

Comparison of LSTM and Transformer Architectures for Polish Language Modeling

Course: Computational Linguistics
Date: 27.10.2025r. **Author:** Nicolas Stupak

1. Introduction

This report presents a comparative analysis of two neural language model architectures: Long Short-Term Memory (LSTM) and Transformer models, both trained for causal language modeling on Polish Wikipedia dataset. The goal is to evaluate their performance in terms of perplexity, training efficiency, and generation quality.

2. Model Architecture

2.1 Shared Configuration

Both models were configured with comparable parameter counts to ensure fair comparison:

Hyperparameter	Value
Embedding Dimension	384
Hidden/Model Dimension	384
Number of Layers	6
Vocabulary Size	200019 (tiktoken encoding for gpt-4o)
Dropout Rate	0.2
Context Window	256 tokens
Optimizer	AdamW
Learning Rate	3×10^{-4}
Batch Size	64

2.2 LSTM Architecture

The LSTM model consists of:

- Token embedding layer (vocab_size × 384)
- 6-layer stacked LSTM with 384 hidden units per layer
- Dropout between LSTM layers (0.2)
- Layer normalization
- Linear output projection to vocabulary

Total Parameters: 160 million

Key Features:

- Sequential processing with recurrent connections
- Hidden and cell states maintain context
- Bidirectional information flow through gates (input, forget, output)

2.3 Transformer Architecture

The Transformer model (GPT-style) consists of:

- Token embedding layer (vocab_size × 384)
- Positional embedding layer (256 × 384)
- 6 transformer decoder blocks, each containing:
 - Multi-head self-attention (6 heads)
 - Feed-forward network (384 → 1536 → 384)
 - Layer normalization and residual connections
- Linear output projection to vocabulary

Total Parameters: 164 million

Key Features:

- Parallel processing via self-attention mechanism
- Causal masking for autoregressive generation
- Positional encodings to capture sequence order

3. Dataset

3.1 Data Source

Polish Wikipedia (plwiki) corpus accessed via the Speakleash library:

- Total documents: ~588k articles
- Language: Polish
- Task: Causal language modeling (next-token prediction)

3.2 Data Split

Split	Proportion	Documents	Tokens (approx.)
Training	90%	~543k	~339 milion tokens
Validation	10%	~30K	~38 milion tokens

3.3 Tokenization

Tokenizer: TikToken (GPT-4o-mini encoding)

- Vocabulary size: ~200k tokens

- Byte-Pair Encoding (BPE) algorithm
- Handles Polish diacritics and morphology

4. Evaluation Results

4.1 Perplexity

Perplexity measures how well the model predicts the test set, with lower values indicating better performance.

Model	Train Loss	Train PPL	Val Loss	Val PPL
Transformer	3.8881	48.82	3.9041	49.60
LSTM	4.8824	131.95	4.9542	141.78

Interpretation:

- Perplexity difference: 285% better for Transformer over LSTM
- The Transformer demonstrates superior language modeling capability

4.2 Time Efficiency

Training Time

Model	Total Training Time	Time per Epoch	Tokens per Second
Transformer	3:20h	6s	107.19
LSTM	2:14h	4.07s	310.30

Analysis:

- Transformer training is slower despite its parallel processing
- GPU utilization: during training both models achieved ~99% GPU usage

Inference Time

Model	Total Inference Time	Tokens Generated	Mean Time per Token	Tokens per Second
Transformer	8.62 s	1000	8.62 ms	115
LSTM	3.19 s	1000	3.19 ms	313

Analysis:

- LSTM is 2.7× faster at inference
- LSTM's sequential nature creates a bottleneck for generation
- Transformer benefits from parallel attention computation during training but is still autoregressive during inference

5. Prompt Completions

Ten Polish prompts were used to evaluate generation quality:

Example 1

Prompt: "Warszawa to stolica"

Transformer: Warszawa to stolica Myko. W swojej przewidciu holenderskiej w Los Angeles związał nie pozwalając na podzięków, że bagne jeżeli że powstały wątpliwe próby naciszenie. Axówalu, który wybuchła się do ugrupowanie białostocki, pyk miejskiego nie obożytnego oddziału Episk nazwilnych. Dziecko ę antidencja kazała o po Stanisław Lud

LSTM: Warszawa to stolica w gminie Doczyzky w Moswersgosania, w powiecie burgorem, w regionie Wileńsku, w skużu, w tym utworzimanu, w departorencie Dokola. 1 lipcem konbacyję w móiergu stółował należał z brytką. W tym czasie powieł wszjął pochynę między firmą kongby. W1 1643 roku miasta studi w

Example 2

Prompt: "Historia Polski rozpoczęła się"

Transformer: Historia Polski rozpoczęła się w okolicy Świętego Gorzów w 1930 we Lwowie, dwuparii spoczy I przełomie 3-1623 mosty je dolinę i SzymanówPolskich. W 1939 pod Czoły Piechoty przy Białeglariczne Tadeusz Habilitowcyli wypowiedzielono dwa stronę polskberyjską za zadanie lotnych lotnictwa do Krzawczy

LSTM: Historia Polski rozpoczęła się w stoskach "Solidozyą". Wprowadzowaniu w 1945 roku przeszedł do lalka miejskiej i w powiecie p osób duszterskie ze Tarnami Jana. W 1983 raz był w parawersizyury jako Irryów Rundenta Peławskich i stanowteryczasowej momentem, a w jego samym jednostkami paraakcji "Na inicjatywnym" na gimfencji Beskii w Pozbanu

Example 3

Prompt: "Największą rzeką w Polsce jest"

Transformer: Największą rzeką w Polsce jest głównie w Rotterdamie World-Br toerman'ektularnych. Jest to granicą administrację tramwajów. Glossiro ojców chł departure przypisywano do pierwotnej marynarki, prowadzi wodnej jednocześnie był trzeci jedynie Harry'ego i Karpat Colts States'a Bierdu. W "szym wielkim tematem warmiledku farnych". Działalność. **LSTM:** Największą rzeką w Polsce jest „Agio”, a następnie białowskim. Dwa jego 4 tys., pracowni opijały między Morzyka Błował. Do tym pięciu starryw do jest słowiany wydarzenia Ody, określający w światach technologiczną delegacji Wrocławiu. Odwieziono nad zwnżący części pieów wojnie, od rodzaju, ale uznasa zamyna pierląskiego do wchododzie od szkociu

Example 4

Prompt: "W roku 1945"

Transformer: W roku 1945 roku, za wojska sowieckie zamieszkiwiora produkującegoieszki i burzę Pancer wskazębowy zalesieni Krzymyskiego oraz dworobw. Okuszcę moce: Pał II, Guzinto, (; ukroczytnica krzyżemica, znajdujący się na południowym: mało kościół III Krzy ruszyżem a następnie na Powązkows

LSTM: W roku 1945 roku. Po pierwszym tej kość 1949, 2021 i 1947 miały przedorganizowanego pieniytowie. W 1647 roku został albumem pracenika Uasencjką. W 2005 roku kierowany w kolejizach

Towienizyst Polski Wnie. Skóźniej była pierwsowany w willie transport (2015), 1998/1999). U 2014 powrorzono treningę wagi Rady

Example 5

Prompt: "Polscy naukowcy odkryli"

Transformer: Polscy naukowcy odkryli ludzi. W Galaxy ówstwach Czeaka próba ten cel wielkieali przez kryłęgarzy (303,33–200 tys. ludzi byli w Stanach Zjednoczonych). Uczył, do niej miot Grażera nieskie podjęli na naturę, a kopcaloążność. DrugaLat Filip.Honorowy pod czasem piezaztę salaryPhot Gonzagiellistowskiej wymiarą Maroy – nazystyk

LSTM: Polscy naukowcy odkryli Polową podróżę. W partkim, Johannes-6 kwedzia przeciwko przekazała od niego ponano 450 500 metrów (vensesis)): znalazając eksportę 700 i osób górku na gminowej, w kraju którego w pobólsku wrocwie, 14 lipca 1923 r. w wieku samą Wojennego, która w wieku samym pozodniysci mieli w organizacji Przewyskowej

Example 6

Prompt: "Kraków jest miastem"

Transformer: Kraków jest miastem Polski i opartalsze komisarza. Po zakończeniu lutowej, dochodządu, frakcjami. Formużył członek Urg Mickiewicza i obrazymonii (erg Menzi) oraz mieszka Zespołu RP i w Katony Hebole.I i Narze Ta panowanie i otrzymał współ Festiwum Thomasa Alberto Od 70 lat. Pochodziły z Celśodość na Sorowie, **LSTM:** Kraków jest miastem obrzające atuską, wysze Nalarczyła, iż stężono, działane w ciągu na grogu pracowników. Ddy krótryńskich została przeniesiony z Szumbą, z którym Mickachowała dostać warne z szżyńowała od rozstrówżelił 2 osoby, a zańu ojca, 4 km, od czym żyzył nie w tej zbółtienia 13 razy i eksper

Example 7

Prompt: "Literatura polska słynie z"

Transformer: Literatura polska słynie z niej sekretarzem. Dziecioryto odegrał po 4 roku w bitwie pod Syczką. Był William Tego w Mielski jako Marian Wierzkom została ranny w narodzinie starostw Europy. Przeprowadził 29 maja 1946 we Lwowie. Swysie pracował własną uczniemry pełnił obowiązkiem do autonomii Stańczędzik. 9 września 194 **LSTM:** Literatura polska słynie z tym nim wylegli w mwację z rzinowych lokalnością, której został wyjeczyć. W roku 30 pierwskania widaku Gjła Humany, jego samaticie barza, którą ogrodu, a także pod czym jednym pełniej zagrała do Turcję wiele poślgiowo się „Milab”. Zapostiosło do. Maże widowęczyzndzka Towanu weszńskojelowego (jedostał

Example 8

Prompt: "Podczas II wojny światowej"

Transformer: Podczas II wojny światowej podziału II przedziału odznaczonych podanie rozpoczęto decyzję o szklonejowanie walk, czwartym przygotowanym poślub, takie oddaniu 21 wykazał obroju przest wieczorem między reszcie 66. W okresie odjaśnieniami. 1 czerwca 1894 Anny – został obiekcie traktowany w ŻGGku i oddziału, chrześcyrowskim **LSTM:** Podczas II wojny światowej, sąca awawił się w jego kolejne Uwinnie samo władze dzolnicza się jedował w tranej Cmentarzu prowozyku z południch Karwickiem zakdze. W marcu 1983 roku występował w gminie Gronau. Kłęgaskuż Cars nie ukończył w dóścjoych skrzecictwa Republy USA. Pod koniec 2010 został założony w województ

Example 9

Prompt: "Polska kuchnia jest znana z"

Transformer: Polska kuchnia jest znana z członkami prz imaginary Rawy. Obaj grzybki miał groć rozwinęczeń drzwi, w czasie społeczeństwo, są krzyżowa popularna polityczna. Drugi autorówokkion przetka. Katolickiego w języku wraca nowy, w czasie Rosyjskich papieża lubymalna skoro, czyli Kotliny Oometryka i Rapp i Pelmacawa Konna. Jemu wielkoanga uważ

LSTM: Polska kuchnia jest znana zdez czym turię stanie lubkowiska Próluwańskiego kraskowa niżwarkami. Liczyksniętnie, z października 1909 roku i pozostonał nagyscy do cędzynarodowej nauką w Komisji Komerytowskiego bursą postaczarowych „W 1897 roku góuke Franciszańskiego) (We. "Ta Swrawiewnego"), „Wspawane Popularacyjne czy ho-

Example 10

Prompt: "Jan Paweł II był"

Transformer: Jan Paweł II był dreemdzemu kościołem. Śmierci mający Lubazwa Budowa – został wybudowany, potem św. we Krzyżem Oficerskiej został od Bard pole łatwy druku służył do dziś i połowy XIV wieku w drugiej, czyli Związku Piusa ostatecznie odjmów Syje. W trakcie obsada do dziś i Święt Conditionsnych. Do bróśł pojemności zrół w Koś

LSTM: Jan Paweł II był powołany do przech. Od listopada 1959 wykołgowisko w Zielziale Komwatowych. W 2010 wygłaczył w East 劳 zajął 6 maja 2005 roku. W 1994 do zatryską pracę w staroli narodowania na Ukrainie w 1988. W listopada 1778 roku odlica podadebiutował w 2001 roku. Andrąc swoim ciężjonizki w

Qualitative Analysis

Coherence: Transformer completions were generally more coherent and contextually relevant, however both models struggled with producing coherent text. **Factual Accuracy:** Both models didn't produce factually accurate information.

Grammatical Correctness: Both models struggled producing grammatically correct Polish sentences, with Transformer showing slightly better fluency. **Long-range Dependencies:** Transformer handled longer contexts better, resembled any consistency compared to LSTM.

6. Discussion

6.1 Key Findings

1. **Perplexity Performance:** The Transformer achieved 63% lower perplexity, indicating better next-token prediction capability.
2. **Training Efficiency:** LSTM trained 32% faster, demonstrating the trade-off between parallelizability and sequential processing.
3. **Inference Speed:** LSTM generated text 2.7× faster, with 313 tokens/second throughput.
4. **Generation Quality:** Qualitatively, Transformer produced more coherent and contextually appropriate completions, particularly for longer sequences.

6.2 Architecture Trade-offs

LSTM Advantages:

- Simpler architecture, easier to implement and debug
- Lower memory footprint (no attention matrices)
- Constant memory per time step

LSTM Disadvantages:

- Sequential processing limits parallelization
- Difficulty capturing very long-range dependencies
- Slower training on modern accelerators

Transformer Advantages:

- Parallel attention mechanism enables efficient training
- Superior at capturing long-range dependencies
- State-of-the-art performance on most NLP tasks

Transformer Disadvantages:

- Quadratic memory complexity with sequence length
- Requires more parameters for comparable performance
- Positional encoding limitations

6.3 Implementation Challenges

1. **Data Processing:** Handling 588k Wikipedia articles required efficient streaming and batching strategies using the Speakease library.
2. **Tokenization:** TikToken's English-centric vocabulary was suboptimal for Polish; a custom BPE tokenizer would likely improve efficiency.
3. **Memory Management:** Training on my local machine was unreachable due to out-of-memory errors.
4. **Evaluation:** Ensuring fair comparison required identical hyperparameters and careful attention to random seeds.

6.4 Insights

- **Polish Language Specifics:** Polish morphological complexity (7 cases, 3 genders, rich inflection) presents unique challenges that Transformer handled better
- **Scalability:** Results suggest that Transformer architecture would benefit more from increased model size and training data

7. Conclusion

This study compared LSTM and Transformer architectures for Polish language modeling. The Transformer demonstrated superior performance with 63% lower perplexity but at cost of 2.7 slower inference.