University of Zurich[UZH]

EPFL AMLD

Transforming Education with Machine Learning

Practical Applications and Case Studies

Rudolf Debelak

University of Zurich, EPFL

Schedule

2:00 PM - Machine learning in education: an overview

2:30 PM - Prediction of student performance (tutorial)

3:30 PM - Coffee break

4:00 PM - Automated grading (tutorial)

5:00 PM - Summary and discussion

Q&A: at the end of each session or during the hand-on tutorials

Example Applications in Education

- In recent years, advanced large language models such as OpenAI's GPT-4o and Google Gemini have been developed, leveraging transformer architectures.
- These models are capable of processing and generating not only text but also other types of media, such as pictures and videos.
- In fields like education, these models facilitate tasks such as item (task) generation and automated essay scoring.
- We will showcase simple, illustrative examples using ChatGPT and Google Gemini on the following slides.

Example Applications in Education

Prompt to ChatGPT:

"You are generating multiple-choice questions for assessing reading comprehension. First, please generate a story of about 200 words aimed at English native speakers around the age of 14. Then, generate a question with four possible response options about the story you wrote. Only one response option should be correct. The question should be answerable based solely on the text, not by common knowledge."

Example Applications in Education

Story

In the small village of Willow Creek, the annual kite festival was the highlight of the year. Every summer, as the winds grew stronger, villagers and visitors alike would gather at the town square, kites in hand. This year, 14-year-old Emma was determined to win the grand prize with her hand-crafted kite, painted with the vibrant colors of the sunset.

Emma had spent weeks perfecting her kite's design, making sure it was both beautiful and aerodynamic. On the day of the festival, the square was bustling with excitement. Kids and adults tested the winds, running about with colorful kite tails fluttering behind them.

As the competition began, Emma launched her kite into the air. It soared high, its colors blazing brightly against the clear blue sky. However, midway through the competition, a sudden gust of wind caught her off guard, and her kite started to dive towards the ground. Heart pounding, Emma managed to regain control, pulling the string at just the right moment to stabilize it.

The crowd cheered as Emma's kite danced gracefully in the wind, finally clinching the top spot. Her months of hard work had paid off, and she proudly held up her prize—a hand-painted trophy that matched the colors of her winning kite.

Multiple-Choice Question

What was the main reason Emma won the kite festival?

A) She was the youngest contestant.
B) Her kite was the most colorful.
C) She expertly regained control of her kite.
D) She had the largest kite.

Correct Answer: C) She expertly regained control of her kite.

Example Applications in Education

Prompt to Google Gemini:

"You are asked to score an essay written by a human student on the topic: 'More and more people use computers, but not everyone agrees that this benefits society...' Write a letter to your local newspaper expressing your opinion on the effects computers have on people, persuading the readers to agree with you." Rate the following essay on a scale of 1 to 5. Describe its strengths and weaknesses.

An example essay:
To: The @ORGANIZATION1 can be a big part mof people's lives, however, some don't realize the negative effect these pieces of technology have. First, staring at a computer screen is unhealthy for your body. Secondly, anyone can post information on a website, therefore, it can be inaccurate. Lastly, people tend rely on computers too much. Computers are negatively impacting our society! The first reason why clearly, computers are not benefiting people is because they can cause unhealthy symptoms. [...]

Example Applications in Education

Score Point 4

This essay takes a clear position on the negative effects of computers and provides adequate support for this stance.

Strengths:

- **Clear Position:** The essay strongly argues that computers negatively impact society.
- **Organization:** The essay is well-organized, with an introduction, three body paragraphs (each addressing a different negative effect), and a conclusion.
- **Supporting Details:** Each reason is supported with examples or scenarios, such as eye strain, inaccurate information, and over-reliance leading to lost work.
- **Awareness of Audience:** The writer uses language and examples that would be relevant to a local newspaper audience.

Example Applications in Education

Weaknesses:

- **Some Generalizations:** While specific examples are provided, some statements are overly general, such as "Computers are obviously unfit for our bodies."
- **Lack of Counterarguments:** The essay does not acknowledge any potential positive effects of computers, which weakens its persuasiveness.
- **Minor Fluency Issues:** There are a few minor errors in grammar and punctuation.

Overall, this essay presents a somewhat developed response with adequate support. It could be improved by addressing counterarguments and refining some of the generalizations. However, it effectively communicates its position and provides sufficient evidence to persuade some readers.

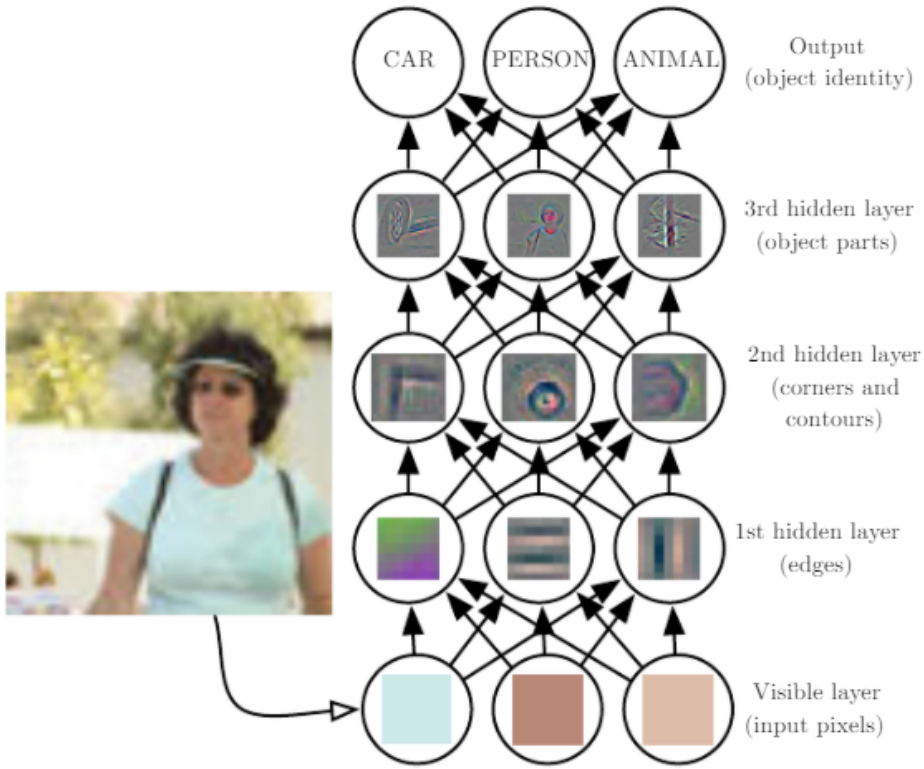
Natural Language Processing in Education

- **Machine Learning in Social and Behavioral Sciences:**
  - Task Automation
  - Here focus on **natural language processing (NLP)** for written and spoken text.
- **Key Applications:**
  - Text classification
  - Summarization of texts
  - Text generation
  - Translation
- **Challenges in Adoption:**
  - NLP models are complex and can be difficult for social scientists to implement.
  - **Transformer models** provide a breakthrough but remain underutilized in the field.

Artificial Neural Networks

- Artificial neural networks play a crucial role in many current machine learning applications, such as text translation, or the categorization of text and pictures.
- The name comes from the idea of simulating neural networks (e.g., of animals or humans). However, this analogy is criticized in the literature.
- The basic design is that features are processed in subsequent layers that transform this information to solve a task at hand.

Artificial Neural Networks



Goodfellow, Bengio, and Courville 2016

Artificial Neural Networks

- In classical applications of machine learning, researchers obtain features from the text by hand, and use machine learning methods for prediction.
- In applying deep learning models, these features are also generated by the models, enhancing their potential generalizability. However, deep learning models have higher computational demands than classical machine learning methods.

Artificial Neural Networks

In general there are four well-known types of neural networks:

- **Feedforward networks**: In these networks, information moves only in one direction—from input nodes, through hidden layers, to output nodes. There is no feedback (i.e., no connections loop back on themselves).
- **Convolutional Neural Networks (CNNs)**: Primarily used for processing data that has a grid-like topology, such as images. CNNs are especially powerful for tasks like image and video recognition.
- **Recurrent Neural Networks (RNNs)**: Designed to handle sequential data, such as text and speech. RNNs are used extensively in natural language processing.
- **Transformers** are primarily used for handling sequential data like text but with improved efficiency and performance over RNNs.

Transfer Learning

- Transfer learning leverages pre-trained models to save on computational resources and training time, particularly beneficial when dealing with large datasets or complex classification tasks like distinguishing between hundreds of categories.
- This approach utilizes the knowledge gained from previous training on a broad dataset and adapts it to a new, often more specific task.

Transfer Learning

- In typical applications, you combine a pretrained foundation model with transfer learning. Via transfer learning, you adapt the model to a task at hand.
- Transfer learning is especially useful in situations with limited data, as they are common in education.
- It starts with layers trained to identify universal features (e.g., textures, shapes) and integrates new layers tailored to the specific task, optimizing for unique challenges without starting from scratch.
- Many foundation models are based on the transformer architecture.

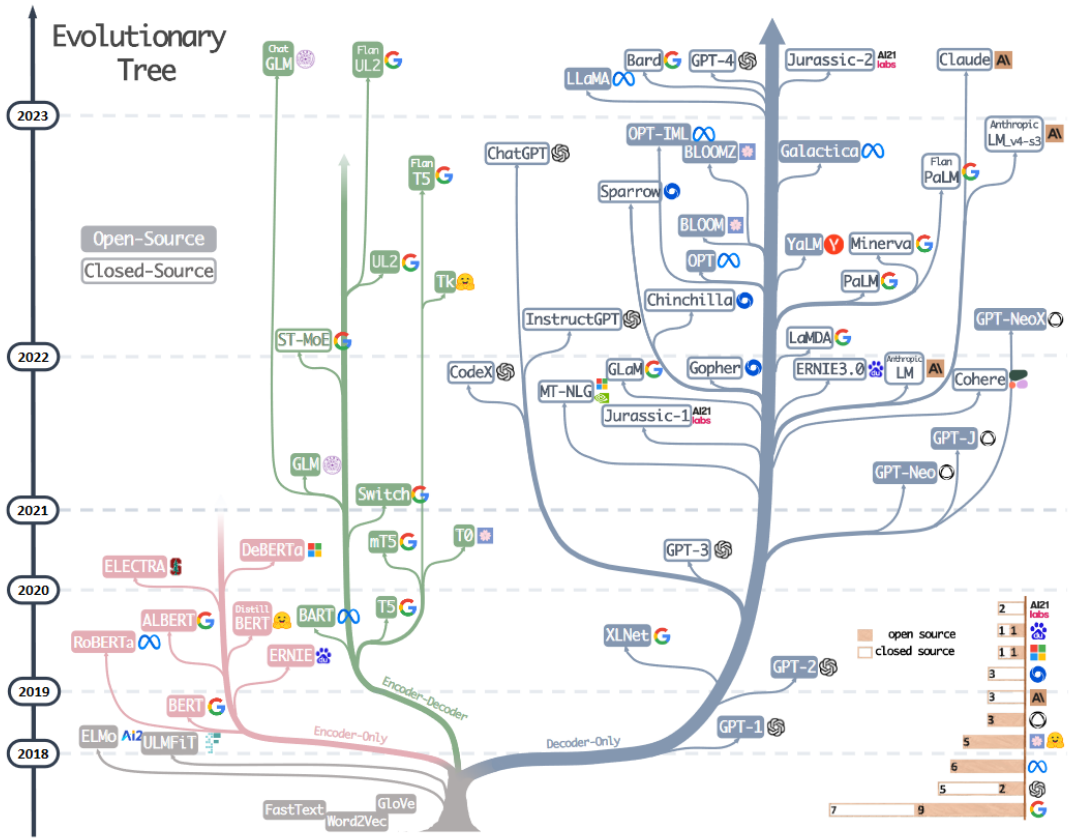
Transformer Models: An Overview

- **What is a Transformer Model?**
  - A deep learning architecture designed to process sequential data like text.
  - Based on **Attention Mechanism**
- **Key Components of Transformer Models:**
  - Encoder
  - Decoder
- Initially used for translation tasks Vaswani et al. 2017

Transformer Models: An Overview

- **Parallel Processing**
- **Evolution of Transformer Models**
  - From encoder-decoder models to separate **encoder-based** and **decoder-based** models.
  - **Encoder-based models:** Focus on text understanding.
  - **Decoder-based models:** Focus on text generation.
  - Modern applications use large language models (LLMs) such as GPT-4o for creative tasks.

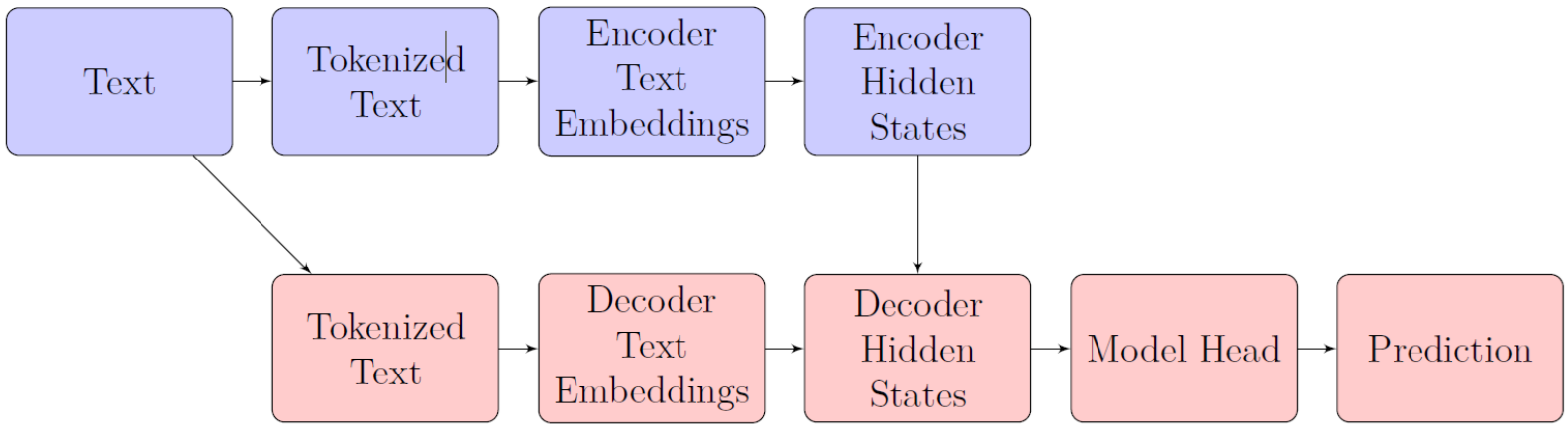
Transformer Models: An Overview



Transformer Models: An Overview

- **Educational Assessments:**
  - Automated **text scoring** and task generation.
  - Supporting teachers by providing feedback and recommendations.
  - Plagiarism detection
  - Personalized feedback

The Overall Structure of Transformer Models



Application Steps of Transformer Models

1. **Task Definition:** The input is a text, the output can be numerical or categorical.
2. **Data Preparation**
3. **Model Selection**: e.g., BERT
4. **Fine-Tuning and Model Evaluation**

Black Box Nature of Transformer Models

- Large language models lack transparency.
- Possibility of prediction errors and systematic biases.
- The application of biased models may reinforce existing biases.
- For instance, in psychological assessment, this may affect their fairness and validity.
- The black-box nature of transformer models can lead to ethical risks.

Ethical Challenges in AI

This directly relates to some ethical challenges of AI systems. These include:

- Bias in training data
- Fairness across demographics
- Transparency and accountability

Examples:

- Model assumes that essays about soccer are written by boys, and essays about gardens are written by girls.
- Model penalizes students who have German as a second language disproportionately.

Solutions:

- Systematic auditing of training data and regular bias testing are critical to ensuring fairness.
- We can address these challenges by a systematic analysis, which reminds of the evaluation of psychological assessments.

Reliability

We now discuss this evaluation in the context of essay scoring.

- Reliability measures the precision of test scores.
- Proposed Methods:
  - Interrater reliability
  - Split-half analysis
  - Robustness analysis: Introduce small perturbations (e.g., typos) and measure stability.
- Examples: Intraclass correlation, split-half correlation.

Validity

- Validity refers to the correctness of the score's interpretation.
- Evidence Types:
  - Evaluating the effect of changes on the scoring.
  - Internal structure: Agreement between model outputs and average human ratings.
  - External variables: Correlation with external criteria.
- Examples: Evaluating the effect of typos on essay scores, prediction error.

Fairness

- Fairness ensures scores are free from bias across groups.
- Proposed Methods:
  - Compare model accuracy across demographic groups.
  - Evaluate bias: Check for systematic over- or under-scoring for specific groups.
- Example: If longer essays are systematically rewarded, this could disadvantage students with limited writing proficiency or non-native speakers.
- This corresponds to professional standards in educational testing (ITC; APA, NCME and AREA).

Additional Evaluations

- **Dataset Auditing:**
  - Identify over- or under-representation of groups = Checking the representativity of the training, validation and test data.
- **Adversarial Testing:**
  - Modifies input data (e.g., changing names) to expose bias.
- **Transparency and Explainability Tools:**
  - SHAP, LIME
  - Attention Visualization

Practical Application

- We showcase the evaluation of a machine learning model for essay grading using the training data from the Automated Essay Scoring Competition hosted by the Hewlett Foundation on Kaggle,
- These data consist of more than 10,000 essays on eight different topics and are available online (https://www.kaggle.com/competitions/asap-aes/data).
- We offer a Jupyter notebook on Google colab to follow the step of the analysis.
- In the exercises, you will explore this data set further using an alternative approach.

Practical Application

In the Jupyter notebook, we investigate the following steps:

- Preprocessing (e.g., getting all grades on a range from 0 to 1)
- Applying a DistilBERT model.
- Comparing the evaluation for different parts of the text.
- Evaluating the effect of inserting random letters.
- Comparing the accuracy of the prediction over different topics.
- In the Jupyter Notebook, we treat our criterion - the variable we want to predict - as metric. In the exercise, we will treat it as categorical.

Practical Application

Possible further checks:

- Dataset Auditing
- Predictive Parity
- Further adverserial Testing
- LIME, SHAP, Attention Visualization

Takeaways: Deep Learning and Neural Networks

- **Neural networks** provide powerful tools to model complex data patterns in domains such as text, images, and audio. They are also capable of **generating new content** such as text and images.
- **Generative AI models** can offer substantial support for research tasks in fields like **education**, from item generation to automated essay and picture scoring.
- Their practical application offers much promise, but requires rigorous testing using methods outlined in this presentation and the exercises.

Limitations

- The proposed methods are specific to numerical outputs from NLP models.
- Lack of general benchmarks for comparing AI models in education.
- Just as psychological tests have frameworks for reliability and validity, AI-based systems require standards for transparency, fairness, and ethical use.
- Future Work:
  - Extend methods to non-numerical outputs (e.g., verbal feedback).
  - Develop benchmarks for essay scoring systems in education.
  - Develop guidelines for data representativity, fairness metrics, and systematic bias detection.

Conclusion

- Proposed methods to evaluate reliability, validity, and fairness in AI-based scoring.
- Empirical results demonstrate the utility of these methods.
- Contribution: Framework for integrating AI into educational assessments while adhering to testing standards.

Summary of Key Takeaways

- Transformer models and other neural networks offer great potential in education and related fields. Compared to classical statistical methods, they can directly use texts, pictures and other media as input.
- They can work on many tasks with near-human performance while being inexpensive.
- Their interpretation is not straightforward, and we need to evaluate their validity and fairness diligently.

Tutorial 1

- Training data from the Automated Essay Scoring Competition hosted by the Hewlett Foundation on Kaggle (https://www.kaggle.com/competitions/asap-aes/data).
- The data set consists of 12976 essays written in English on eight different topics.
- All essays were rated by at least two human raters, whose average rating served as target score. Furthermore, all personally identifying information was removed.

Tutorial 1

- Use the provided Jupyter Notebook to run the analysis from the presentation in the workshop. Here, we treat the target score as metric. Make sure you understand what each step is doing.

https://tinyurl.com/AMLD25
AMLD Workshop - Excercise 1.ipynb

Tutorial 2

- AMLD Workshop - Excercise 2.ipynb
- Use the provided Jupyter Notebook to run an alternative analysis, where we treat the target score as categorical. Again, we carry out several analysis steps, including:
  - Comparing the evaluation score for parts of the essays. (Reliability)
  - Comparing the prediction accuracy of DistilBERT and human raters for groups of essays. (Fairness, Validity)
  - Assessing the robustness of the prediction against minor changes. (Validity)

  Which results suggest that the model is reliable, valid and fair?

References I


Goodfellow, Ian, Yoshua Bengio, and Aaron Courville (2016). *Deep Learning*. http://www.deeplearningbook.org. MIT Press.


Vaswani, Ashish et al. (2017). "Attention is All you Need". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc.