From Single Purpose to Multi-task & Multi-modal

Frédéric Odermatt
The classical NLP pipeline

NLP Pipeline (pre Deep Learning)

Parts-of-Speech Tagging
Lemmatization
Dependency Labeling
Named Entity Labeling
Semantic Role Labeling
Coreference

Input Sentence
Sentence Embedding
BERT redisCOVERS THE CLASSICAL NLP PIPELINE

Base: 12x
Large: 24x

$[e_1, e_2, \ldots, e_L]$
BERT redisCOVERS the classical NLP pipeline
BERT rediscovers the classical NLP pipeline

\[ h_1^{(8)} := h_1^{\text{Layer 8}} \]

\[ h_1^{(1)} h_2^{(1)} h_3^{(1)} h_4^{(1)} h_5^{(1)} \]

\[ h_1^{(2)} h_2^{(2)} h_3^{(2)} h_4^{(2)} h_5^{(2)} \]

\[ \vdots \]

\[ h_1^{(L)} h_2^{(L)} h_3^{(L)} h_4^{(L)} h_5^{(L)} \]

\[ \mathbf{MLP} \rightarrow \text{predict} \]

\[ h_{i,\tau} = \sum_{l=0}^{L} s_{\tau}^{(l)} h_i^{(l)} \quad \tau: \text{task} \]

where \( s_{\tau} = \text{softmax}(a_{\tau}) \)

\[ \mathbf{MLP} \rightarrow \text{predict} \]
BERT redisCOVERS THE CLASSICAL NLP PIPELINE

\[ h_{1,\tau} = \sum_{l=0}^{L} s_\tau^{(l)} h_1^{(l)} \]

\( h_1^{\text{Layer 8}} \)

\( e_1 \)

\( e_2 \)

\( e_3 \)

\( e_4 \)

\( e_5 \)

I

eat

strawberry

ice

cream
BERT rediscovers the classical NLP pipeline

NLP Pipeline (pre/early Deep Learning)

- Parts-of-Speech Tagging
- Lemmatization
- Dependency Labeling
- Named Entity Labeling
- Semantic Role Labeling
- Coreference

Input Sentence

Example: Parts-of-Speech Tagging

is it a...

noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, interjection

Sentence Embedding
BERT redisCOVERs the classical NLP pipeline

Parts of Speech Tagging:

- $i = anchor\_layer$
- $\text{eat}$
- strawberry
- ice
- $\text{cream}$

$e_{3} = h_{3}^{l}$

$h_{3,\tau} = \sum_{l=0}^{L} s_{\tau}^{(l)} h_{3}^{(l)}$

$e_{1}^{s}$

$e_{2}^{s}$

$e_{4}^{s}$

$e_{5}^{s}$

is noun? [yes/no]

2 Layer MLP + sigmoid

through backprop optimize
- $s_{\tau} \in [0,1]^{L}$
- 2 Layer MLP
BERT rediscovers the classical NLP pipeline

**NLP Pipeline (pre/early Deep Learning)**

- Parts-of-Speech Tagging
- Lemmatization
- Dependency Labeling
- Named Entity Labeling
- Semantic Role Labeling
- Coreference

**Parts of Speech Tagging:** “I eat strawberry [ice] cream” → Noun

Training: is it a noun? y/n, is it a verb? y/n, is it a...

**Coreference Resolution:** “I haven’t seen [Jack] in the office today, so [he] might be working from home” → True

Training: do these two things refer to the same entity? y/n
BERT rediscovers the classical NLP pipeline

**Analysis 1**

Center of Gravity:

$$E_s[l] = \sum_{l=0}^{L} l \cdot s_r^{(l)}$$

$$h_{i, \tau} = \sum_{l=0}^{L} s_r^{(l)} h_i^{(l)}$$

<table>
<thead>
<tr>
<th></th>
<th>F1 Scores</th>
<th>Expected layer &amp; center-of-gravity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parts-of-Speech Tagging</strong></td>
<td>88.5 96.7</td>
<td>11.68</td>
</tr>
<tr>
<td><strong>Constituency Parsing</strong></td>
<td>73.6 87.0</td>
<td>13.06</td>
</tr>
<tr>
<td><strong>Dependency Labeling</strong></td>
<td>85.6 95.5</td>
<td>13.75</td>
</tr>
<tr>
<td><strong>Named Entity Labeling</strong></td>
<td>90.6 96.1</td>
<td>13.16</td>
</tr>
<tr>
<td><strong>Semantic Role Labeling</strong></td>
<td>81.3 91.4</td>
<td>13.63</td>
</tr>
<tr>
<td><strong>Coreference</strong></td>
<td>80.5 91.9</td>
<td>15.80</td>
</tr>
</tbody>
</table>
M(akridakis) Competitions

Time-series forecasting: “How good are we at it?”

M1 1982
M2 1993
M3 2000
M4 2018
M5 2022
Multi-Modal Deep Learning
How we perceive the world

speech data

human filter

Natural Language

computer vision
CM3: Rethinking domains

CM3: A Causal Masked Multimodal Model of the Internet [Jan 22]

Training on text → Training on HTML source code

• move images to tokens using VQ-VAE-GAN
• includes hyperlinks, markup, etc.
CM3: Unconditional Image Generation

(a) A mountain of olive trees on the way to Cabo de la Vela
(b) Spain Europa Amenacer Winter
(c) blog TIGI Bed Head Tie Dye Spray Hair Spray Hairspray ml
(d) birthday invitation printable christmas gift for birthday party

<img src=""
CM3: Image Infilling

input:
<alt="group of people...", img= [10, 31, mask, mask, 391, 01]

input:
<img=[10, 31, mask, mask, 391, 01]

group of people windsurfing over the beach and water in the ocean.

the wooden park benches are painted dark purple.

some bread is on a plate with jam, an apple, yogurt and orange juice.

a nice looking hotel room with a neatly done bed, coffee table, and a chair.
Multi-Task Learning

Connections to
Multi-Modal Learning
&
Distributed Learning
Mixture of Experts

Non-Sparse

\[ G_{\text{base}} = \text{Softmax}(x \cdot W_g) \]

Sparse (ex: \(k = 2\))

\[ G = \text{Softmax}(\text{Top}_k(H(x))) \]

\[ H(x)_i = (x \cdot W_g)_i + \mathcal{N}(0,1) \cdot \text{Softplus}(x \cdot W_{\text{noise}})_i \]
Five types of Parallelism

Batch size 1

Data parallel = SPMD

GPU 1

Gradient Updates Buffer

GPU N

Batch size B

Model Serialization

GPU 1

$\vdots$

$\vdots$

$\vdots$

GPU 2

$\vdots$

Batch size 1

Data parallel = SPMD

GPU 1

Gradient Updates Buffer

GPU N

Batch size B

Model Serialization

GPU 1

$\vdots$

$\vdots$

$\vdots$

GPU 2

$\vdots$
Pathways Architecture

[Mar 22]

Model & Data Parallel

Claim: as fast as single program multiple data (SPDM)
Pathways Language Model (PaLM)

[Apr 22]

100% accelerator utilization (computation)
6144 TPU v4 Chips
540B parameters
150+ NLP tasks

Tags for tasks:
- **traditional NLP**: context-free question answering, reading comprehension, summarization, ...
- **logic, math, code**: algorithms, logical reasoning, mathematical proof, ...
- **understanding the world**: causal reasoning, common sense, visual reasoning, ...
- **understanding humans**: emotional understanding, intent recognition, humor, figurative language, ...
- **pro-social behavior**: emotional intelligence
- **other**: riddle, multilingual

Performance on 58 Tasks

![Graph showing performance on 58 tasks with different models and parameter sizes.](image)
Pathways Language Model (PaLM)

**Code Generation**

```python
def exercise5():
    """Marie ordered one chicken meal that costs $12, 5 packs of milk that costs $3 each, 4 apples that cost $1.50 each, and some boxes of pizza. Marie paid a total of $50. How many boxes of pizza did Marie order if each box costs $8.50?"""

    total_paid = 50
    cost_of_pizza_per_box = 8.50
    cost_of_meal = 12
    cost_of_milk_per_pack = 3
    cost_of_apples_per_apple = 1.50
    packages_of_milk = 5

    return float(total_paid - cost_of_meal - cost_of_milk_per_pack * 
                   packages_of_milk - cost_of_apples_per_apple * 
                   4) / cost_of_pizza_per_box
```
Pathways Language Model (PaLM)

Translation

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>en</td>
<td>fr</td>
<td>32.9&lt;sup&gt;a&lt;/sup&gt;</td>
<td>38.5</td>
<td>28.3&lt;sup&gt;b&lt;/sup&gt;</td>
<td>37.5</td>
<td>33.9&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>44.0</td>
<td>45.6&lt;sup&gt;c&lt;/sup&gt;</td>
</tr>
<tr>
<td>en</td>
<td>de</td>
<td>25.4&lt;sup&gt;a&lt;/sup&gt;</td>
<td>31.8</td>
<td>26.2&lt;sup&gt;b&lt;/sup&gt;</td>
<td>31.8</td>
<td>26.8&lt;sup&gt;a&lt;/sup&gt; (11)</td>
<td>37.4</td>
<td>41.2&lt;sup&gt;d&lt;/sup&gt;</td>
</tr>
<tr>
<td>en</td>
<td>ro</td>
<td>16.7&lt;sup&gt;a&lt;/sup&gt;</td>
<td>24.2</td>
<td>20.6&lt;sup&gt;b&lt;/sup&gt;</td>
<td>28.2</td>
<td>20.5&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>28.7</td>
<td>33.4&lt;sup&gt;e&lt;/sup&gt;</td>
</tr>
<tr>
<td>fr</td>
<td>en</td>
<td>35.5&lt;sup&gt;a&lt;/sup&gt;</td>
<td>41.1</td>
<td>33.7&lt;sup&gt;b&lt;/sup&gt;</td>
<td>37.4</td>
<td>38.0&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>42.8</td>
<td>45.4&lt;sup&gt;f&lt;/sup&gt;</td>
</tr>
<tr>
<td>de</td>
<td>en</td>
<td>38.9&lt;sup&gt;a&lt;/sup&gt;</td>
<td>43.8</td>
<td>30.4&lt;sup&gt;b&lt;/sup&gt;</td>
<td>43.9</td>
<td>40.6&lt;sup&gt;a&lt;/sup&gt; (11)</td>
<td>47.5</td>
<td>41.2&lt;sup&gt;g&lt;/sup&gt;</td>
</tr>
<tr>
<td>ro</td>
<td>en</td>
<td>36.8&lt;sup&gt;a&lt;/sup&gt;</td>
<td>39.9</td>
<td>38.6&lt;sup&gt;b&lt;/sup&gt;</td>
<td>42.1</td>
<td>37.3&lt;sup&gt;a&lt;/sup&gt; (9)</td>
<td>43.8</td>
<td>39.1&lt;sup&gt;h&lt;/sup&gt;</td>
</tr>
</tbody>
</table>

BLEU score
Pathways Language Model (PaLM)

Joke Explanation

Input: I was going to fly to visit my family on April 6th. My mom said "Oh great, your step dad's poetry reading is that night!" So now I'm flying in on April 7th.

Model Output: The joke is that the speaker's mother is trying to get them to go to their step dad's poetry reading, but the speaker doesn't want to go, so they are changing their flight to the day after the poetry reading.
Pathways Language Model (PaLM)

Future Work

PaLM paves the way for even more capable models by combining the scaling capabilities with novel architectural choices and training schemes, and brings us closer to the Pathways vision:

Google’s Pathway Vision

Pathways: A single model that can generalize across millions of tasks.
GATO
[12 May 22]
BERT redisCOVERS THE CLASSICAL NLP PIPELINE

<table>
<thead>
<tr>
<th>POS</th>
<th>The important thing about Disney is that it is a global [brand]₁. → NN (Noun)</th>
<th>Const.</th>
<th>The important thing about Disney is that it [is] a global brand]₁. → VP (Verb Phrase)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depend.</td>
<td>[Atmosphere]₁ is always [fun]₂ → nsubj (nominal subject)</td>
<td>Entities</td>
<td>The important thing about [Disney]₁ is that it is a global brand. → Organization</td>
</tr>
<tr>
<td>SRL</td>
<td>[The important thing about Disney]₂ [is]₁ that it is a global brand. → Arg₁ (Agent)</td>
<td>SPR</td>
<td>[It]₁ [endorsed]₂ the White House strategy... → {awareness, existed, after, ...}</td>
</tr>
<tr>
<td>Coref.₀</td>
<td>The important thing about [Disney]₁ is that [it]₂ is a global brand. → True</td>
<td>Coref.ₚ</td>
<td>[Characters]₂ entertain audiences because [they]₁ want people to be happy. → True</td>
</tr>
<tr>
<td>Rel.</td>
<td>The [burst]₁ has been caused by water hammer [pressure]₂. → Cause-Effect(e₂, e₁)</td>
<td></td>
<td>Characters entertain [audiences]₂ because [they]₁ want people to be happy. → False</td>
</tr>
</tbody>
</table>

![Diagram showing pre-trained encoder and MLP with spans and contextual vectors]
BERT redisCOVERS THE CLASSICAL NLP PIPELINE

Analysis 2

access to more and more hidden states
- **Semantic Role Labeling**: In natural language processing, *semantic role labeling* (also called shallow semantic parsing or slot-filling) is the process that assigns labels to words or phrases in a sentence that indicates their *semantic role* in the sentence, such as that of an agent, goal, or result.

- **Semantic Proto-Roles**
  
  For decades researchers have debated the number and character of thematic roles required for a theory of the syntax/semantics interface. AGENT and PATIENT are canonical examples, but questions emerge such as: should we have a distinct role for BENEFICIARY? What about RECIPIENT? What are the boundaries between these roles?

<table>
<thead>
<tr>
<th>Role property</th>
<th>Q: How likely or unlikely is it that...</th>
</tr>
</thead>
<tbody>
<tr>
<td>instigated</td>
<td>Arg caused the Pred to happen?</td>
</tr>
<tr>
<td>volitional</td>
<td>Arg chose to be involved in the Pred?</td>
</tr>
<tr>
<td>awareness</td>
<td>Arg was/were aware of being involved in the Pred?</td>
</tr>
<tr>
<td>sentient</td>
<td>Arg was sentient?</td>
</tr>
<tr>
<td>moved</td>
<td>Arg changes location during the Pred?</td>
</tr>
<tr>
<td>phys_existed</td>
<td>Arg existed as a physical object?</td>
</tr>
<tr>
<td>existed_before</td>
<td>Arg existed before the Pred began?</td>
</tr>
<tr>
<td>existed_during</td>
<td>Arg existed during the Pred?</td>
</tr>
<tr>
<td>existed_after</td>
<td>Arg existed after the Pred stopped?</td>
</tr>
<tr>
<td>changed_poss</td>
<td>Arg changed possession during the Pred?</td>
</tr>
<tr>
<td>changed_state</td>
<td>The Arg was/were altered or somehow changed during or by the end of the Pred?</td>
</tr>
<tr>
<td>stationary</td>
<td>Arg was stationary during the Pred?</td>
</tr>
</tbody>
</table>
Figure 2. Our approach uses a convolutional VQGAN to learn a codebook of context-rich visual parts, whose composition is subsequently modeled with an autoregressive transformer architecture. A discrete codebook provides the interface between these architectures and a patch-based discriminator enables strong compression while retaining high perceptual quality. This method introduces the efficiency of convolutional approaches to transformer based high resolution image synthesis.
Modular Architecture for Autonomous AI

- **Configurator**
  - Configures other modules for task
- **Perception**
  - Estimates state of the world
- **World Model**
  - Predicts future world states
- **Cost**
  - Compute “discomfort”
- **Actor**
  - Find optimal action sequences
- **Short-Term Memory**
  - Stores state-cost episodes