Announcements:

- **Next week Tue: Final project presentations**
  - Video recording of 10 min or less, followed by feedback
  - Expect ~85% of final report materials, including main findings and telling the story
  - Ask for feedback where you need it for finishing your final report (due Sunday)
  - Remember: Course participation & feedback to others is part of your grade
- **Next week Sun: Final report due**
  - **New section on Ethical Considerations**
- **Also due on Sun:**
  - **Summary of Individual Contribution to Project**
  - **Final Reflection**
  - (optional) Project product for difference audience

Technical Writing for Data Science

CSE481DS Data Science Capstone

Tim Althoff

PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

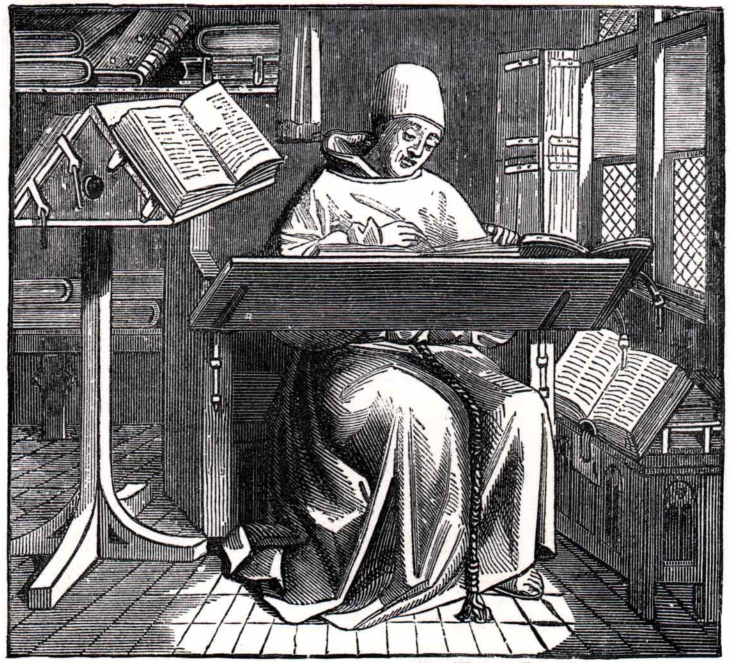
Who cares?

(or, “why shouldn’t I spend this lecture quietly doing homework?”)

- Nobody will know what you did if you can’t communicate it
  - because you can’t have 1:1 conversations with everyone, you have to write it down
- Writing plays a major role in how someone judges your idea
  - if your writing is very unclear, people will not trust your argument
- Writing can change your research
  - it’s an organizational tool that can point out flaws in your research and tell you which experiments or analyses you’re missing

Context

- Writing style depends strongly on the field and audience.
- Today's lecture focuses on data science, data mining, ML, NLP venues.
- While we focus on academic (paper) writing here, the same principles apply to any other technical communication including reports, blog posts, executive summaries, etc.


SCRIPTORIUM MONK AT WORK. (From *Lacroix.*)

Bad at writing?

- Writing is a **skill**. Skills require **practice**.
  - You will get better by doing (and being bad at first)
  - You will get better by getting feedback
  - You will get better by reading **good writing**!
- Not a native English speaker?
  - ***Not a problem!***
  - Good research writing is about **good ideas** and **clear thinking**, not a big mental lexicon

But first

Before you write a word

Your goal

- **You are writing for your readers.**
  - To convey a **message**
    - To teach your reader something
    - To convince your reader of something
    - To explain how you reached your conclusion
  - Be **clear**, even at the cost of precision
  - ***Not your job: to show how clever you are***
- You are **not** primarily writing for you. But you are kind of writing for you (more on this later)

Your message

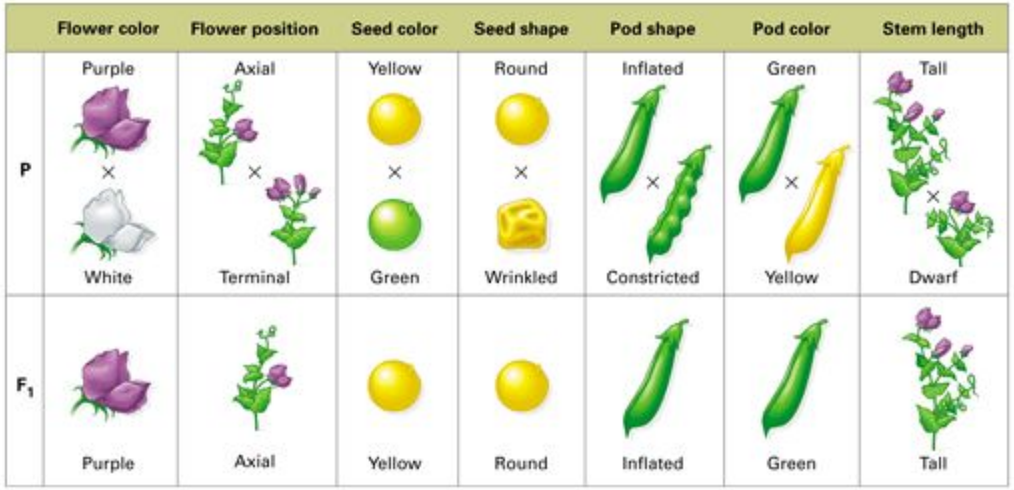
- Figure out what **your message** is. Keep it in mind.
- Make sure the reader knows what this is. **Be 100% explicit.**
  - *"The main idea of this paper is…"*
  - *"The goals of this article are to characterize the core ideas of X and provide a taxonomy of various approaches."*
  - *"In this section we present the main contributions of this paper…"*
- This belongs at the **beginning** of the paper (more later)
- Good ideas that are not distilled = **bad paper**!

Tale of 2 cites

Mendelian genetics (Gregor Mendel, 1822-1844)

- the prevailing belief at the time was in only "blended" inheritance
  - the two parents' traits are blended, in the way that two color paints might be
- 10 year research study on inheritance in pea plants, wrote up findings (~40 page paper) and presented them
  - followed 3 generations + 3 hybrid generations
  - identified multiple characteristics with discrete classes (e.g. "white flower" vs. "purple flower")
  - derived patterns of inheritance for those traits
  - explained "skipping generations"

Main idea: genetics for many traits in pea plants is discrete, and follows consistent rules of dominance

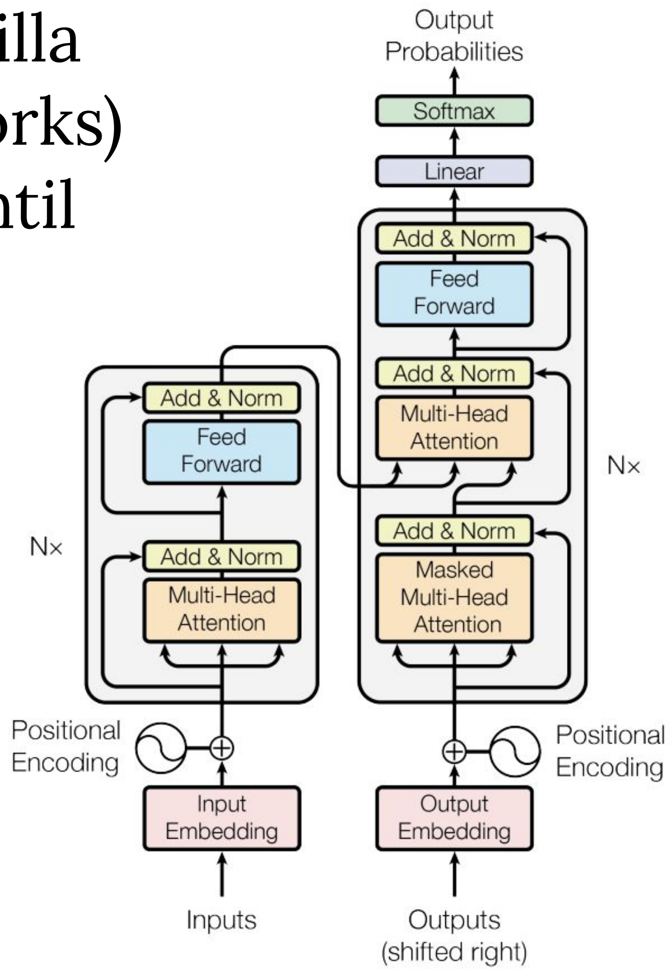


Tale of 2 cites

Transformers (Vaswani, et al., 2017)

- most common architectures at the time were vanilla RNNs and LSTMs (long short term memory networks)
- released in August 2017 but not widely adopted until 2019 (after BERT, around GPT2)
- paper was on neural machine translation tasks

Main idea: our new architecture (Transformers) outperform the existing architectures on NMT tasks (and therefore should be adopted and/or studied further)



Your Reader

To successfully communicate to someone, you should know:

- What do they **know**?
  - Vocabulary and Notation
  - Concepts
  - Prior Work / State of the art
- What do they **think**?
  - Opinions
  - Common assumptions
- What do they **expect**?
  - Format and Style
  - Other conventions

What do I need to explain?

What do I need to address?

How should this look?

Respect your Reader

Anything you've seen another author do that makes their paper hard to read – don't do that:

- Don't bore your reader – ***do get to the point***
- Don't make the reader work more than necessary — ***do organize your writing logically***
- Don't be too harsh — ***do treat people's theories, methods, models, etc. with respect***
- Don't belabor – ***do make your point well and thoroughly, and move on***
- Do ***not*** overestimate your readers
  - We are ***not*** as knowledgeable as you!
  - We will read your paper in **minutes**, **hours**, or **days** … *You have worked on it for **weeks**, **months**, or **years**!*

Who is Your Reader?

- Conference paper
  - (*home conference*) you, if you spent the last year doing something else
  - (*a new conference*) pick an author who publishes there, and imagine them reading the paper
- Journal article
  - someone working in the journal subfield, but *on different problems*
- Dissertation / Book
  - someone from a broad field (Computer Science, Physics) in the future
    - Trick: *Imagine reading your dissertation in 10 years*
    - Anything you depend on that is “hot right now” needs to be contextualized and explained in terms of **stable common ground**

Your paper in this class ≃ DS conference paper

Tale of 2 cites

Mendelian genetics

- primarily wrote for: other scientists interested in inheritance at the time
- was read by: generations of scientists studying genetics

Transformers

- primarily wrote for: NLP researchers, especially MT researchers, in 2017
- was read by: the entirety of a massive ML field, including non-researchers

Questions you need to answer

- Why is this paper important?
  - Are you introducing a new problem?
    - Is the problem obviously important?
    - Do you need to convince them it's important?
  - Are you introducing a new technique?
    - Benefits relative to alternative techniques
    - Costs relative to alternative techniques [be honest]

- What is difficult to understand?
  - Algorithms [correctness, complexity]
  - Theorems [proofs, intuitions]
  - Models [assumptions]
  - Process [data, steps, dependencies]

Tale of 2 cites

Mendelian genetics

- New theory & supporting experiments
- Hardest to explain
  - experimental design
  - conclusions

Transformers (NLP)

- New technique
- Hardest to explain
  - architecture
  - implementation/experimental details

Structure [conference paper]

- Title (1000 readers)
- Abstract (4-8 sentences, 100 readers)
- Introduction (1 page, 100 readers)
- The problem (1 page, 10 readers)
- Our idea (2 pages, 10 readers)
- The details (5 pages, 3 readers)
- Related work (1-2 pages, 10 readers)
- Conclusions and further work (0.5 pages)

Structuring a Paper: From old to new

- Start with the **known**
  - Identify a practical problem in need of solving
  - Identify an example illustrating some unexplained phenomenon
    - unexplained pattern of results
    - inconsistency between theory and reality, or among existing theories or findings
- Progress **logically** to **new** material
  - What is your proposed solution/explanation?
  - How do you express your solution formally and in relation to past work?
  - Why did you choose this solution?
  - What did you do to realize this solution (experiment, proof, etc.)?
  - Results
  - Analysis

Structuring a Paper: Logical flow

- What is logical structure?
  - Getting to the **main ideas** in the most direct way
- What is ***not*** logical structure?
  - Recapitulating how you (or the field) got to an idea
    - "First I ran this experiment, and then it didn't work but I don't know why, so I ran this one, and then I was confused so I ran this one, which told me <x>, so then I compared it to this other thing…"
  - Building a paper around your own anxieties
    - "Here are all the ways I've been criticized and my arguments against them, please believe me."

Writing for you

What part of this is supposed to be helpful for you?

Writing:

- helps you clarify and organize your ideas
- helps you *evaluate* your work (e.g., shows you what you're missing)
- enables you to get feedback from others (more later)

Introduction and Abstract

or, the only parts that
99% of readers will look at

(of course we read project reports in detail :))

The Introduction

- Identify the **problem** you are solving
- Clearly list **your contributions**
  - Your contributions drive the structure of the whole paper
  - **For a survey paper**: Your contribution is a convenient way of understanding a bunch of related techniques / problems
  - You don't need to list *everything*
- For an 8-page paper: intro gets **one page**
  - Longer paper -> longer intro but it's not a linear growth

$$max\_intro\_pages = \sum_{i=1}^{\log_2 \frac{total\_pages}{4}} \frac{1}{i}$$

Do not make the reader guess what your contributions are!

How to structure your introduction

- Following Jennifer Widom's "patented five-point structure for Introductions"
- Also works for abstracts (~1 sentence instead of ~1 paragraph)

1. *What is the* ***problem****?*
2. *Why is it* ***interesting and important****?*
3. *Why is it* ***hard****?* (E.g., why do naive approaches fail?)
4. *Why hasn't it been* ***solved before****?* (Or, what's wrong with previous proposed solutions? How does mine differ?)
5. *What are the* ***key components*** *of my approach and results?* (Or, what are your key contributions?) Also include any specific limitations.

There are no rules about how much space each question gets.

Don't: "the rest of this paper is …"

- Not a laundry list:

> "The rest of this paper is structured as follows. Section 2 introduces the problem. Section 3 … Finally, Section 8 concludes".

- Instead, **use forward references from the narrative in the introduction**.
  The introduction should give a road map of the whole paper, and therefore forward reference every important part.

The most common of these approximations is the max-derivation approximation, which for many models can be computed in polynomial time via dynamic programming (DP). Though effective for some problems, it has many serious drawbacks for probabilistic inference:

1. It typically differs from the true model maximum.

2. It often requires additional approximations in search, leading to further error.

3. It introduces restrictions on models, such as use of only local features.

4. It provides no good solution to compute the normalization factor $Z(f)$ required by many probabilistic algorithms.

In this work, we solve these problems using a Monte Carlo technique with none of the above drawbacks. Our technique is based on a novel Gibbs sampler that draws samples from the posterior distribution of a phrase-based translation model (Koehn et al., 2003) but operates in linear time with respect to the number of input words (Section 2). We show that it is effective for both decoding (Section 3) and minimum risk training (Section 4).

The current standard approach

Problems of standard approach that they are solving.

Map of the paper with forward references

They didn't mention the conclusion!

Tale of 2 cites: Mendel

Context

EXPERIENCE OF ARTIFICIAL FERTILIZATION, such as is effected with ornamental plants in order to obtain new variations in color, has led to the experiments which will here be discussed. The striking regularity with which the same hybrid forms always reappeared whenever fertilization took place between the same species induced further experiments to be undertaken, the object of which was to follow up the developments of the hybrids in their progeny.

Prior approaches

To this object numerous careful observers, such as Kölreuter, Gärtner, Herbert, Lecoq, Wichura and others, have devoted a part of their lives with inexhaustible perseverance. Gärtner especially in his work *Die Bastarderzeugung im Pflanzenreiche* [The Production of Hybrids in the Vegetable Kingdom], has recorded very valuable observations; and quite recently Wichura published the results of some profound investigations into the hybrids of the Willow. That, so far, no generally applicable law governing the formation and development of hybrids has been successfully formulated can hardly be wondered at by anyone who is acquainted with the extent of the task, and can appreciate the difficulties with which experiments of this class have to contend. A final decision can only be arrived at when we shall have before us the results of *detailed experiments* made on plants belonging to the most diverse orders.

& difficulties

Those who survey the work done in this department will arrive at the conviction that among all the numerous experiments made, not one has been carried out to such an extent and in such a way as to make it possible to determine the number of different forms under which the offspring of the hybrids appear, or to arrange these forms with certainty according to their separate generations, or definitely to ascertain their statistical relations.

It requires indeed some courage to undertake a labor of such far–reaching extent; this appears, however, to be the only right way by which we can finally reach the solution of a question the importance of which cannot be overestimated in connection with the history of the evolution of organic forms.

This work

The paper now presented records the results of such a detailed experiment. This experiment was practically confined to a small plant group, and is now, after eight years' pursuit, concluded in all essentials. Whether the plan upon which the separate experiments were conducted and carried out was the best suited to attain the desired end is left to the friendly decision of the reader.

Tale of 2 cites: Transformers

Prior approaches

Recurrent neural networks, long short-term memory [13] and gated recurrent [7] neural networks in particular, have been firmly established as state of the art approaches in sequence modeling and transduction problems such as language modeling and machine translation [35, 2, 5]. Numerous efforts have since continued to push the boundaries of recurrent language models and encoder-decoder architectures [38, 24, 15].

& difficulties

Recurrent models typically factor computation along the symbol positions of the input and output sequences. Aligning the positions to steps in computation time, they generate a sequence of hidden states h_t, as a function of the previous hidden state h_{t-1} and the input for position t. This inherently sequential nature precludes parallelization within training examples, which becomes critical at longer sequence lengths, as memory constraints limit batching across examples. Recent work has achieved significant improvements in computational efficiency through factorization tricks [21] and conditional computation [32], while also improving model performance in case of the latter. The fundamental constraint of sequential computation, however, remains.

Attention mechanisms have become an integral part of compelling sequence modeling and transduction models in various tasks, allowing modeling of dependencies without regard to their distance in the input or output sequences [2, 19]. In all but a few cases [27], however, such attention mechanisms are used in conjunction with a recurrent network.

This work

In this work we propose the Transformer, a model architecture eschewing recurrence and instead relying entirely on an attention mechanism to draw global dependencies between input and output. The Transformer allows for significantly more parallelization and can reach a new state of the art in translation quality after being trained for as little as twelve hours on eight P100 GPUs.

An example focusing on “this work”

We present the Branch-Train-Merge (BTM) algorithm for learning this set of specialized LMs. Our procedure repeatedly expands the ELMFOREST by adding one or more new ELMs completely in parallel. Each new ELM in the ELMFOREST is first *branched* by initializing a new LM with an average of the parameters of the most relevant LMs in the current set, then further *trained* on new domain data with a standard cross-entropy loss, and finally *merged* into the model by simply adding it to the current ELMFOREST (Figure 3 provides a schematic of this process). BTM is initialized with a single LM that is trained on heterogeneous data to establish strong shared representations for future domain specialization, a process that we explore extensively in our ablation analysis.

When evaluated in- and out-of-domain, ELMFORESTs trained with BTM outperform monolithic GPT-style transformer LMs (GPT-LMs) and a previous domain-specialized mixture-of-experts baseline (DEMIX; Gururangan et al. 2022) across a range of computational budgets – up to 1.3B parameters per ELM trained for 7000 GPU-hours in aggregate (Figure 1; §4.2). These gains are biggest for ELMFOREST ensembles, which use all of the model parameters, but also hold when we collapse the models by averaging parameters.

We also perform detailed analysis to understand which aspects of BTM are most important for these gains. Ensembled ELMFORESTs outperform ensembling across random data splits, suggesting that domain specialization is a critical component to our approach (§5.1). We also show that performance is robust to a range of initializations, including the choice of the compute budget allocation (§5.2) and data (§5.3) for training the initial LM. Our ELMFORESTs are also able to forget domains by removing the relevant ELM, as long as they were not included in the initialization phase (§5.3).

Finally, we perform a preliminary scaling study on a training corpus with 192B whitespace-separated tokens (§6.3). Building on our findings, we use BTM to incrementally train a total of 64 experts which form a ELMFOREST. Our scaled ELMFOREST performs comparably with a 1.3B parameter TRANSFORMER-LM trained with 2.5 times the total GPU hours. We find that benefits of BTM increase with the number of domains in the training corpus.

Look at all of these forward references

Abstracts

- Abstracts typically follow the structure of the introduction closely
- They should answer the same questions as the introduction but are more brief
- This brevity should eliminate a lot of details, but should retain every *major* point
- Trick: write the introduction, then summarize each paragraph or idea into one sentence for an abstract. Or,
- Write the abstract first, giving each point one sentence, then expand each sentence into a paragraph or several.

Science paper abstracts

Annotated example taken from *Nature* 435, 114–118 (5 May 2005).

Intro for broad audience	One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.	During cell division, mitotic spindles are assembled by microtubule-based motor proteins[1,2]. The bipolar organization of spindles is essential for proper segregation of chromosomes, and requires plus-end-directed homotetrameric motor proteins of the widely conserved kinesin-5 (BimC) family[3].
	Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.	Hypotheses for bipolar spindle formation include the 'push–pull mitotic muscle' model, in which kinesin-5 and opposing motor proteins act between overlapping microtubules[2,4,5].
What is the problem?	One sentence clearly stating the **general problem** being addressed by this particular study.	However, the precise roles of kinesin-5 during this process are unknown.
What are main results?	One sentence summarizing the main result (with the words "**here we show**" or their equivalent).	Here we show that the vertebrate kinesin-5 Eg5 drives the sliding of microtubules depending on their relative orientation.
What do results add?	Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.	We found in controlled *in vitro* assays that Eg5 has the remarkable capability of simultaneously moving at ~20 nm s^{-1} towards the plus-ends of each of the two microtubules it crosslinks. For anti-parallel microtubules, this results in relative sliding at ~40 nm s^{-1}, comparable to spindle pole separation rates *in vivo*[6]. Furthermore, we found that Eg5 can tether microtubule plus-ends, suggesting an additional microtubule-binding mode for Eg5.
Implications	One or two sentences to put the results into a more **general context**.	Our results demonstrate how members of the kinesin-5 family are likely to function in mitosis, pushing apart interpolar microtubules as well as recruiting microtubules into bundles that are subsequently polarized by relative sliding.
	Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline, may be included in the first paragraph if the editor considers that the accessibility of the paper is significantly enhanced by their inclusion. Under these circumstances, the length of the paragraph can be up to 300 words. (This example is 190 words without the final section, and 250 words with it).	We anticipate our assay to be a starting point for more sophisticated *in vitro* models of mitotic spindles. For example, the individual and combined action of multiple mitotic motors could be tested, including minus-end-directed motors opposing Eg5 motility. Furthermore, Eg5 inhibition is a major target of anti-cancer drug development, and a well-defined and quantitative assay for motor function will be relevant for such developments.

Label the abstract

Identify the:

Problem

Solution

Main results

We present Branch-Train-Merge (BTM), a communication-efficient algorithm for embarrassingly parallel training of large language models (LLMs). We show it is possible to independently train subparts of a new class of LLMs on different subsets of the data, eliminating the massive multi-node synchronization currently required to train LLMs. BTM learns a set of independent EXPERT LMs (ELMs), each specialized to a different textual domain, such as scientific or legal text. These ELMs can be added and removed to update data coverage, ensembled to generalize to new domains, or averaged to collapse back to a single LM for efficient inference. New ELMs are learned by *branching* from (mixtures of) ELMs in the current set, further *training* the parameters on data for the new domain, and then *merging* the resulting model back into the set for future use. Experiments show that BTM improves in- and out-of-domain perplexities as compared to GPT-style Transformer LMs, when controlling for training cost. Through extensive analysis, we show that these results are robust to different ELM initialization schemes, but require expert domain specialization; LM ensembles with random data splits do not perform well.

Tale of 2 cites: Transformers

Identify the:

Problem

Solution

Main results

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Tale of 2 cites: Mendel

Context

EXPERIENCE OF ARTIFICIAL FERTILIZATION, such as is effected with ornamental plants in order to obtain new variations in color, has led to the experiments which will here be discussed. The striking regularity with which the same hybrid forms always reappeared whenever fertilization took place between the same species induced further experiments to be undertaken, the object of which was to follow up the developments of the hybrids in their progeny.

Prior approaches

To this object numerous careful observers, such as Kölreuter, Gärtner, Herbert, Lecoq, Wichura and others, have devoted a part of their lives with inexhaustible perseverance. Gärtner especially in his work *Die Bastarderzeugung im Pflanzenreiche* [The Production of Hybrids in the Vegetable Kingdom] has recorded very valuable observations; and quite recently Wichura published the results of some profound investigations into the hybrids of the Willow. That, so far, no generally applicable law governing the formation and development of hybrids has been successfully formulated can hardly be wondered at by anyone who is acquainted with the extent of the task, and can appreciate the difficulties with which experiments of this class have to contend. A final decision can only be arrived at when we shall have before us the results of *detailed experiments* made on plants belonging to the most diverse orders.

& difficulties

Those who survey the work done in this department will arrive at the conviction that among all the numerous experiments made, not one has been carried out to such an extent and in such a way as to make it possible to determine the number of different forms under which the offspring of the hybrids appear, or to arrange these forms with certainty according to their separate generations, or definitely to ascertain their statistical relations.

It requires indeed some courage to undertake a labor of such far–reaching extent; this appears, however, to be the only right way by which we can finally reach the solution of a question the importance of which cannot be overestimated in connection with the history of the evolution of organic forms.

This work

The paper now presented records the results of such a detailed experiment. This experiment was practically confined to a small plant group, and is now, after eight years' pursuit, concluded in all essentials. Whether the plan upon which the separate experiments were conducted and carried out was the best suited to attain the desired end is left to the friendly decision of the reader.

Tale of 2 cites: Mendel

Related Works

how you fit in to the field

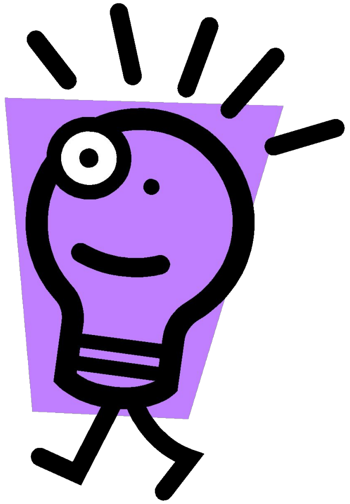
Structure [Conference Paper]

- Abstract (4 sentences)
- Introduction (1 page)
- **~~Related work~~**
- The problem (1 page)
- My idea (2 pages)
- The details (5 pages)
- Related work (1-2 pages)
- Conclusions and further work (0.5 pages)

Related work at the beginning?

Your reader

We adopt the notion of transaction from Brown [1], as modified for distributed systems by White [2], using the four-phase interpolation algorithm of Green [3]. Our work differs from White in our advanced revocation protocol, which deals with the case of priority inversion as described by Yellow [4].



Your idea

Related work at the beginning?

- Problem 1: the reader knows nothing about the problem yet; so your (carefully trimmed) description of various technical tradeoffs is absolutely incomprehensible

→ **Reader feels stupid**

- Problem 2: describing alternative approaches gets between the reader and your idea

→ **Reader feels bored**

Do not put obstacles in your reader's way

How to Write about Related Work

1. Make a laundry list with **all** the relevant works.
2. As you write, **move** citations into the paper.
   - The most important papers your paper is “conversing with” go in the introduction.
   - Papers that are part of your narrative should be smoothed in where they fit naturally (problem, idea, details, …).
3. **Smooth** the “leftovers” into a coherent, organized related work section that discusses more distant works and larger context, also potential confusions. Tuck it at the end of the paper.
   - *Dyer et al. (2013) use similar terminology to refer to a different idea in a different context …*

Related Work: Community Norms

- Related work at the end is common for NLP conferences
- Some venues expect related work as Section 2
- Some are flexible (data mining conferences like WebConf and KDD)
- Journals don't have a related work section at all and expect more natural integration into introduction and discussion.

Lesson: Understand your audience and their community norms. Follow them.

- If your related work section appears early, use it to:
  - Set up necessary context for the reader ("Background and related work")
  - Clarify your contributions and novel ideas
  - If you could simply move the related work later without making it harder for the reader, you should strongly consider to do so.

Other sections

lightning round

Title

- First filtering step
- Include relevant keywords
- The most acceptable place to be daring in style
  - But don't push it
- Not worth spending *that* much time on

Problem, Method, Details

- Be as clear as possible
- Organize and reorganize
- Get feedback. So much feedback (more later).

Conclusion, Future Work

- Be brief
- Drum up excitement
  - Make this paper sound exciting
  - Make people want to read your next paper
  - Make people want to collaborate with you or build off your work

Acknowledgements

- Be generous
  - don't forget anyone, even if you don't feel like they "helped that much"
  - it costs you nothing to give an acknowledgement
- Be brief
  - it's not an Oscar speech

Appendix

- Grab bag of results, etc., that don’t fit well in the main paper
- Can be less polished than the main paper
- Great place for negative results

General Advice

that you really should take

Examples

- Pick examples that
  - Illustrate the easy case easily
  - Illustrate the simplest complicated case easily
  - Are **concrete**

John	proved	correctness	is better than	w1	w2	w3
PN	VBD	NN		t1	t2	t3

  - You don't need to cover all of the most complicated cases!
- Use a running example
  - Return to the same example throughout the paper
- Structure
  - Concrete → abstract

Prefer intuition over formal definitions.

Tale of 2 cites: Mendel

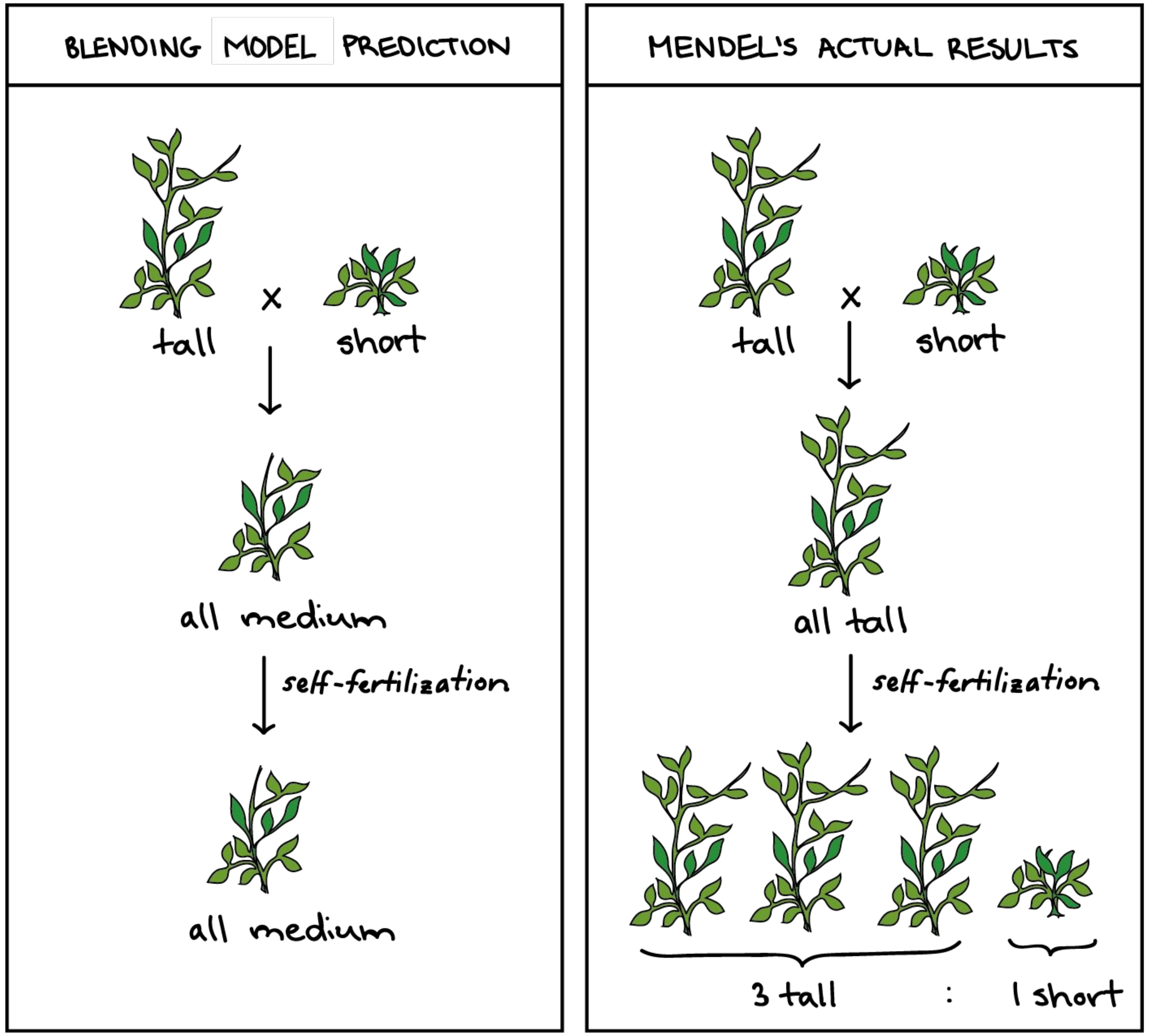


Figure 3; Section 4.1 (hypothetical)

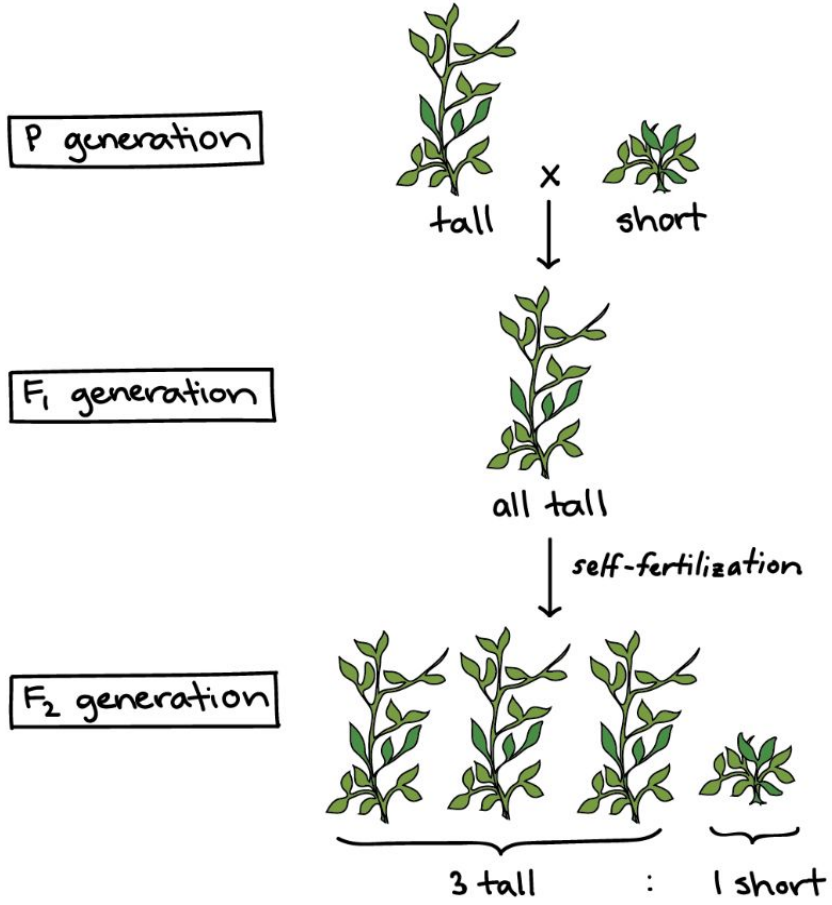
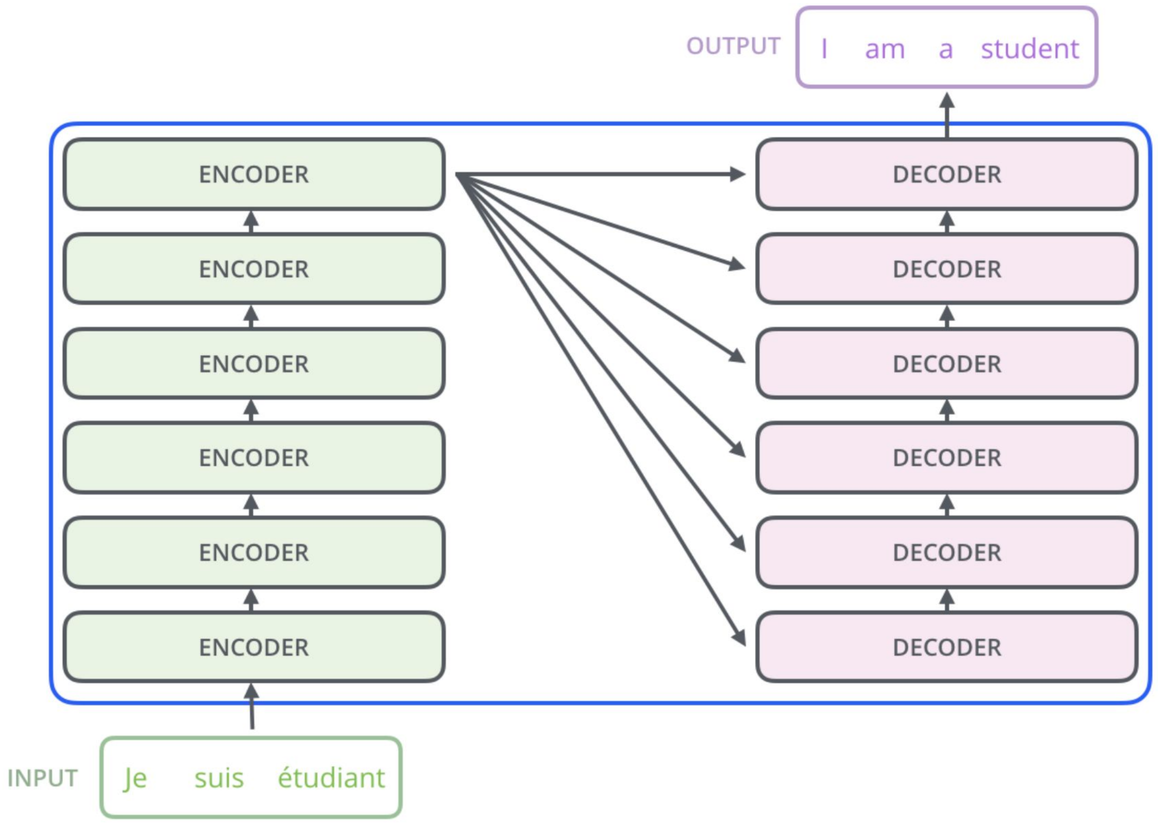
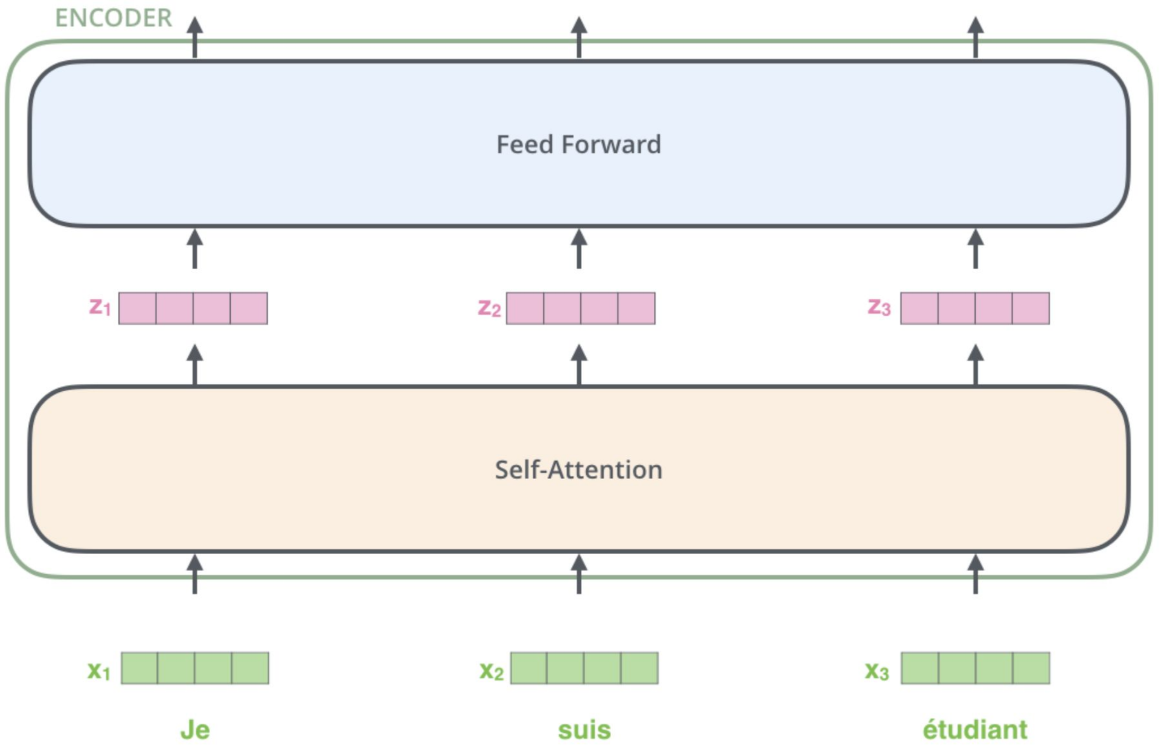


Figure 7; Section 4.4 (hypothetical)

Tale of 2 cites: Transformers





Start early, draft all the time

- Writing is one of the best ways to develop your ideas.
  - So take advantage of it throughout your research process
- You do not need to have a completely focused idea when you **start**, but you **must** have a completely focused idea when you **finish**.
  - Starting the night before a deadline will not get you this.

Getting Feedback

- Who should you ask? Depends on your goals, but people who span the range of your desired audience across:
  - familiarity with your work
    - and with this project in particular
  - familiarity with this subfield
  - years spent publishing (seniority)
  - bias and preconceptions towards this work

Get your paper read by as many colleagues and friends as possible!

Getting Feedback

- What kind of feedback do you prioritize?
  - Sources of: confusion, misunderstanding, boredom
  - ("**I got lost here**" is much more important than "bayes should be capitalized".)
- **Suggestion**: Ask your reader to explain parts of your contribution back to you
  - Did they get it right? If not, you may want to edit.
  - An expert can check details, but the logic of any paper should be comprehensible to a non-expert.
- **Remember**: Each reader can only read your paper for the first time once!

Confidence and Hedging

You want to be cautious about what you cannot claim, and confident in what you can.

- When to hedge:
  - Empirical science is about failing to refute an idea, ***not about proving that an idea is correct***.
    - Rule of thumb: never use the word “prove” unless you are writing a proof
  - Your language around conclusions should signal your awareness of this, e.g., “we have found evidence supporting …” never “we proved that …”
  - Writing guides advise caution in making scientific assertions.

Confidence and Hedging

- When not to hedge:
  - Established facts:
    - Leave your beliefs out of it; focus instead on the reasons for those beliefs.
  - Watch out for verbs like *believe* and *seem.*
  - If you overdo it with hedging language, your reader will get tired; use workarounds that state facts when you can, for example:
    - “We believe that” -> “our conjecture is that …” or “we hypothesize that …”
    - “it seems that <x> is related to <y>” -> “a possible explanation is …”
    - “It’s possible that” -> “future work could explore …”

Confidence and Hedging: Causality

Casual language	Non-causal language
• Causes	• Associated
• Effects, modifies	• Related
• Increases/decreases	• Correlated
• Elevates/reduces	• Predicts
• Makes	• Higher
• Impro	• Lower
•	• Linked to
	• Varies with
n	
nduces	
• Effective in	
• Is attributable to, contributes to	
• Leads to	
• Responsible for	

Harder to claim

Advice: Verbs

- Avoid ***Present / describe*** and friends.
  - E.g. "*We now present the wombat feature…*"
  - Did you invent it? Are you reviewing it? **Present** is ambiguous.
  - **Use a non-ambiguous verb!**
- **Use strong verbs.**
  - E.g., "*We introduce the novel GAGA algorithm*" is stronger than "*We propose the GAGA algorithm.*"
  - Good verbs: **introduce, validate, verify, demonstrate, show, prove**
- **The passive voice is okay!**

Advice: Nouns

- Avoid **pronoun this**. *"This raises questions…"*
- Prefer instead **demonstrative this**: *"This pattern of results raises questions…"*
- (Smith et al., 2012) is **not** a noun. However, *Smith et al. (2012) offered an intriguing solution to the problem of nouns.*

Advice: Adjectives & Adverbs

- Avoid value-judgment adjectives.
  - Bad: We present an important algorithm.
  - Good [**verifiably true**]: We present a novel algorithm.
  - Better [**true and precise**]: We present a novel, polynomial time decoding algorithm using a linear program relaxation of the ILP.
- Use adverbs ***sparingly***.

Advice: Discourse Connectives

- The end of every sentence is an opportunity for a reader to get bored and give up.
- Discourse connectives signal the logical relationship that the next sentence will have to what came before. This keeps them going:

  However,
  As a result,
  Therefore,
  Similarly,
  On the other hand,

- Using the *wrong* discourse connective will confuse your reader.
  - "Experiment A suggests <x>. However, experiment B suggests <x>" – what??

Simple is better than smart (and convoluted)

NO	YES
The object under study was displaced horizontally	The ball moved sideways
On an annual basis	Yearly
Endeavour to ascertain	Find out
It could be considered that the speed of storage reclamation left something to be desired	The garbage collector was really slow

The details (or, "polish")

- There are hundreds of little conventions good writers follow, often compulsively, such as:
  - Spelling, punctuation, grammar norms
  - Citation styles (e.g., know where the parentheses go)
  - Mathematical notation
  - Use of italics, boldface, abbreviations, …
  - Managing tables and figures: self-contained, clear captions; references in the main text; ease of reading; font size; color-blind-friendly palettes, …
- Making these things perfect **will not save an unclear paper**!
- However, a lack of polish will distract readers from your ideas and make it harder for them to trust you!
  - Be the kind of author/scientist who pays attention to details!

Needs polish

https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9136710

VIII. CONCLUSION

Typically, the artificial intelligence tool is arriving at the clinical field in present times. It is presently a reality that we should face to encourage their appearance. In this survey, we discuss various AI techniques that help in speeding up researches and assisting in the current COVID-19 crisis. Also, various learning techniques were emphasized. Cloud computing plays a vital role in virtualization since everyone is in isolation. We discuss various areas in which cloud computing can assist in concurring with this current pandemic. Consolidating enormous information and AI could prompt a significant achievement for the two patients and experts. In any case, even though we distinguished huge numbers of the main impetuses for the usage of AI in the clinical framework, the previously mentioned hindrances could likewise prevent it, particularly if the estimations of the partners are not regarded. Simulated intelligence and huge information must

Needs polish

https://www.sciencedirect.com/science/article/pii/S1877042813028322

5. Conclusion

According to provided regression analysis using presented equation it was found that unemployment rate, Gini coefficient and Tax Freedom Day explains 70.7 percent of emigration reasons. Therefore, in order to decrease emigration rate in Lithuania unemployment rate should go down at least to 8.5 percent, the level of Gini coefficient decrease to 30.

One of suggestions for Lithuania tax system trying to decrease migration rate, it would be to reduce tax paying time by 5 working days and to reach an average of EU27. Speaking about taxation issues it is possibility to think about VAT decreasing as well. Regression analysis showed when VAT increases more than 18.5 percent a number of emigrants start growing rapidly. VAT is 21 percent at the moment in Lithuania. It would be useful to return it to 19 percent as it was before January of 2009. It would help people to survive taxis and increase purchasing power.

Needs polish

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3489488/pdf/nihms192342.pdf

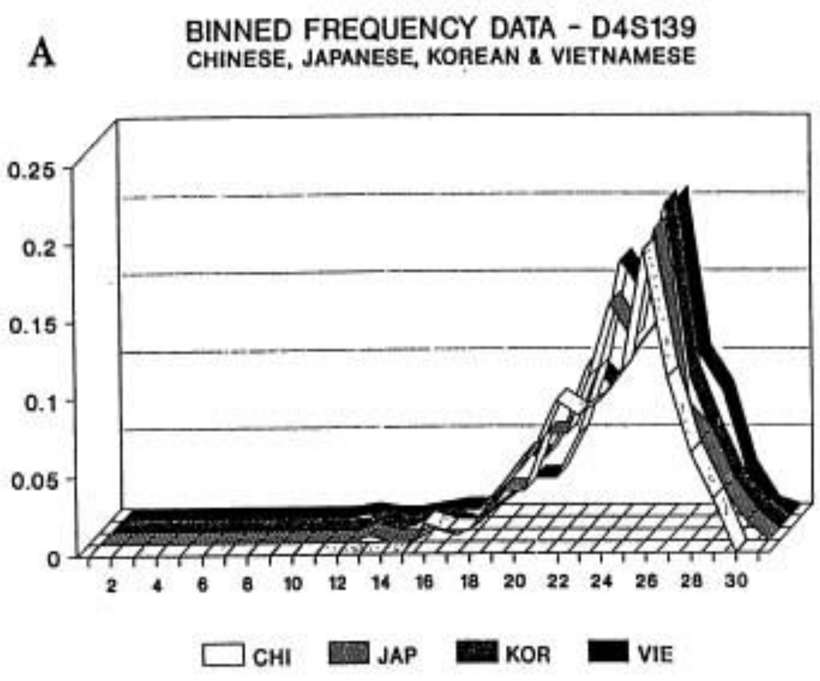
A third innate system of representation with numerical content: Natural language quantifiers

Please dwell upon the experiments that reveal the set-size signature of parallel individuation. Infants choose a set of 3 crackers over a set of 2 crackers or a single cracker; and shown 3 balls placed into a box, having retrieved 2 or 1 of these, they search for the remaining ones. However, they are at chance in a comparison of 1 and 4 crackers, and shown 4 balls placed into a box, they are satisfied after retrieving only 1. These are very counterintuitive phenomena. Not only are infants failing to encode the set of four as approximately four (using analog magnitudes), they are failing to encode it as "plural," for if they had done so, they would represent it as more than one.

The infants in these studies are 12- to 14-months of age; the failures at 1 vs. 4 comparisons in the box-search task are also observed at 16, 18, and 20 months of age (Barner, Thalwitz, Wood & Carey, 2007). In spite of these failures, we now know that there are circumstances in which prelinguistic infants and non-human primates reveal representations of the singular-plural distinction. In the above studies, the individuals move independently of each other, encouraging the infants to deploy parallel individuation. If sets of objects move as coherent wholes (e.g., glued to a platform, such that the items move together), young infants and rhesus macaques distinguish singletons from sets of more than one and fail to distinguish among plural sets of different numerosities, at least for small sets (e.g., 2, 3, 4, and 5; Barner, Wood, Hauser & Carey, 2008; Barner, Thalwitz, Wood & Carey, unpublished data). Again, we do not know why monkeys and infants do not draw on analog magnitudes on these tasks, but the data indicated that they do not, whereas their pattern of behavior reflects a categorical distinction between singletons, on the on hand, and pluralities, on the other.

Needs polish

https://www.jstor.org/stable/2246327



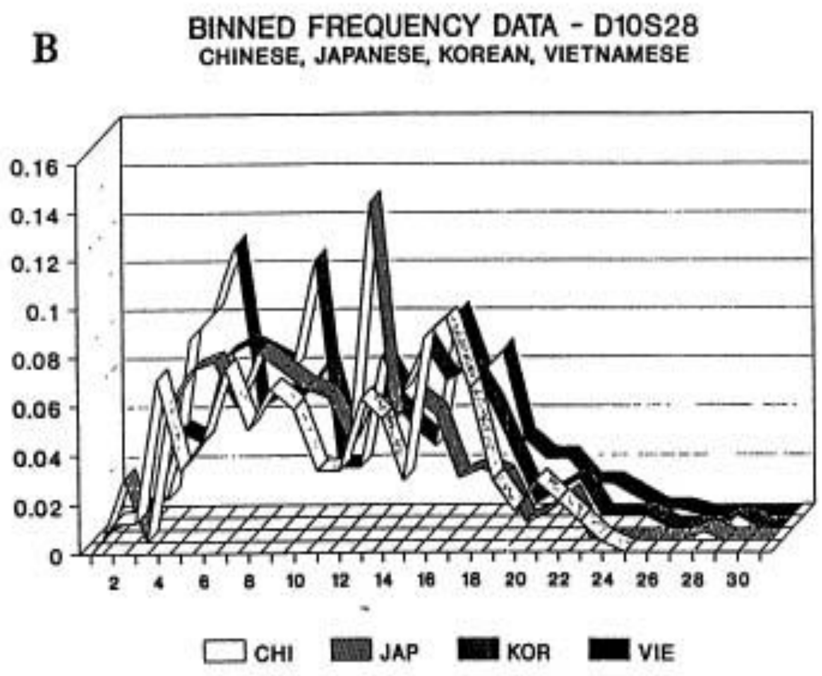


FIG. 4. *Fixed bin distribution (histogram) for two loci and four Asian subpopulations (used with permission from John Hartmann): the boundaries of the 30 bins (vertical axis) are determined by the FBI; these bins are not of equal length. Sample sizes (numbers of individuals) for Chinese, Japanese, Korean and Vietnamese are* 103, 125, 93 *and* 215 *for* D4S139 *and* 120, 137, 100 *and* 193 *for* D10S28. *The horizontal axis is the bin number; bins are not of equal length.*

Needs polish??

https://www.theliberatedmathematician.com/wp-content/uploads/2015/11/PiperThesisPostPrint.pdf

1.5.1 Historical Truths and That Time I Was Wrong

A first question to be asked might be "What has already been done?" Without the shape condition, the question of "how many (and how)" S_n-number fields are there (ordered by discriminant) has already been answered for $n = 3, 4, 5$. In each case, the first step was a parametrization that allows you to look at forms instead of number fields [DF64, Bha04, Bha08]. Counting results were done in [Dav51b, Dav51c, DH71, Bha05, Bha10]. With the shape condition, the question was answered in [Ter97] for $n = 3$. (I should also note that in [BST13] they rewrite things we need for $n = 3$ from [Dav51b, Dav51c, DH71, DF64] in an easier-to-use way so I often use that reference for myself.)

What does that give us? Well, first, I thought I was supposed to read Terr's thesis [Ter97] and magically generalize it to $n = 4$ (the case I worked on). This was folly. Then, I thought I was supposed to rewrite [Bha05] adding "and shape in W" everywhere. This is what I did and I alternated between feeling the task was impossibly hard and trivially, plagiarizingly easy (common feelings for grad students). And then one day (and we won't say which day), my advisor tells me I should just "use" what is known and "make an argument" to prove my result. MIND = BLOWN.

Needs polish??

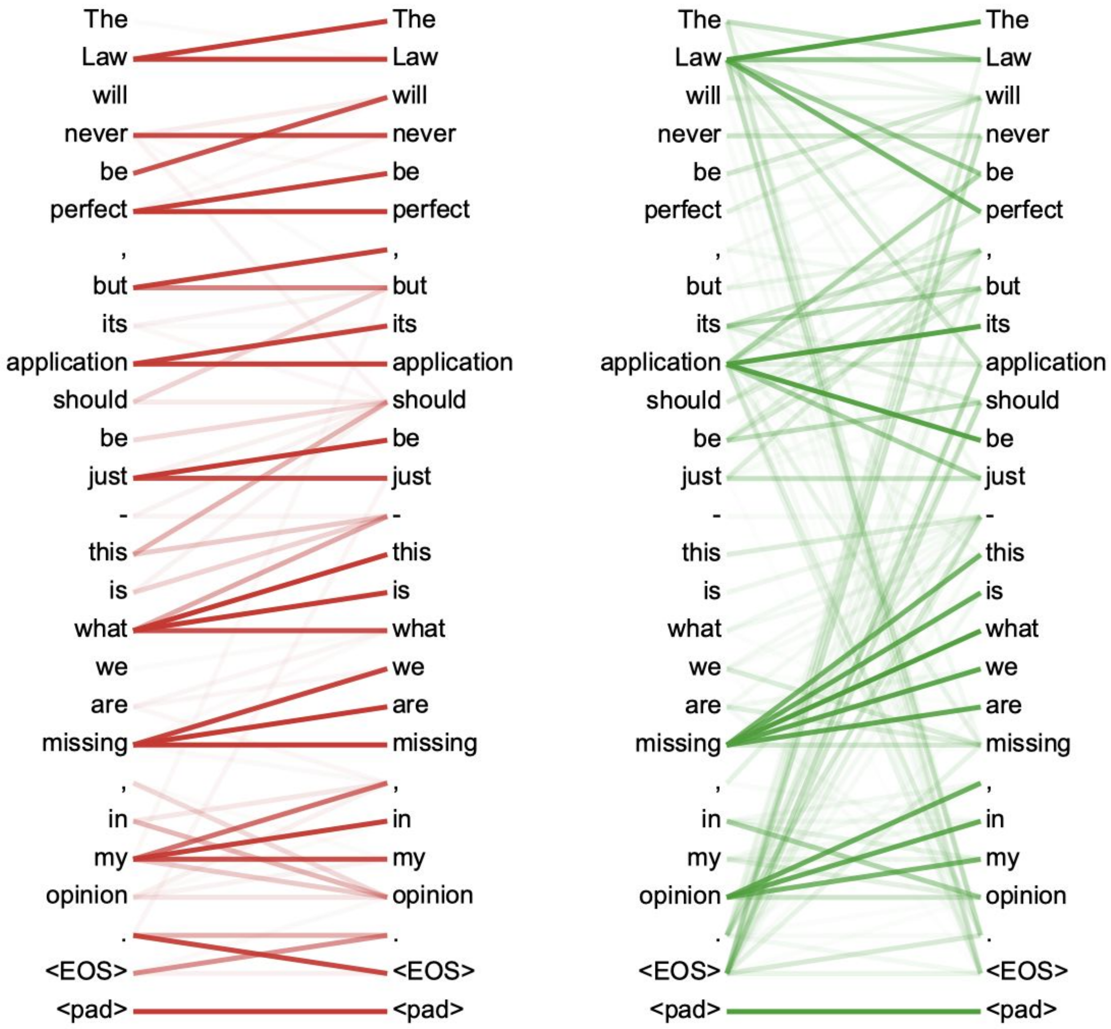
In this section we compare various aspects of self-attention layers to the recurrent and convolutional layers commonly used for mapping one variable-length sequence of symbol representations $(x_1, ..., x_n)$ to another sequence of equal length $(z_1, ..., z_n)$, with $x_i, z_i \in \mathbb{R}^d$, such as a hidden layer in a typical sequence transduction encoder or decoder. Motivating our use of self-attention we consider three desiderata.

One is the total computational complexity per layer. Another is the amount of computation that can be parallelized, as measured by the minimum number of sequential operations required.

The third is the path length between long-range dependencies in the network. Learning long-range dependencies is a key challenge in many sequence transduction tasks. One key factor affecting the ability to learn such dependencies is the length of the paths forward and backward signals have to traverse in the network. The shorter these paths between any combination of positions in the input and output sequences, the easier it is to learn long-range dependencies [12]. Hence we also compare the maximum path length between any two input and output positions in networks composed of the different layer types.

As noted in Table 1, a self-attention layer connects all positions with a constant number of sequentially executed operations, whereas a recurrent layer requires $O(n)$ sequential operations. In terms of computational complexity, self-attention layers are faster than recurrent layers when the sequence length n is smaller than the representation dimensionality d, which is most often the case with sentence representations used by state-of-the-art models in machine translations, such as word-piece [38] and byte-pair [31] representations. To improve computational performance for tasks involving very long sequences, self-attention could be restricted to considering only a neighborhood of size r in the input sequence centered around the respective output position. This would increase the maximum path length to $O(n/r)$. We plan to investigate this approach further in future work.

Needs polish??



https://arxiv.org/pdf/1706.03762.pdf

Figure 5: Many of the attention heads exhibit behaviour that seems related to the structure of the sentence. We give two such examples above, from two different heads from the encoder self-attention at layer 5 of 6. The heads clearly learned to perform different tasks.

Your Voice

- **Scientific:** A key tenet of science is that true findings are true no matter who found them; we write with some personal distance from the content, and this establishes trust.
  - Never: *happily, our method worked better than the baseline*
  - Informal language and slang will deplete reader trust
- **Personal:** Readers suffer if all papers sound the same.
  - Avoid: clichés, tropes, catch-phrases, repetition, dry writing without variation, clumsy mimicry of science-like language.
  - Take this with a grain of salt: never write a sentence that more than a very few people could have written.

Your scientific voice needs to be professional but also engaging. Proofread by reading your paper out loud.

~~Your voice~~ not a good scientific voice

- “With the addition of this feature, we are fortunately able to achieve a 3 point improvement in performance.”
- “Thus, we have established the viability of our approach.”
- “This outdated baseline method underperformed.”
- “It was really weird that we found no correlation between these features.”
- “[Names] et al (2020) put forth a nonsensical argument regarding this phenomenon.”
- “These results indicate the promise of our method.”

Ethics

Not an afterthought

Research Ethics

- Computer Science research and products are impacting individuals and our society both positively and negatively. It is our responsibility to consider and navigate any ethical concerns.
- Therefore, each final report needs to contain a section on Ethical Considerations. This is also required by many conferences and journals.
    - If you find that there are minimal risks, state that and explain how you came to this conclusion
    - If there are any potential risks, discuss these and what could be done to mitigate them.

Principle-based Ethics Framework

How do we systematically assess potential risks?
Principle-based ethics framework, following Coghlan et al., 2023; Floridi & Cowls, 2019; Beauchamp & Childress, 2001

1. **Non-maleficence:** Avoid causing physical, social or mental harm to users
2. **Beneficence:** Ensure that interventions do good or provide real benefit to users
3. **Respect for Autonomy:** Respect users' values and choices
4. **Justice:** Treat users without unfair bias, discrimination or inequity
5. **Explicability:** Provide to users sufficient *transparency* about the nature and effects of the technology, and be *accountable* for its design and deployment

You don't need to address *all* of these principles – focus on the relevant ones.

Questions to ask yourself

Modified from ACL Ethics Review Questions (https://2021.aclweb.org/ethics/Ethics-review-questions/)

- How would the technology be deployed in actual use cases? Does your research reflect how the technology would be deployed?
- What are the possible harms:
  - when the technology is being used as intended and functioning correctly?
  - when the technology is being used as intended but gives incorrect results?
  - when the technology is being intentionally misused?
- If the system learns from user input once deployed, what are the checks and limitations to the learning process?
- Will the harms identified fall disproportionately on populations that already experience marginalization or are otherwise vulnerable?
- If there are harms, what are the potential mitigation strategies?
- If there are human subjects in our research, what are the effects on them?

Tale of 2 cites

Mendelian genetics & Transformer (Vaswani, et. al., 2017)

- For each paper, separately:
  - How would you write an ethics section for this paper?
  - If the authors had applied the questions listed previously to their thinking, prior to starting their research, what might have changed?
  - How would you answer those questions now?

Start *now*

- When should these questions be asked?
  - Probably *not* for the very first time when all of the research is done
  - Probably not only once, even!
  - Ideally, you keep the impact of your work in mind throughout your research

How to write an ethics section

https://aclanthology.org/2023.acl-long.555.pdf

Cognitive Reframing of Negative Thoughts through Human-Language Model Interaction

Ashish Sharma[♠] **Kevin Rushton**[◇] **Inna Wanyin Lin**[♠] **David Wadden**[♣]
Khendra G. Lucas[◇] **Adam S. Miner**[♥♡] **Theresa Nguyen**[◇] **Tim Althoff**[♠]
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington
[◇]Mental Health America [♣]Allen Institute for Artificial Intelligence
[♥]Department of Psychiatry and Behavioral Sciences, Stanford University
[♡]Center for Biomedical Informatics Research, Stanford University
{ashshar,althoff}@cs.washington.edu

"...[W]e conduct a human-centered study of **how language models may assist people in reframing negative thoughts**... we define a **framework** of seven linguistic attributes that can be used to reframe a thought. We develop automated **metrics** to measure these attributes and **validate** them with expert judgements from mental health practitioners. We **collect a dataset** of 600 situations, thoughts and reframes from practitioners and use it to **train a retrieval-enhanced in-context learning model** that effectively g**enerates reframed thoughts and controls their linguistic attributes**...
[W]e conduct an IRB approved randomized **field study** on a large mental health website with over **2,000 participants**..."

How to write an ethics section

10 Ethics Statement

Intervention in high-risk settings such as mental health necessitates ethical considerations related to safety, privacy and bias. There is a possibility that, in attempting to assist, AI may have the opposite effect on people struggling with mental health challenges. Here, in active collaboration and consultation with mental health professionals and clinical psychologists, we took several measures to minimize these risks.

Crisis Resources. We made it very explicit that the model should not be used as a "cry for help" outlet and should not be used in cases of suicidal ideation and self-harm. Also, we provided two crisis resources – Crisis Text Line (crisistextline.org) and 988 Suicide and Crisis Lifeline (988lifeline.org) – to our participants at the start of the study.

Informed Consent from Participants. We obtained informed consent from all participants in our randomized field study (Appendix H). All participants were 18 years of age and older. Participants were informed that they will be interacting with an AI-based model that automatically generates reframed thoughts and is not monitored by a human. Also, they were informed about the possibility that some of the generated content may be upsetting or disturbing.

Safety Measures. To minimize harmful LM-generated reframings, we filtered out any response that contained suicidal ideation or self-harm-related words or phrases. For this, we created a list of 50 regular expressions (e.g., to identify phrases like "*feeling suicidal*", "*wish to die*", "*harm myself*") using suicidal risk assessment lexicons such as Gaur et al. (2019). An LM-generated response that matched any of the regular expressions was filtered out and not shown to the participants. Also, participants were given an option to flag inappropriate reframing suggestions through a "Flag inappropriate" button (Appendix C).

Privacy. We did not collect any privately identifiable information in our randomized field study and removed any user identifiers before conducting our data analysis. All research data was stored within a separate secure computing environment and only trained research personnel were provided access to data. The situations and thoughts collected in §4.1 went through an anonymization process, where we manually removed any user identifiers and replaced any specific identifiable information including locations, names, etc. with their more general version, following Matthews et al. (2017).

Summary

- If you remember nothing else from today:
  - Write for your **readers**, not yourself
  - Communicate **one** main message
  - Identify your contributions
  - Use clear, concrete examples
  - Move from known and the concrete to the abstract and the new
  - Use precise language and hedge sparingly

- Final reports are due in ~12 days. Start writing now!

Acknowledgements & Further Material

- Philip Resnik (UMD, Chris's PhD advisor)
- Simon Peyton Jones (MSR Cambridge)

  http://research.microsoft.com/en-us/um/people/simonpj/papers/giving-a-talk/giving-a-talk.htm

  [Bonus: how to give a research talk;
  how to write a research proposal;
  video of him talking about good writing]

  [*Several slides are taken from SPJ's posted talk*]

  http://www.cs.jhu.edu/~jason/advice/how-to-write-a-thesis.html

- Jason Eisner (JHU, Noah's PhD advisor)

Acknowledgements & Further Material

- Geoffrey K. Pullum (Edinburgh)

  http://www.lel.ed.ac.uk/grammar/passives.html

- Steven Pinker (Harvard), *The Sense of Style*
- Jennifer Widom (Stanford)
  - https://cs.stanford.edu/people/widom/paper-writing.html

Today's slides were adapted from Noah Smith and Chris Dyer

This talk and its contents

If you were paying attention, you may have noticed that this lecture was structured like a paper:

- Abstract / Introduction
- Method overview
- Details
- Conclusion
- Acknowledgements / Related Works

With running examples throughout

Break ☺️

Your Turn ☺️

Activity

- Get into your project groups
- [~30 min] Write a draft abstract for your project report in Google Doc
  - You will need this for your final report.
- [~15min] Give feedback to another group. Decide within your group who will cover which group.
- Remember:

1. *What is the problem?*
2. *Why is it interesting and important?*
3. *Why is it hard?* (E.g., why do naive approaches fail?)
4. *Why hasn't it been solved before?* (Or, what's wrong with previous proposed solutions? How does mine differ?)
5. *What are the key components of my approach and results?* (Or, what are your key contributions?) Also include any specific limitations.
6. What are the *implications* of your findings?

Next steps for your report: Thinking about Abstract & Introduction

1. *What is the problem?*
2. *Why is it interesting and important?*
3. *Why is it hard?* (E.g., why do naive approaches fail?)
4. *Why hasn't it been solved before?* (Or, what's wrong with previous proposed solutions? How does mine differ?)
5. *What are the key components of my approach and results?* (Or, what are your key contributions?) Also include any specific limitations.
6. What are the *implications* of your findings?

Look out for the course survey next week!

Your participation and feedback is critical!

Thank you!

Thank you for sharing your feedback with us!

http://bit.ly/cse481ds-au23-feedback