Large Language Models, Seven Plus Minus Two Law, Fuzzy Logic, Zipf Law, and Principal Components Analysis of Word Embedding: How Is All This Possibly Related

Miroslav Svitek[1], Olga Kosheleva[2], and Vladik Kreinovich[3]

[1]Faculty of Transportation Sciences,
Czech Technical University in Prague, Konviktska 20,
110 00 Praha 1, Czech Republic, miroslav.svitek@cvut.cz
[2,3]Departments of [2]Teacher Education and [3]Computer Science
University of Texas at El Paso,
500 W. University, El Paso, Texas 79968, USA
olgak@utep.edu, vladik@utep.edu

1. Hallucinations are a problem for Large Language Models (LLMs)

- Large Language Models are fascinating.
- They produce poems, texts, class curricula, it looks like they can produce almost anything we want.
- However, what they produce is not always reliable.
- Reasonably often, they produce answers that are smooth and may seem reasonable, but are, in reality, wrong.
- This phenomenon is known as *hallucinations*.
- When hallucinations were first detected, the hope was that additional training will deal with this phenomenon.
- However, this did not happen.
- In spite of all the further training, the hallucination rate remains at the approximately 15% level.

2. We humans can often detect hallucinations, and what does that mean

- In many cases, humans users can easily detect hallucinations.
- They do it by using simple logic to compare LLMs with facts that we know.
- And, by the way, the LLMs knows the same facts.
- However, they lack an ability to compare its conclusions with these facts.
- From this viewpoint, the main reason for hallucinations is that LLMs:
  - while making perfect statistical conclusions,
  - are not very good in thinking logically.
- We humans can both provide some statistical conclusions and we can also use logic.
- Statistical conclusion is something that all animals do.

3. We humans can often detect hallucinations, and what does that mean (cont-d)

- The understanding of this started with Pavlov's experiments.
- In these experiments, dogs:
  - learned artificially introduced statistical dependencies
  - after a reasonably small (in comparison with neural networks) number of iterations.
- Can animals make logical conclusions? Doubtfully.
- Even modern humans, logic-trained at schools, are not very good in logic; see, e.g., Kahneman's work.
- Most probably, our ancestors were even worse.
- LLMs use a lot of optimization to process data – they use practically all the knowledge from everywhere in the world.
- So, it looks like 15% is the best we can do if only use statistics, but not logic.

4. We humans can often detect hallucinations, and what does that mean (cont-d)

- What is 15%? It is approximately one out of seven.
- So this means that if we only use statistics, then one wrong answer out of seven is the best we can do.

5. What are the possible biological consequences of this fact?

- How is this related to humans?
- Our ancestors were not very good in logical reasoning.
- So, they had to live with this limitation.
- They has to take into account that 1/7 of their decisions would be wrong.
- And evolution should have adjusted our brains to this fact.
- What can this imply?
- We cannot reach error rate lower that 1/7.
- This means that it makes no sense to view and consider things with better accuracy.
- A similar example:
  - if we want to compute the distance with accuracy 10%,
  - there is no need to measure velocity or time with higher accuracy.

6. What are the possible biological consequences of this fact (cont-d)

- And what does this mean that we have accuracy 1/7?
- t means that, e.g., on the interval $[0, 1]$ (or on any other interval), we can only distinguish at most 7 different values.
- This is exactly what psychologists observe.
- The famous "seven plus minus seven law" states that, in general, we can only consider 7 plus minus 2 different options.
- Hereby we perceive seven major colors, we have seven days in a week, etc.
- So maybe the LLMs hallucination rate is an explanation for the 7 ± 2 law?

7. How is this all related to fuzzy?

- We are trying:
  - to understand how people think, and
  - to explain why they think and reason that way.
- In this analysis, it is reasonable to use fuzzy techniques.
- Indeed, these techniques were specifically invented:
  - to describe imprecise ("fuzzy") human statements and human reasoning
  - in precise terms.
- Let us use fuzzy techniques to brainstorm about uncertainty of human reasoning.
- Let us start with a situation in which we have no knowledge about some statement.
- We have no reasons to believe that this statement is true.

8. How is this all related to fuzzy (cont-d)

- If we had some reasons, our degree of confidence in this statement would be closer to 1.
- We also have no reasons to believe that this statement is false.
- If we had some reasons, our degree of confidence in this statement would be closer to 0.
- In such situations, it is reasonable to describe our resulting degree of confidence in this statement by a value which is equally distant from 0 and 1.
- In other words, by the value 0.5.
- This value thus corresponds to *unknown*.
- Suppose now that we gained some knowledge.
- This means that instead of "unknown", we have a smaller degree of uncertainty.

9. How is this all related to fuzzy (cont-d)

- This can be naturally described as "somewhat unknown".
- How can we describe the hedge "somewhat"?
- A usual way in fuzzy technique is:
  - to use x^2 to describe "very" and
  - to use the inverse operation – square root – to describe "somewhat".
- If we take the square root of 0.5, we get the degree close to 0.7.
- So, the remaining degree of uncertainty is $1 - 0.7 = 0.3$.
- With this degree of uncertainty, we get $1/0.3 \approx 3$ different levels.
- What if we gain even more knowledge?
- In this case, we again apply the square root – this time to the square root of 0.5.

10. How is this all related to fuzzy (cont-d)

- This way, we get approximately 0.84, with the remaining degree of uncertainty $1 - 0.84 = 0.16$.
- This is again close to 1/7 (or maybe to 1/6).
- So this is maybe why we have 1/7?

11. Comments

- Many fuzzy papers mention the 7 ± 2 law.
- They use it to explain why, usually, we form 7 ± 2 natural language terms to describe the value of each quantity.
- Examples: very small, small, etc.
- This way, this law explains the empirical success of such fuzzy models.
- What we decided is to do it the other way around: use fuzzy techniques to explain the 7 ± 2 law itself.
- Why apply twice and not more times?
- Some arguments in favor of two times are given in our recent paper.
- But what if we still apply the operation one more time?
- This time, we will get 0.917, so the remaining degree of uncertainty is 0.083.

12. Comments (cont-d)

- This is almost exactly 1/12.
- This may be the reason why 12 is often used by us – as in a dozen or as in a musical scale.

13. This somewhat explains 7, but how can we explain plus minus 2?

- Of course, 0.15 is an approximate number, and, correspondingly, 7 is an approximate number.
- How accurate is it?
- Usually, we have less uncertainty about our uncertainty than we have uncertainty about the actual value.
- So, to gauge how uncertain we are about number 7, we need to use the previous – higher – level of uncertainty.
- On this level, the relative uncertainty was about 0.3.
- So, the absolute uncertainty with which we take the value 7 is $\pm 7 \cdot 0.3$.
- This is exactly 7 ± 2 that psychologists have observed.

14. Comment

- So, we can keep in mind, at the same time, only 7 ± 2 objects, between $7 - 2 = 5$ and $7 + 2 = 9$.
- This means that some people can keep no more than 5, others can keep up to 9.
- This may be a reason why in Islam:
  - where it is emphasized that all the wives should get the same good attention and care,
  - a person can have no more than 4 wives.
- This way:
  - even a person who can take into account only up to 5 objects,
  - shall be able to take into account both himself and all his wives.

15. How is this related to word embedding

- Researchers is natural language processing have found a way to check how close are different concepts.
- For this purpose, they characterize each term by several numerical quantities.
- This way, each word is represented by a tuple consisting of several numbers.
- In this sense, words are *embedded* into a multi-dimensional space.
- It turns out that distance in this space is a good indication of how close the original concepts are.
- For example, the word “doctor” appears close to the related word “nurse”.

16. How is this related to word embedding (cont-d)

- Then, they use Principle Component Analysis PCA:
  - to reduce the dimension of the data space
  - while preserving the notion of closeness as accurately as possible.
- It turns out that we can retain practically all information about closeness if we only keep the three main dimensions.
- A natural question is: why 3?

17. Zipf's law can help

- To explain, let us yet another empirical law – Zipf Law.
- It says that if we sort features of objects by importance, the importance of the i-th term is proportional to $1/i$.
- This law was first described in linguistics:
  - if we sort all the words from a language by their frequency,
  - then the frequency of the i-th word is proportional to $1/i$.
- Later on, it turned out that this law is ubiquitous: e.g., it describes the distribution of companies by size.
- In our case, Zipf law says that when we apply PCA to word embedding, the contribution of the i-th dimension is proportion to $1/i$.
- The usual Euclidean distance is the sum of the squares of the differences.
- According to the 7 ± 2 law, we can perceive 7 factors.

18. Zipf's law can help (cont-d)

- So the contribution of all 7 dimensions is equal to

$$1 + \frac{1}{2^2} + \ldots + \frac{1}{7^2} \approx 1.51.$$

- By the same law:
  - it is sufficient to have the sum of fewer terms,
  - as long as the resulting sum is approximately equal to this number, with accuracy of 1/7.
- In other words, it is sufficient to make sure that the sum of the terms corresponding to selected dimensions is larger than or equal to

$$1.51 - \frac{1.51}{7} \approx 1.30.$$

- For two dimensions, we have

$$1 + \frac{1}{2^2} = 1.25 < 1.30.$$

19. Zipf's law can help (cont-d)

- So using only two dimensions is not enough.
- However, for three dimensions, we already have

$$1 + \frac{1}{2^2} + \frac{1}{3^2} = 1.3611\ldots > 1.30.$$

- This explains why empirically, three dimensions are sufficient to describe our commonsense concept of closeness between concepts.

20. Acknowledgments

This work was supported in part:

- by the US National Science Foundation grants:
  - 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science),
  - HRD-1834620 and HRD-2034030 (CAHSI Includes),
  - EAR-2225395 (Center for Collective Impact in Earthquake Science C-CIES),
- by the AT&T Fellowship in Information Technology,
- by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Focus Program SPP 100+ 2388, Grant Nr. 501624329,
- and by the European Union under the project ROBOPROX (No. CZ.02.01.01/00/22 008/0004590).