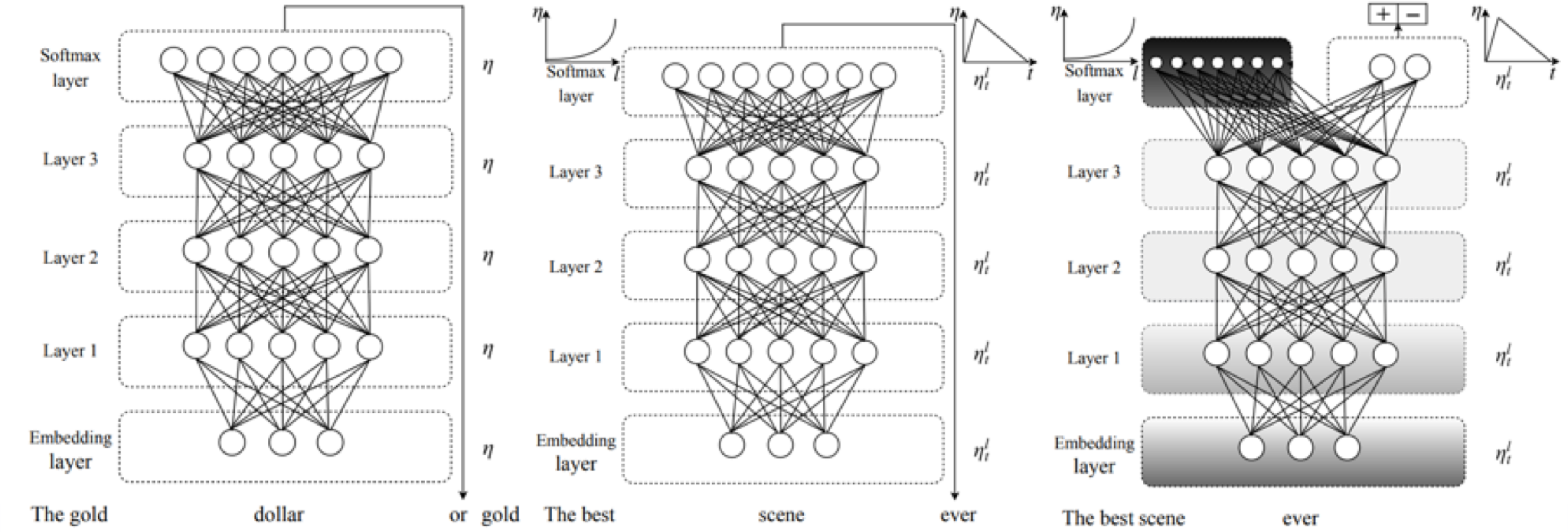
Masked:

- The <?> brown fox jumped over the <?> dog.
- Similar setup as before.



In NLP - Unsupervised



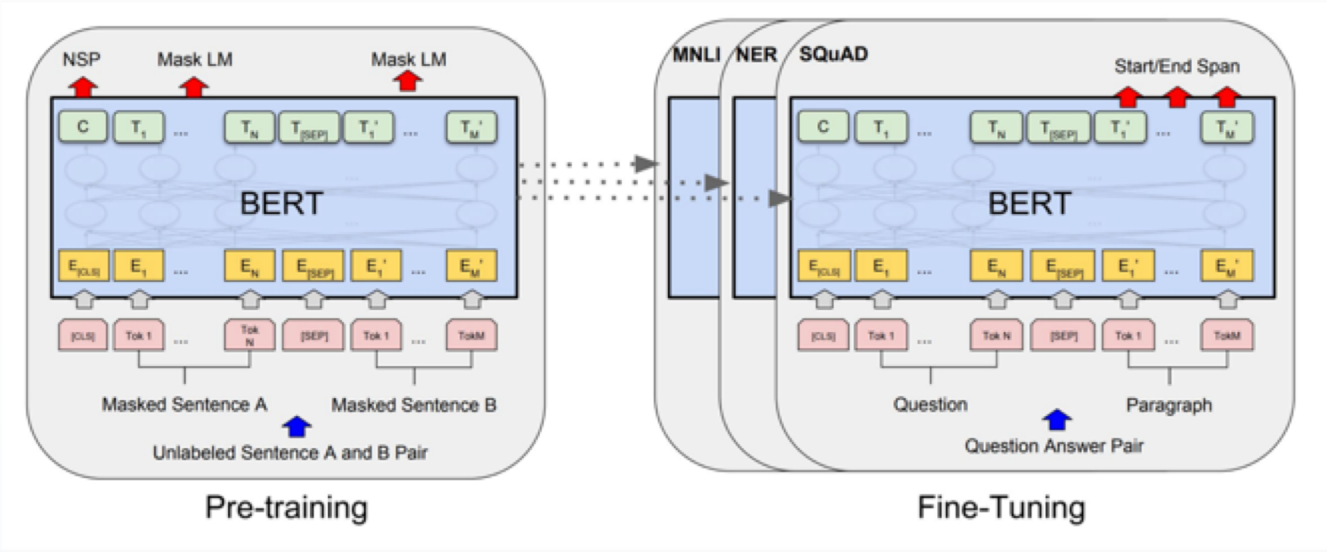


(a) LM pre-training (b) LM fine-tuning (c) Classifier fine-tuning

In NLP — BERT

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin **Ming-Wei Chang** **Kenton Lee** **Kristina Toutanova**
Google AI Language
{jacobdevlin,mingweichang,kentonl,kristout}@google.com



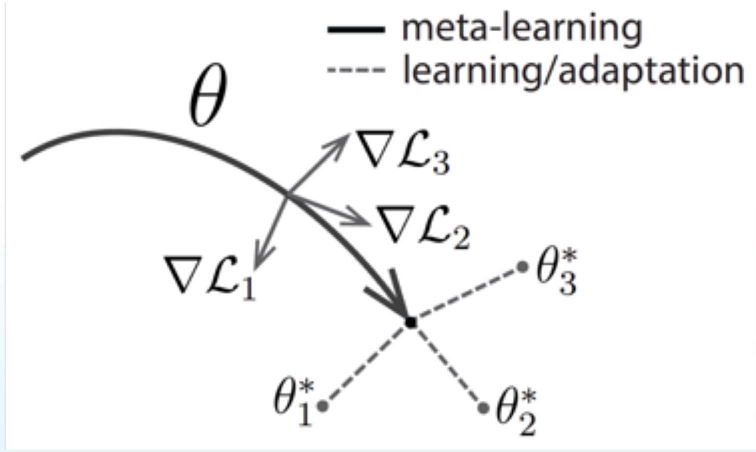
Solution 2 – MAML

Model Agnostic Meta-Learning (MAML)

Model-Agnostic Meta-Learning for Fast Adaptation of Deep Networks

Chelsea Finn [1] **Pieter Abbeel** [1,2] **Sergey Levine** [1]

- Intentionally train the network to be a good adaptor
- Scalable
- Beneficial
- Efficient



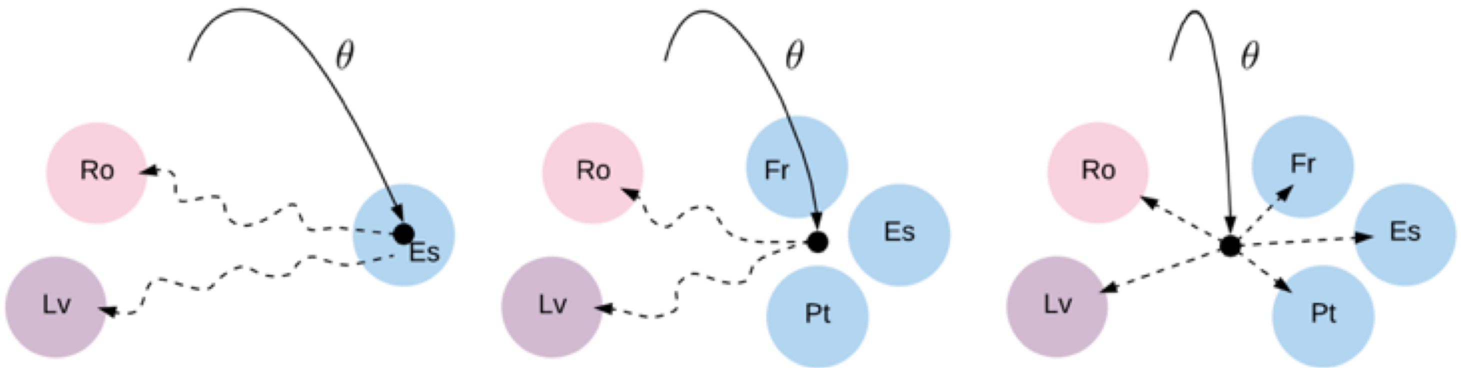
Focus is less on a real learnt task

In NLP — Low Resource MT

Meta-Learning for Low-Resource Neural Machine Translation

Jiatao Gu*[†], Yong Wang*[†], Yun Chen[†], Kyunghyun Cho[‡] and Victor O.K. Li[†]

[†]The University of Hong Kong
[‡]New York University, CIFAR Azrieli Global Scholar
[†]{jiataogu, wangyong, vli}@eee.hku.hk
[†]yun.chencreek@gmail.com
[‡]kyunghyun.cho@nyu.edu



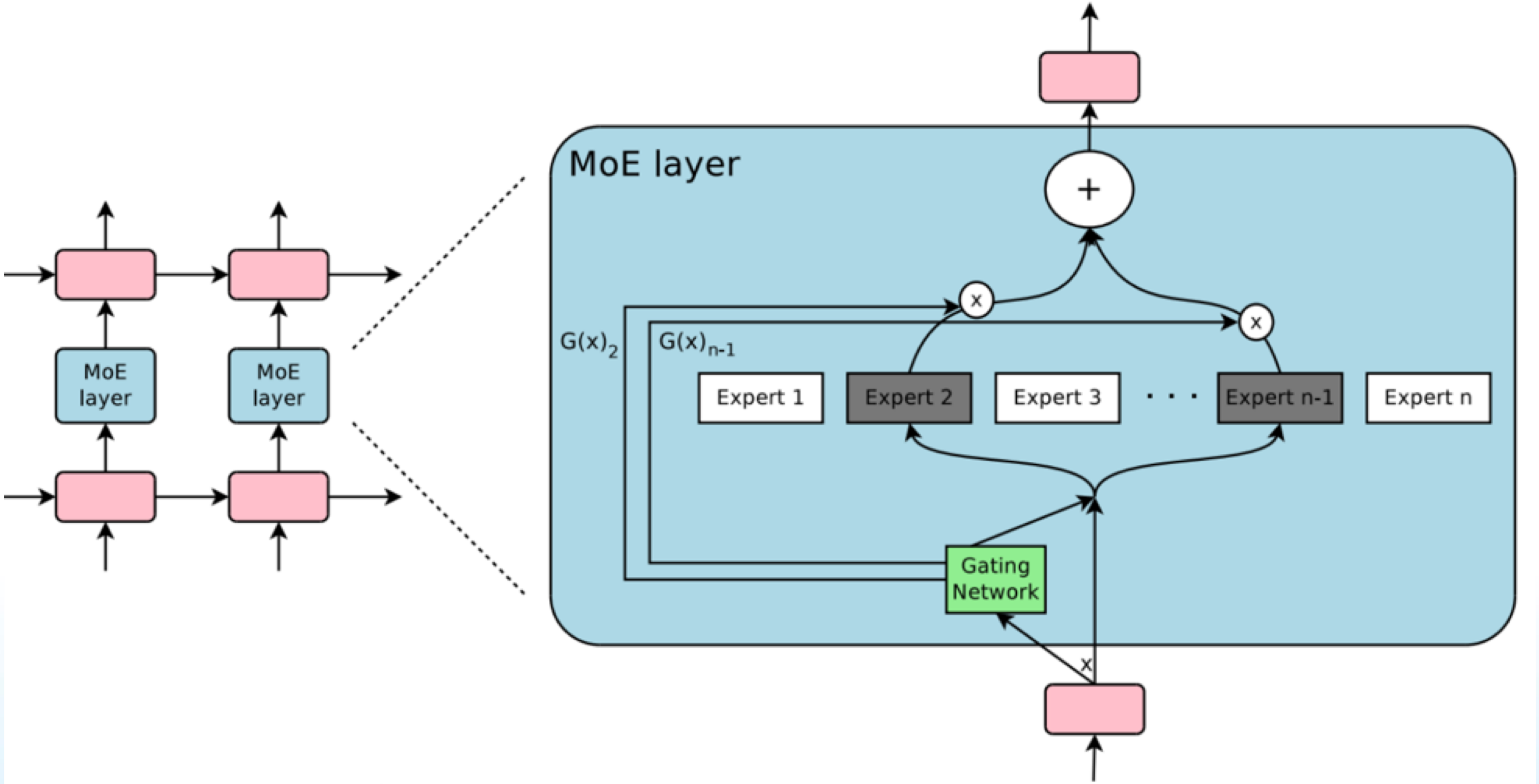
(a) Transfer Learning (b) Multilingual Transfer Learning (c) Meta Learning

Efficient Assimilation

What We Desire

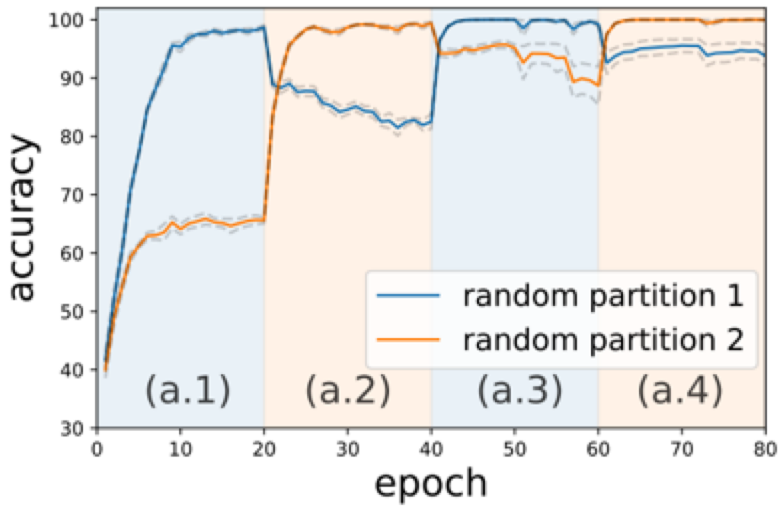
- Train a network on task(s), when new task(s) is presented, the model adapts to perform well on new tasks **AND** maintains performance on old ones.
- Performance on old tasks is at least as good as before assimilation.
- Performance on new tasks is at least as good as them being trained in isolation.
- Assimilation is beneficial:
  - A sizable fraction of tasks benefit from assimilation over their individual models.
- Assimilation is *efficient*:
  - Speed of learning is not negatively impacted for new tasks
  - Capacity of the model does not increase linearly with added tasks

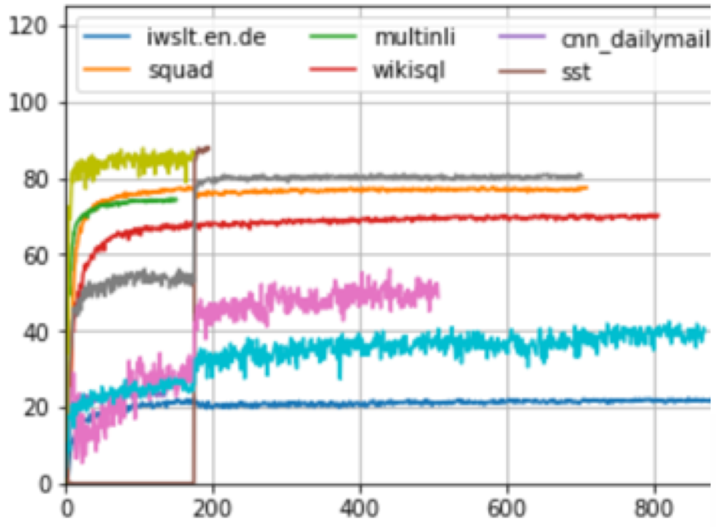
Solution 1: Mixture of Experts



Solution 2: Continual Learning

- Recall:
  - For multitasking, round-robin through all tasks, one mini-batch at-a-time, is a strong baseline.
  - Requires all tasks to be present a-priori.
- If new task appears, pretend it was always around — Simply add it to the round-robin list.
- Very strong baseline.

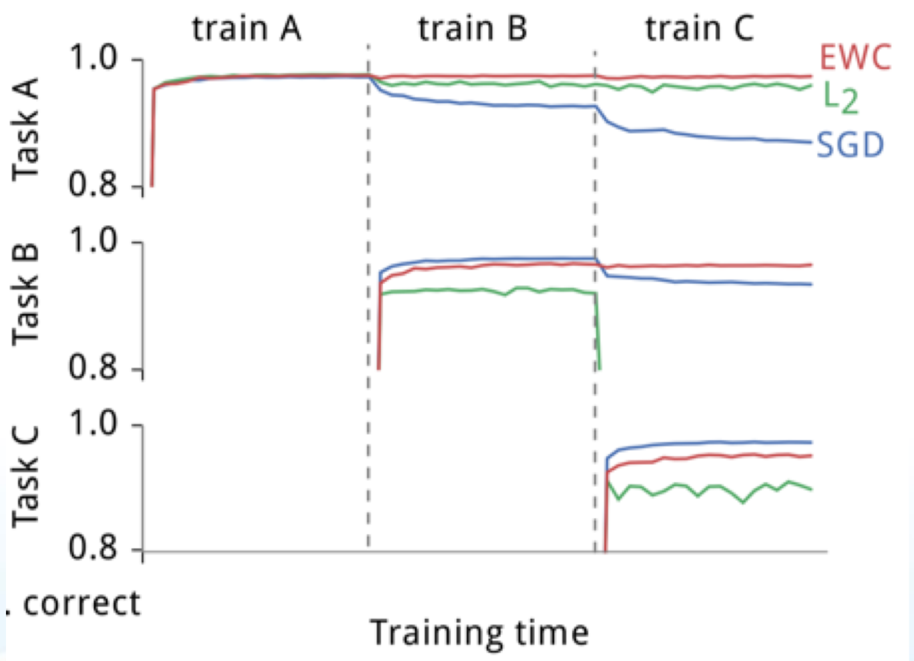




- Requires data from all tasks to still be available.

Solution 3: Catastrophic Forgetting

- Catastrophic Forgetting: A network trained with only task A and then trained only with task B tends to forget task A rapidly.
- (Almost as-if) network weights over-written rather than gracefully changed.
- Bad! Want to keep performance on task A.
- Solution: encourage grace in parameter changes.
- Elastic Weight Consolidation (EWC) & Beyond

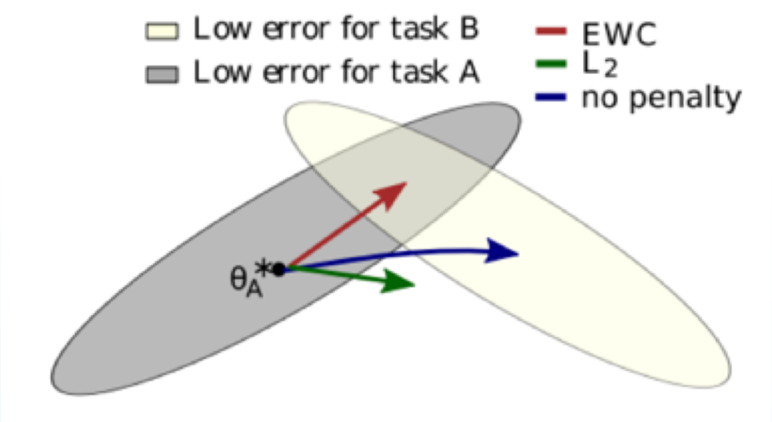


Solution 3: Catastrophic Forgetting

Overcoming catastrophic forgetting in neural networks

James Kirkpatrick[a], Razvan Pascanu[a], Neil Rabinowitz[a], Joel Veness[a], Guillaume Desjardins[a], Andrei A. Rusu[a], Kieran Milan[a], John Quan[a], Tiago Ramalho[a], Agnieszka Grabska-Barwinska [a], Demis Hassabis[a], Claudia Clopath[b], Dharshan Kumaran[a], and Raia Hadsell[a]

[a]DeepMind, London, N1C 4AG, United Kingdom
[b]Bioengineering department, Imperial College London, SW7 2AZ, London, United Kingdom



$$\mathcal{L}(\theta) = \mathcal{L}_B(\theta) + \sum_i \frac{\lambda}{2} F_i (\theta_i - \theta^*_{A,i})^2$$

Solution 4: Adapters

Learning multiple visual domains with residual adapters

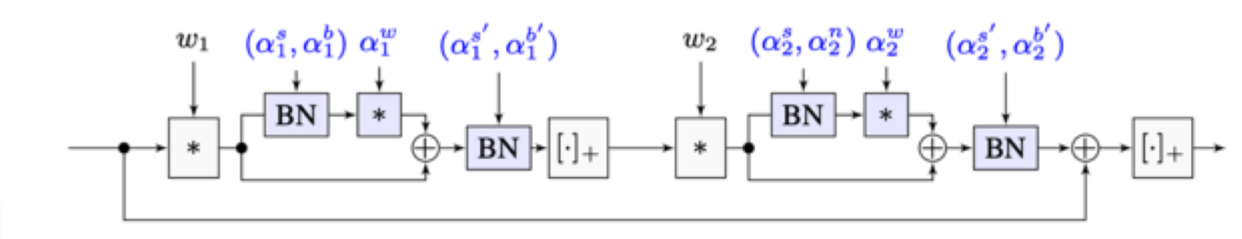
Sylvestre-Alvise Rebuffi[1] **Hakan Bilen**[1,2] **Andrea Vedaldi**[1]

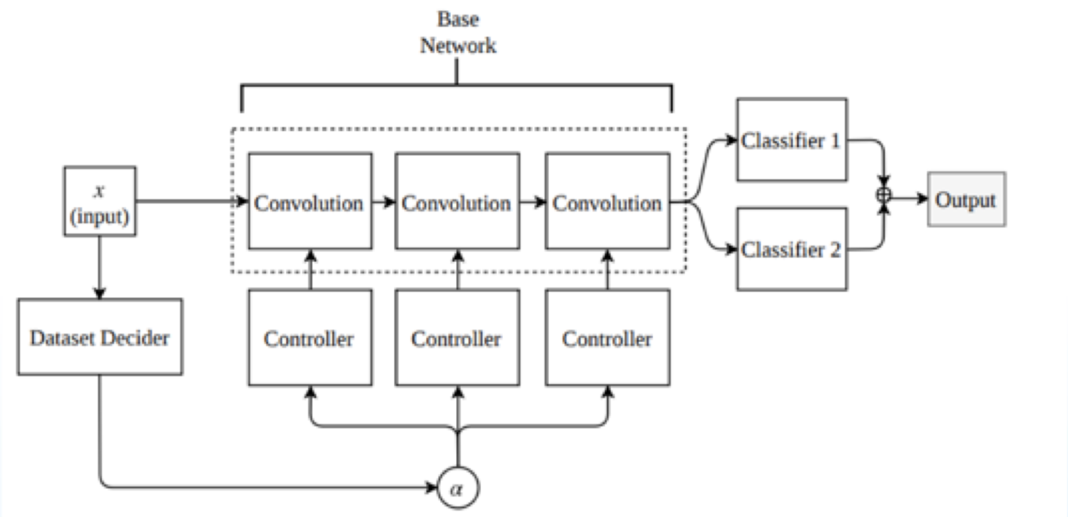
[1] Visual Geometry Group
University of Oxford
{srebuffi,hbilen,vedaldi}@robots.ox.ac.uk

[2] School of Informatics
University of Edinburgh

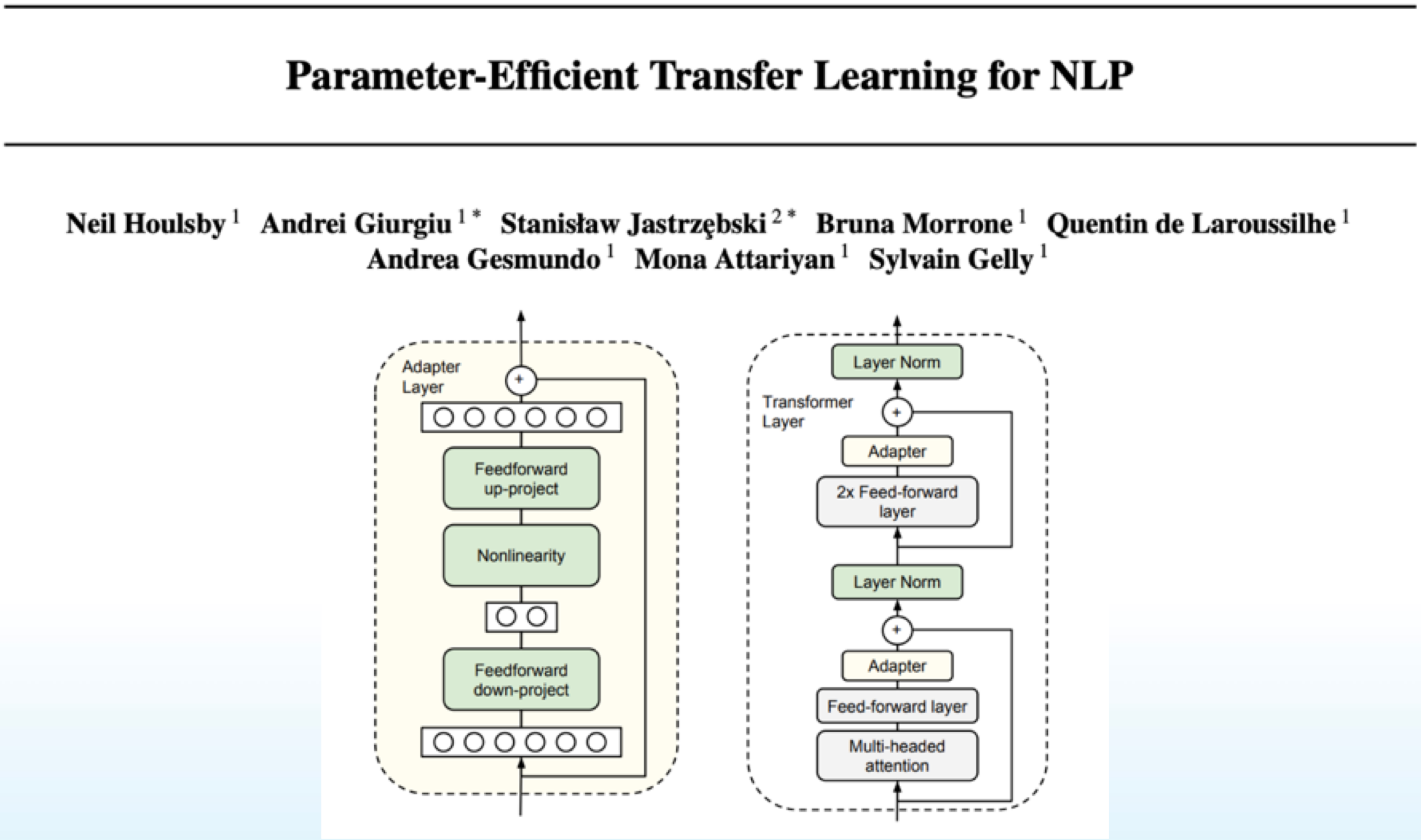
Incremental Learning Through Deep Adaptation

Amir Rosenfeld John K. Tsotsos
Department of Electrical Engineering and Computer Science
York University, Toronto, ON, Canada
amir@eecs.yorku.ca,tsotsos@cse.yorku.ca





Solution 4: Adapters

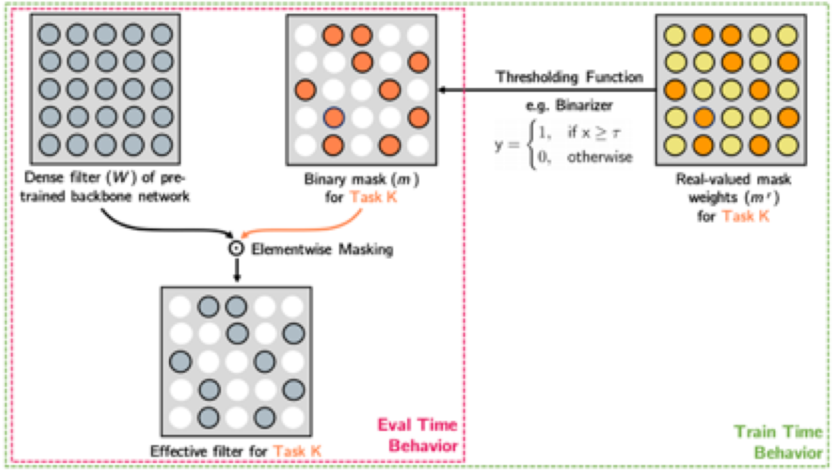


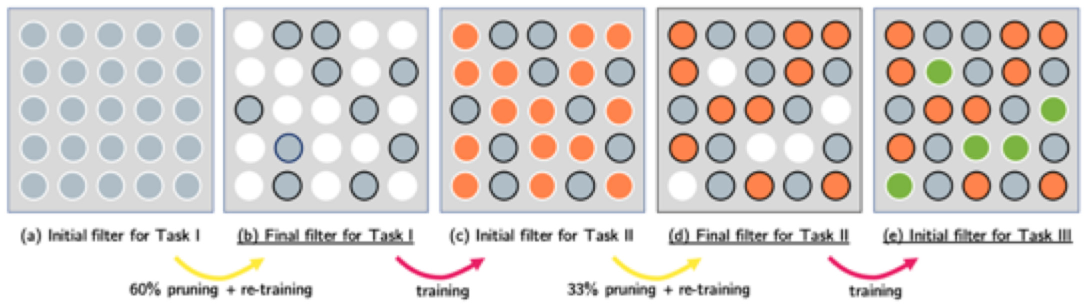
Solution 5: Masking

Piggyback: Adapting a Single Network to Multiple Tasks by Learning to Mask Weights

Arun Mallya, Dillon Davis, Svetlana Lazebnik

University of Illinois at Urbana-Champaign





Many, Many Other Approaches

- Learning without Forgetting (LwF)
- PathNet
- GeppNet
- Fixed Expansion Layer (FEL)
- FearNet
- Incremental Class Learning
- Pseudo-replay/rehearsal
- Gradient Episodic Memory
- Incremental Moment Matching
- Architecture Search

Efficient Zero-Shot Learning

What We Desire

- Learning without labels — rely on descriptions of the classes instead
- Ability to provide descriptions in an *intuitive* manner (e.g., natural language or attributes).
- Efficient use of terse descriptions

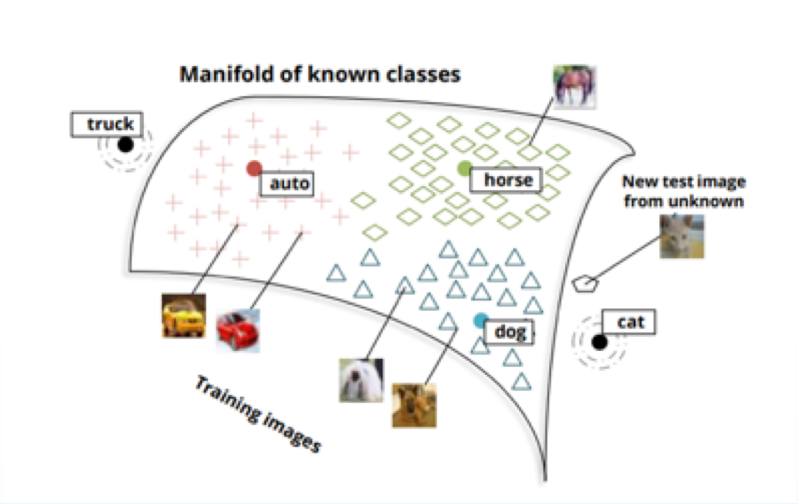
In Vision

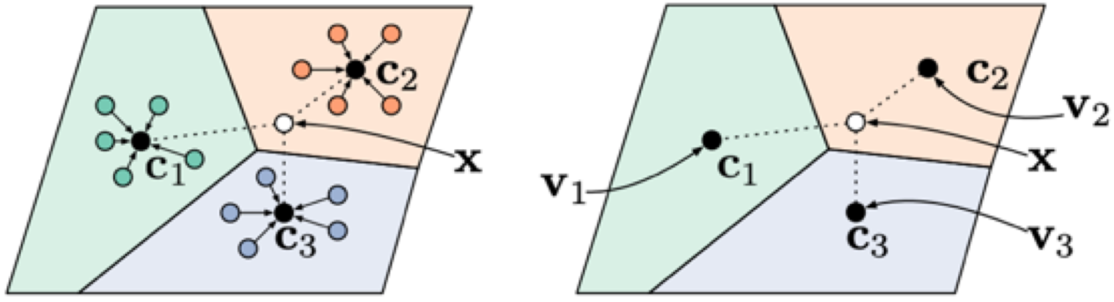
Zero-Shot Learning Through Cross-Modal Transfer

Richard Socher, Milind Ganjoo, Christopher D. Manning, Andrew Y. Ng
Computer Science Department, Stanford University, Stanford, CA 94305, USA
richard@socher.org, {mganjoo, manning}@stanford.edu, ang@cs.stanford.edu

Prototypical Networks for Few-shot Learning

Jake Snell, University of Toronto*
Kevin Swersky, Twitter
Richard S. Zemel, University of Toronto, Vector Institute



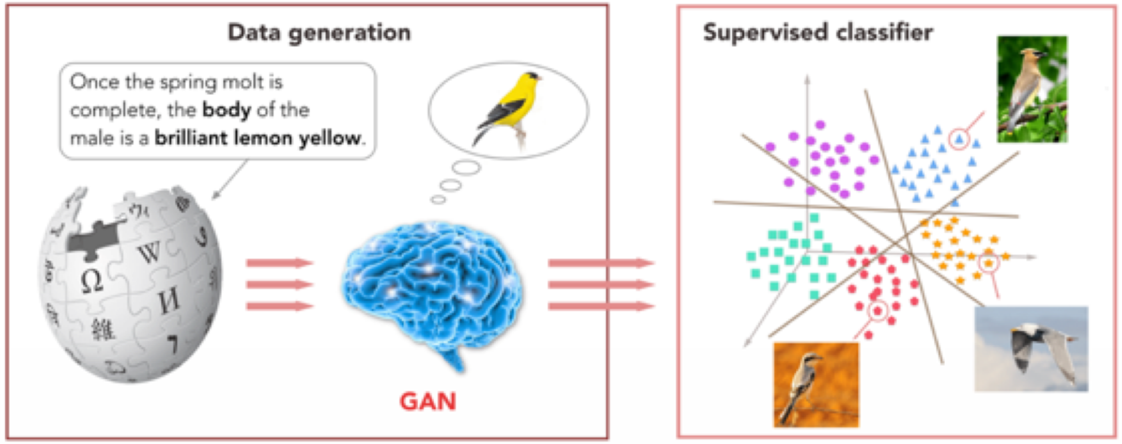


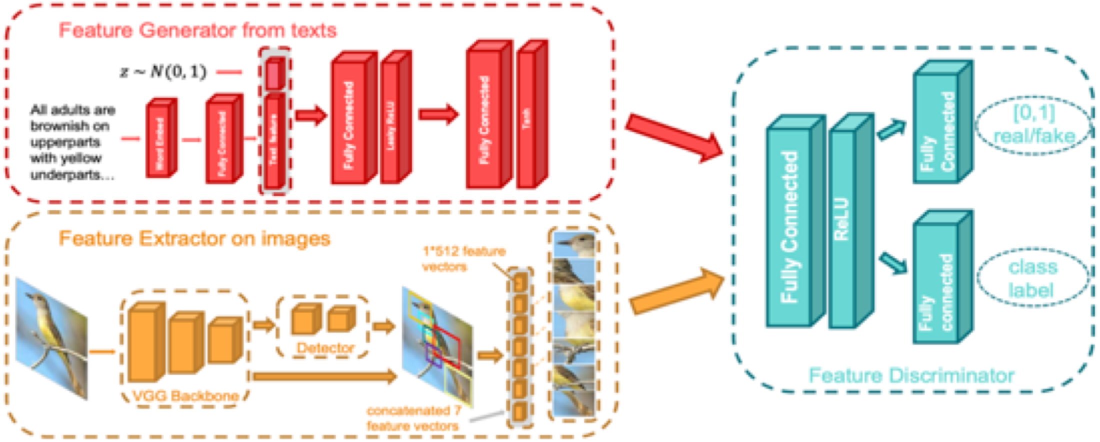
(a) Few-shot (b) Zero-shot

In Vision

A Generative Adversarial Approach for Zero-Shot Learning from Noisy Texts

Yizhe Zhu[1], Mohamed Elhoseiny[2], Bingchen Liu[1], Xi Peng[1] and Ahmed Elgammal[1]
yizhe.zhu@rutgers.edu, elhoseiny@fb.com,
{bingchen.liu, xipeng.cs}@rutgers.edu, elgammal@cs.rutgers.edu
[1]Rutgers University, Department of Computer Science, [2] Facebook AI Research

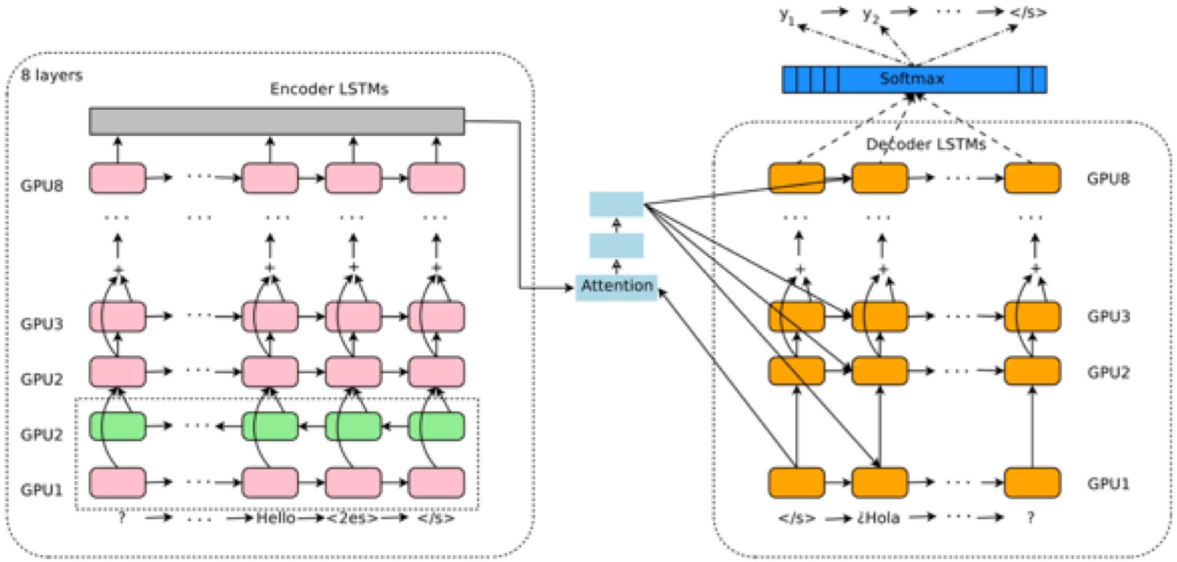




In NLP – Machine Translation

Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat
melvinp,schuster,qvl,krikun,yonghui,zhifengc,nsthorat@google.com
Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, Jeffrey Dean

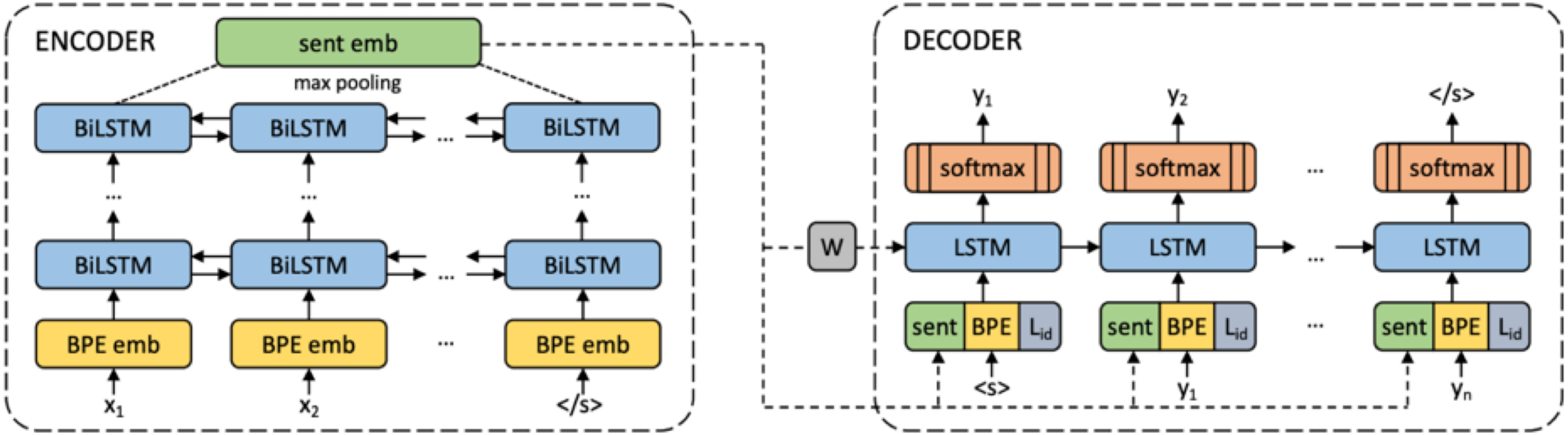


In NLP – Cross-Linguality

Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond

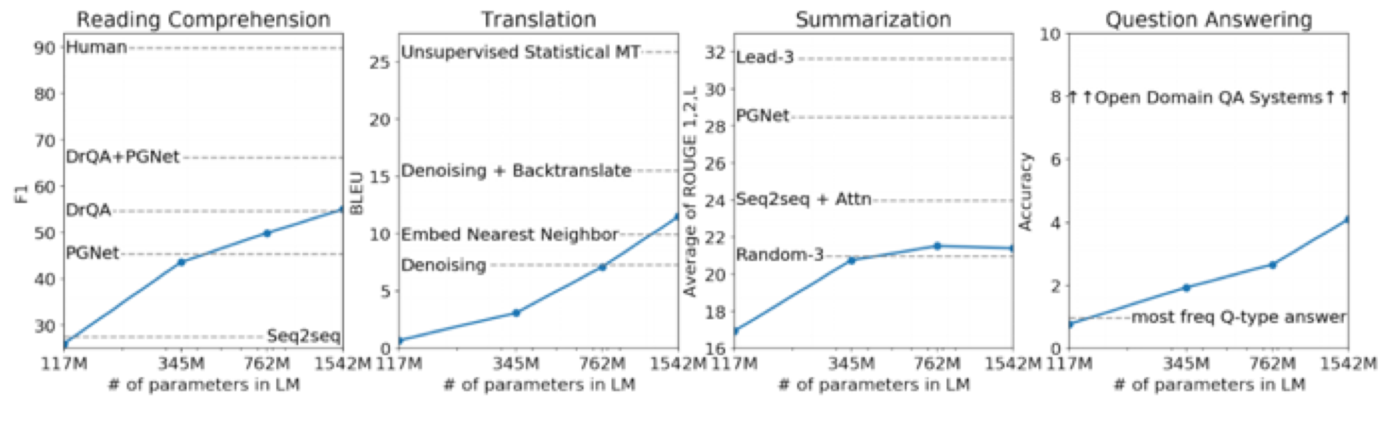
Mikel Artetxe
University of the Basque Country (UPV/EHU)*
mikel.artetxe@ehu.eus

Holger Schwenk
Facebook AI Research
schwenk@fb.com



In NLP – decaNLP and GPT-2

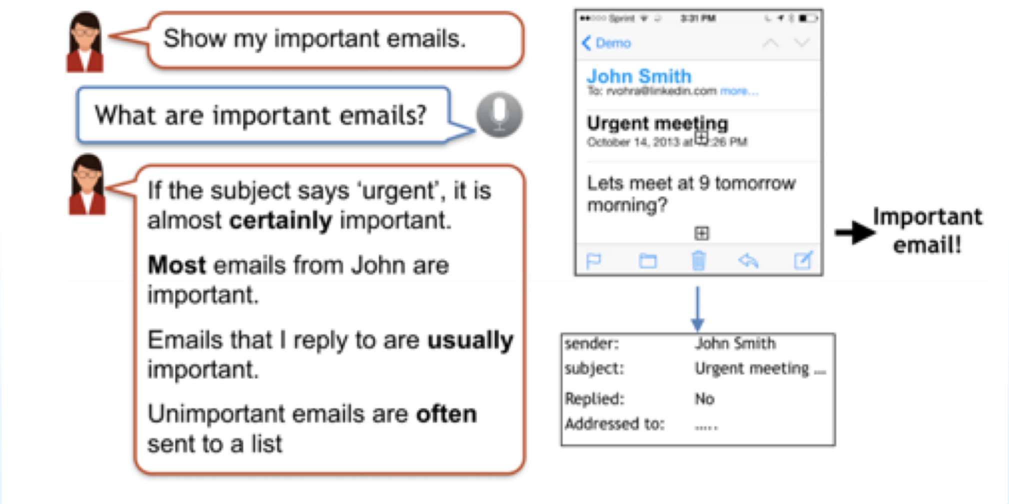
- decaNLP:
  - Trained on 10 NLP tasks jointly.
  - Has seen span-extractive question answering & sentiment analysis
  - Can reasonably answer queries like:
    - John gave a talk but no one clapped. Would John be happy or **sad**?
- GPT-2
  - Trained on a large amount of unsupervised language modeling data
  - Can zero-shot on several tasks



In NLP – New Classification Tasks

Zero-shot Learning of Classifiers from Natural Language Quantification

Shashank Srivastava Igor Labutov Tom Mitchell
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
ssrivastava@cmu.edu ilabutov@cs.cmu.edu tom.mitchell@cmu.edu



Open Questions

Adaptation:

- Robust adaptation on new tasks with limited training data (or which guides data collection)
- Adaptation to more difficult tasks. <Muppet> for multi-document multi-lingual video captioning

Assimilation:

- Still a long way to go…
- Have to choose between desiderata; not possible to satisfy them all

Zero-Shot Learning:

- Again, long way to go.
- Humans can do this: describe the task in natural language and do it!
- Q: Describe why Adam is a good optimizer.
  - R: Mark as correct if answer talks about adapting to curvature, using moments or momentum, and not needing hyperparameter tuning. Deduct points if answer talks about regularizing effect or being cheaper than SGD.

thank you