**Future of NLP – 4 Problems**

**––––––––––––––––––––**

*https://www.youtube.com/watch?v=G5lmya6eKtc*

* **Problem of bigger & bigger models (up to 10 billion parameters)**
  + -> No academics in GLUE leaderboard, as comp. requirements are too big; Narrows research
  + -> Bad for the environment ; up to 3 car lifetimes CO2 for 1 huge training w/ architecture search
    - Neural nets are over-parametrized
    - -**> Distillation:** 
      * DistillBERT:
      * You take the good model as a teacher and train the student model to produce the same results
      * -> Cross Entropy btw output of teacher & student
    - **-> Pruning:**
      * Remove weights from the teacher model
        + Remove attention weights
        + Remove normal weights X Might not help that much removing matrix weights, as already optimized for GPU computation (Sparse matrices are often just as difficult to train -> try to structure sparsity by e.g. removing blocks /// switch chips designed for sparse matrices)
        + -> Layer Pruning: Removing a full layer; You need to train the model to be resilient to losing layers by using layer dropout during training

Works well because of the residual connections

* + - **-> Quantization:**
      * Instead of e.g. Float32 using int8 -> Q8Bert by Intel
* **Problem of more data:**
  + - XLNet vs Bert – Was XLNet better cause of its ‘autoregressive architecture’ or because of more data?
      * -> Settled by RoBERTa, which was Bert, but trained on more data -> outperformed XLNet
    - Winograd Challenge: Understanding the meaning behind words -> The trophy did not fit into the suitcase cuz it was too big: -> What was too big, the trophy or the suitcase?
      * Solved by collecting & training sentences with two times the same word and replacing one of them with it;; Scientifically just used more data, not really enabled common sense
    - Power Law: Pretty much sticking to the hyperparams from ‘Attention is all you need’, but just increasing model size means better results (this has been happening since 2018)
    - We should compare apples to apples; how do models compare after the same amount of data trained on – ‘Sample efficiency’
  + **Problem of Robustness**
    - Not able to do ‘Zero-shot learning’ -> You always need to add that one layer and train it for transfer learning; Called Task-specific components; Limits efficient models
    - If you vary the fine tuning random\_seed big changes in performance
    - -> Add word types for each word during training & other linguistic informations
    - -> Add common sense (DEF: Basic level of practical knowledge shared among most people & encountered in everyday situations)
      * Reporting Bias – We do not state the obvious:
        + E.g. we know that ‘sheep are white’, but in texts we never really say that, because it’s to obvious – its an oxymoron in fact – Rather even texts talk about a black sheep, cuz its a common expression
        + -> If you ask one of those models, what color is a sheep – It’ll probably say black, cuz it appears more often in texts
        + IDEAS:

Add Knowledge Base (i.e. when it sees a sheep in the text, it goes into the knowledge base and checks what it knows)

Multimodal learning (i.e. it also learns from images)

Interactive/Human-in-the-loop training -> Model might ask the human

-> Many common-sense datasets have been built

ATOMIC

COMET

WINOGRANDE

* **Problem of Continual Learning:**
  + - BERT was trained on 2018 data – It will always think the president of the US is Trump

**Sources:**

<https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html>

<http://jalammar.github.io/illustrated-transformer/>