European Language Resource Coordination
Connecting Europe Facility

European Commission

LEGAL FAQ: SHARING LANGUAGE MODELS

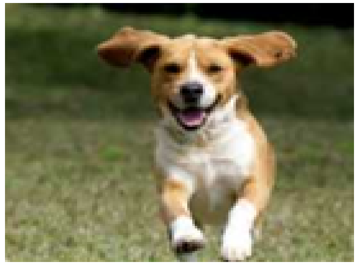
Mickael Rigault, ELDA

Khalid Choukri, ELDA

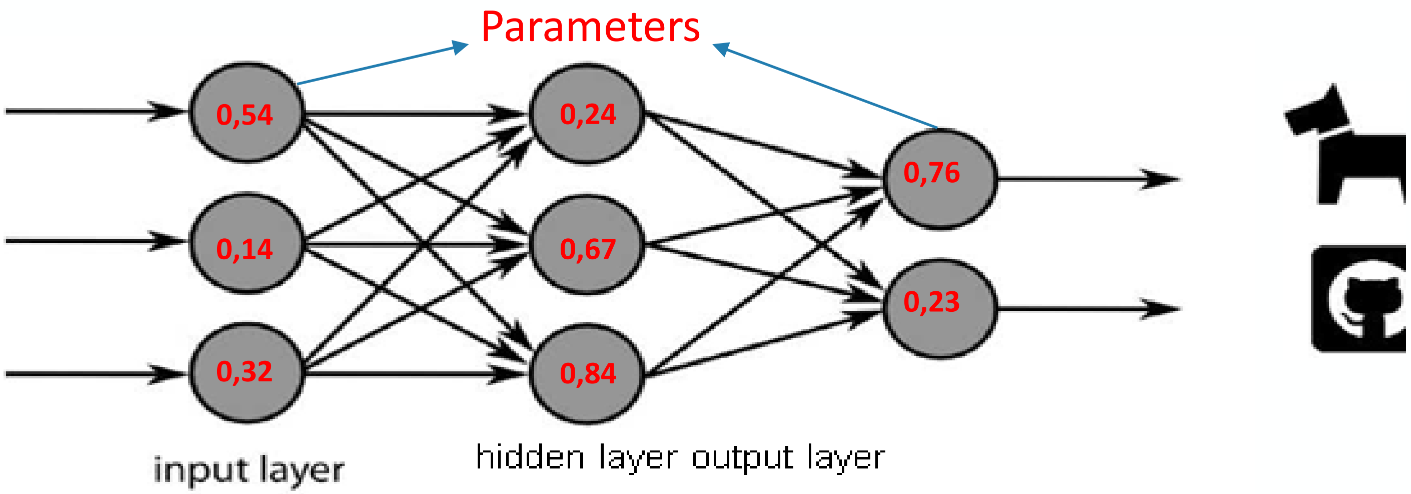
TECHNICAL DEFINITION OF A LANGUAGE MODEL

- Definition: It is the result of learning methods that allows the analysis and modelling of patterns of languages
- Important features of Language Models (LM):
  - LM are trained on Language Resources (LR). They are often tuned with specific resources (e.g. in a particular domain).
  - LM are trained with algorithmic methods (e.g. rule-based systems, neural networks).
  - LM Training approaches
    - Supervised Training : The algorithm learns on a labelled data set, providing an answer key that the algorithm can use to evaluate its accuracy on training data)
    - Unsupervised Training : The unsupervised model provides unlabeled data that the algorithm tries to make sense of by extracting features and patterns on its own).

Example of an Image Recognition Model



$$\begin{bmatrix}7\\2\\2\\4\\112\\3\\4\end{bmatrix}$$



Input

Encoded input

Architecture

Output

Source:
Wikimedia
pixy.org

Legal Considerations for LM Components

- Language Models (LM) are composed of different objects relevant to IP Protection, including:
  - An untrained model encoded into a Neural network architecture
  - Input data used to train the LM
  - Training algorithms coded into source code
  - Training model resulting from the input data and new parameters obtained through the training
  - Applications based on models (i.e. specialised engines relying on models)
  - Tuning data (domain specific data for finetuning the LM)

- Legal questions surrounding LM
  - Usability of input data (training & tuning data) fed into LM
  - Redistribution of LM & applications built on top of LM

Best Practices when selecting Input Data

- Language models are based on data so be sure that input data is usable to train a model
  - Proprietary licenses depending on the clauses
  - Data made available under permissive licences (Creative Commons and others) especially regarding derivatives.
  - Definition "Derivative": a work based upon one or more preexisting works, such as a translation, musical arrangement, dramatization, fictionalization, sound recording etc. or any other form in which a work may be recast, transformed, or adapted

  ➔ **Trained Model**: Derivative made of the untrained model + input data used
- Be careful with some Creative Commons Designations
  - ND: Forbid to redistribute derivatives (including the model)
  - SA: Obligation to share the derivatives under the same licence as the original
  - NC: Forbid commercial redistribution of the original data and derivatives (if commercial exploitation is intended)

Best Practices for Selecting LMs

- Conditions for sharing the model embedded with training algorithms and parameters
  - If you use open-source licenses prefer licenses tailored for software & database sharing (e.g. MIT, Apache, GNU-GPL, BSD)
    - Apache 2.0 and GNU-GPL 3.0 allow uses of patented elements
    - MIT & BSD do not allow use of patented elements
    - Allow wide reuse of the model (e.g. BERT is available under Apache 2.0 License)
    - May compel contributors to document modifications made to the model
  - A model and its applications can be made available under a proprietary license based on the access to the LM or distribution of the parameters

GDPR Issues Related with Language Models

GDPR applies in all cases where personal data is processed.

If personal data is included in the input data:

- LM training purpose falls under the scope of the GDPR
- Data Processor is responsible for complying with GDPR
  - Carry out Data Protection Impact Analysis before training (DPIA)
  - Inform Natural Persons on the purpose and their rights
  - Set up security measures (anonymisation, access rules, data storage)
  - Pay attention to transfers made to third countries and processors