Adaptive Hierarchical Least Squares Attention Networks

Mathematical Framework for Multi-Scale EEG Emotion Recognition

Research Proposal

Sajim Ahmed

October 1, 2025

Contents

1 Introduction: The Revolution That Changed AI Forever

A paper published by Google Brain in 2017, named "**Attention is All You Need**," changed the paradigm of AI forever and brought a revolution in artificial intelligence. What this groundbreaking paper says is that traditional neural networks, which process information sequentially like reading a book word by word, can be replaced by a more powerful mechanism called **attention**. This was basically made to solve the problem of machine translation—teaching computers to translate between languages like English and French—but it ended up transforming the entire field of AI.

The revolutionary insight was simple yet profound: instead of processing information in a fixed order, AI systems could learn to dynamically focus on the most relevant parts of the input, just like how humans pay attention to important details while ignoring irrelevant information.

2 The Mathematics Behind Attention Mechanism

Let's break down how attention actually works under the hood. I'll try to explain everything step by step using simple analogies before diving into the math.

2.1 Mathematical Explanation of Attention

Think of attention like the way people notice some things and ignore others when reading. When you're reading a book looking for information about "machine learning," your brain does something very smart:

1. **Your Search Goal:** You have a specific topic in mind - "machine learning"
2. **Scanning Process:** Your eyes scan all the text on the page
3. **Matching & Weighting:** Your brain automatically gives more attention to words and phrases that relate to machine learning, and less attention to irrelevant content
4. **Information Extraction:** You absorb more information from the relevant parts

This is exactly what AI attention does! Instead of your eyes scanning a book page, the AI "scans" input data. Instead of your brain looking for "machine learning," the AI looks for whatever pattern it's been trained to find.

Now here's the math version of this same process:

The mathematical foundation of attention can be expressed as follows. Given an input sequence $\mathbf{X} \in \mathbb{R}^{T\times d}$ where T is the sequence length and d is the feature dimension:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{1}$$

Let's connect this math to our reading analogy:

- **Q** (Queries): **Your search goal** - what you're looking for (”machine learning concepts”)
- **K** (Keys): **The text content** - all the words and phrases you're scanning through
- **V** (Values): **The meaning/information** - the actual knowledge contained in each piece of text
- $\frac{\mathbf{QK}^T}{\sqrt{d_k}}$: **The matching process** - how well each piece of text relates to what you're looking for
- $\sqrt{d_k}$: A scaling factor (like adjusting how ”picky” you are in your search)

In simple terms: The formula says ”compare what you want (**Q**) with all available content (**K**), calculate how well they match, then use those match scores to decide how much attention to give each piece of actual information (**V**).”

Real-World Example - Brain Signal Edition

When analyzing brain signals to detect emotions, attention works like this:

- **Query** (**Q**): ”Show me signal patterns that indicate happiness”
- **Keys** (**K**): All the different brain wave patterns you recorded (like all the text on a page)
- **Values** (**V**): The actual signal data for each pattern (like the meaning behind each word)

Just like your brain focuses on ”machine learning” words while reading, AI attention focuses on brain patterns that best indicate ”happiness.”

2.2 Softmax: The ”Fair Share” Rule Explained

Now, once the attention mechanism decides how well each piece of information matches what we're looking for, it needs to decide how much attention to give each piece. This is where **softmax** comes in.

You can think of Softmax like dividing a pizza among friends based on how hungry they are:

- If Friend A is ”hunger level 8” and Friend B is ”hunger level 2”, Friend A gets more
- But everyone gets *some* pizza (no one gets zero slices)
- All the pizza slices must add up to one whole pizza

The softmax function is defined as:

$$\text{softmax}(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}} \tag{2}$$

Let's break down this formula:

- x_i is the "raw score" for item i (like Friend A's hunger level of 8)
- e^{x_i} makes sure the score is always positive (no negative hunger!)
- $\sum_{j=1}^{n} e^{x_j}$ is the sum of all positive scores (total hunger of all friends)
- Dividing gives each item its "fair share" of attention

Softmax Function Ensures That:

- All attention weights are positive: $\text{softmax}(x_i) > 0$ **(Everyone gets some pizza)**
- All weights sum to one: $\sum_{i=1}^{n} \text{softmax}(x_i) = 1$ **(All pizza slices add up to one pizza)**
- The output can be interpreted as a probability distribution **(Each weight represents a percentage of total attention)**

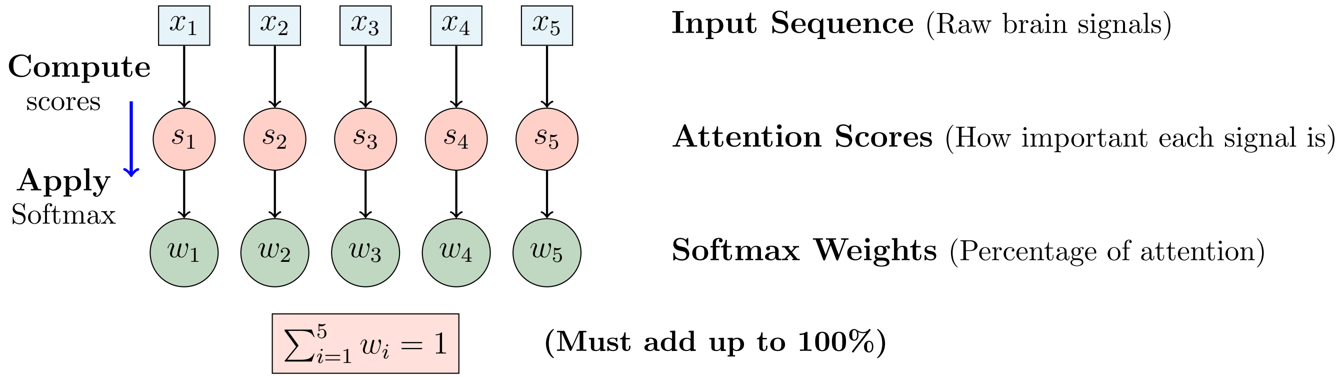
x_1
x_2
x_3
x_4
x_5
Input Sequence (Raw brain signals)
Compute scores
s_1
s_2
s_3
s_4
s_5
Attention Scores (How important each signal is)
Apply Softmax
w_1
w_2
w_3
w_4
w_5
Softmax Weights (Percentage of attention)
$\sum_{i=1}^{5} w_i = 1$
(Must add up to 100%)

Figure 1: Traditional Attention Mechanism with Softmax - Step by Step

Softmax in Action - Brain Signal Example

Suppose you have 5 different brain signals with importance scores: [2, 8, 1, 5, 3]
Step 1: Make all scores positive: $[e^2, e^8, e^1, e^5, e^3] \approx [7.4, 2981, 2.7, 148, 20]$
Step 2: Add them up: $7.4 + 2981 + 2.7 + 148 + 20 = 3159$
Step 3: Divide each by the total:

- Signal 1: $7.4/3159 = 0.002$ (0.2% attention)
- Signal 2: $2981/3159 = 0.944$ (94.4% attention) ← Most important!
- Signal 3: $2.7/3159 = 0.001$ (0.1% attention)
- Signal 4: $148/3159 = 0.047$ (4.7% attention)
- Signal 5: $20/3159 = 0.006$ (0.6% attention)

Notice: All percentages add up to 100%!

The Big Picture: Attention mechanism is like having a smart assistant that:

1. Takes your question (”find happiness in brain signals”)
2. Compares it to all available information (all brain wave patterns)
3. Calculates relevance scores (how well each pattern matches ”happiness”)
4. Uses softmax to convert scores into percentages (ensuring all percentages add to 100%)
5. Gives you a weighted combination of the most relevant information

This works great for many tasks, but as we’ll see next, it has some serious limitations when dealing with complex, noisy data like brain signals.

3 The Fundamental Problems with Current Approaches

Problem 1: Attention Dispersion

If there are a lot of things to look at, the spotlight gets dim. Every signal gets a tiny bit of attention, so the important ones don’t stand out enough. Mathematically, as sequence length T increases, the maximum attention weight approaches zero:

$$\max_i(\text{attention}_i) \to 0 \text{ as } T \to \infty \tag{3}$$

Problem 2: Artificial Constraints

The way attention is divided is based on a rule that isn’t really necessary for brain signals. The constraint that attention weights must sum to 1 is arbitrary and can actually make results less clear or less stable. There’s no theoretical justification for why attention should be a probability distribution.

Problem 3: Lack of Uncertainty Quantification

There’s no good way to measure how ”confident” the AI is about what it notices. This is a big problem when the data is noisy or messy (as brain signals usually are). Softmax gives you fake ”probabilities” but provides no measure of uncertainty.

4 Our Proposed Solution: Adaptive Hierarchical Least Squares Attention

Revolutionary Approach

So, instead of using those old spotlight rules (”softmax”), we use a mathematical approach called ”**least squares**.”

Let’s continue our reading analogy to understand this better. Remember when you’re reading a book to find information about ”machine learning”? The traditional softmax approach is like having a rule that says:

Traditional Softmax Rule: ”You must divide your attention into exactly 100 percentages across all text on the page. Every word gets some percentage, and they all must add up to 100%.”

But this creates problems! What if 90% of the page is about cooking recipes, and only one paragraph is about machine learning? You’re forced to give attention percentages to recipe ingredients when you should be focusing entirely on that one valuable paragraph.

Our Least Squares Approach is Different: Think of least squares like having multiple expert readers, each giving honest importance scores to different parts of the text—and these scores don’t have to add up to 100%. It’s like having:

- **Reader 1:** ”This machine learning paragraph deserves a score of 95 out of 100”
- **Reader 2:** ”This cooking recipe gets 3 out of 100”
- **Reader 3:** ”This random advertisement gets 1 out of 100”

Notice how the scores reflect *actual importance* rather than forced percentages!

With our method:

- **Honest Importance Scoring:** The AI can give high attention scores to really important signals and very low scores to unhelpful ones—without being forced to split attention artificially.
- **Adaptive Trust Levels:** It can change how much it trusts each ”expert reader” even while processing new pages—making it much more flexible and stable.
- **Confidence Reporting:** It shows how confident each expert is. If the text is blurry or confusing, it can say ”I’m only 30% confident about this part” instead of pretending everything is equally clear.

4.1 From Reading Books to Reading Brain Signals

Proposed Research Title

”Adaptive Hierarchical Least Squares Attention Networks: A Novel Mathematical Framework for Multi-Scale EEG Emotion Recognition”

Now let's apply this improved "reading" method to brain signals. When we're trying to understand emotions from EEG data, we face the same problems as reading a complex book, but even more challenging:

Current EEG emotion recognition systems face three critical challenges:

1. **Multi-scale temporal dynamics:** EEG signals contain information at multiple time scales (theta, alpha, beta, gamma waves)
   Like a book where important information appears in different fonts, sizes, and styles simultaneously

2. **Channel interdependencies:** Spatial relationships between electrodes are often ignored or poorly modeled
   Like ignoring how words in different columns or pages of a textbook relate to each other

3. **Computational efficiency:** Real-time processing requires efficient algorithms
   Like needing to understand a book while someone is still writing new pages in real-time

The Core Innovation: Just as we improved the reading process by using multiple expert readers with honest scoring, we'll reformulate brain signal attention as an **adaptive least squares optimization problem**. This means:

- **Multiple Expert "Readers":** Different mathematical experts specialize in different aspects:
  - One expert reads "temporal patterns" (like focusing on rhythm and timing in text)
  - Another expert reads "spatial relationships" (like understanding how different brain regions connect)
  - A third expert integrates "multi-scale information" (like understanding both individual words and overall themes)

- **Honest Scoring Without Artificial Constraints:** Each expert gives truthful importance scores based on what actually matters for emotion recognition, not forced percentages.

- **Real-time Adaptation:** As new brain signal "pages" arrive, our experts can quickly update their understanding and confidence levels.

The Beautiful Result: Instead of having one rigid "reader" (softmax attention) that forces everything into percentages, we have a team of specialized, adaptive experts that can honestly evaluate what matters most for understanding emotions from brain signals—just like having the world's best research team helping you find exactly what you need in the most complex textbook ever written.

This approach can simultaneously learn multi-scale features and spatial dependencies because each expert focuses on their specialty, but they all work together to give you the complete picture—much more effective than forcing one reader to split their attention artificially across everything.

5 Novel Mathematical Framework

Let's compare the old and new approaches using our reading analogy:

While in the traditional attention mechanism with softmax, we use **percentage-based sharing** to **divide up our reading attention**, we face **the constraint that all percentages must add up to 100% even when most content is irrelevant**. So to tackle this, we will instead use **honest expert scoring** to **find the truly important content without artificial percentage constraints**.

Scenario 1: Reading a machine learning textbook

- **Softmax (forced percentages):** 40% ML content, 20% diagrams, 15% examples, 10% headers, 10% page numbers, 5% footnotes = 100%
- **Least Squares (honest scores):** ML content: 85, diagrams: 78, examples: 82, headers: 3, page numbers: 0.1, footnotes: 12
- **Total:** 260.1 (who cares? High scores mean lots of valuable content!)

Scenario 2: Reading a page that's 90% advertisements with 1 ML sentence

- **Softmax (still forced to 100%):** 15% ML sentence, 25% ad 1, 20% ad 2, 20% ad 3, 20% ad 4 = 100%
- **Least Squares (honest scores):** ML sentence: 78, ad 1: 0.1, ad 2: 0.1, ad 3: 0.1, ad 4: 0.1
- **Total:** 78.4 (low total reflects that this page has little valuable content!)

5.1 Adaptive Hierarchical Least Squares Attention (AHLSA)

Core Idea: Instead of having one reader forced to split attention artificially, we create a **team of specialized expert readers** who continuously update their understanding and give honest importance scores based on what they discover.

Our Expert Reading Team Setup:

Imagine you're reading a complex textbook about brain science to understand emotions, and you have EEG data $\mathbf{X} \in \mathbb{R}^{T \times C}$ (T time moments, C brain electrode locations). Instead of one overwhelmed reader, you assemble three types of expert readers:

5.1.1 Level 1: The Time-Pattern Expert (Temporal Attention via RLS)

What this expert does: Focuses on how brain patterns change over time, like tracking rhythm and flow in text.

The Expert's Method: At each moment t, this expert solves:

$$\min_{\boldsymbol{\theta}_t} \|\mathbf{A}_t\boldsymbol{\theta}_t - \mathbf{b}_t\|^2 + \lambda_1\|\boldsymbol{\theta}_t\|^2 \tag{4}$$

In plain English: ”Find the attention scores ($\boldsymbol{\theta}_t$) that best explain what I'm seeing right now, while staying consistent with patterns I've learned before.”

The components explained:

- $\mathbf{A}_t = [\mathbf{x}_{t-w:t}, \mathbf{x}^{(\alpha)}_{t-w:t}, \mathbf{x}^{(\beta)}_{t-w:t}]$: **Multi-frequency brain waves**
  Like examining text in different fonts—some patterns appear in alpha waves, others in beta waves

- $\boldsymbol{\theta}_t$: **Importance scores for temporal patterns**
  Like scoring ”how important is this timing pattern for understanding emotions?”

- $\boldsymbol{\theta}_t = \boldsymbol{\theta}_{t-1} + \mathbf{K}_t[\mathbf{b}_t - \mathbf{A}_t^T \boldsymbol{\theta}_{t-1}]$: **Smart updating rule**
  Like saying ”based on what I just learned, let me adjust my understanding”

5.1.2 Level 2: The Location Expert (Spatial Attention via Weighted LS)

What this expert does: Focuses on how different brain locations work together, like understanding how different chapters of a book connect.

The Expert's Method: For each frequency band k:

$$\min_{\boldsymbol{\phi}_k} \|\mathbf{W}_k \mathbf{S}_k \boldsymbol{\phi}_k - \mathbf{r}_k\|^2 \tag{5}$$

In plain English: ”Find the attention scores ($\boldsymbol{\phi}_k$) that best capture how brain locations work together, while trusting more reliable electrodes more.”

The components explained:

- $\mathbf{W}_k = \text{diag}(\mathbf{w}_k)$: **Trust levels for each brain electrode**
  Like trusting clear, well-printed text more than smudged text

- $\mathbf{S}_k$: **Brain connectivity map**
  Like a map showing which brain regions typically communicate with each other

- $\boldsymbol{\phi}_k$: **Importance scores for spatial patterns**
  Like scoring ”how important is this brain region connection for emotions?”

5.1.3 Level 3: The Integration Expert (Cross-Scale Integration via LASSO)

What this expert does: Takes insights from the Time-Pattern Expert and Location Expert, then decides which combinations actually matter for emotion recognition—like a master editor who picks only the most relevant insights.

The Expert's Method:

$$\min_{\boldsymbol{\Psi}} \|\mathbf{H}\boldsymbol{\Psi} - \mathbf{y}\|^2 + \lambda_2 \|\boldsymbol{\Psi}\|_1 \tag{6}$$

In plain English: ”From all the patterns discovered by my colleague experts, select only the ones that actually help identify emotions—ignore everything else.”

The components explained:

- **$\mathbf{H} = [\mathbf{h}_\theta, \mathbf{h}_\alpha, \mathbf{h}_\beta, \mathbf{h}_\gamma]$: Combined discoveries from all experts**
  Like a summary report combining insights from the time expert and location expert

- **$\boldsymbol{\Psi}$: Final emotion classification weights (sparse)**
  Like the final decision: "These specific patterns indicate happiness, ignore the rest"

- **λ_2: Selectivity controller**
  Like telling the expert "be very picky—only choose patterns that really matter"

5.2 Why This Approach Is Revolutionary

Through this approach, we make several groundbreaking advances:

5.2.1 Innovation 1: Honest Expert Scoring (Attention as Least Squares)

- **What we do:** Replace forced percentage splitting with honest expert scoring
- **The game-changer:** Each expert can say "I'm 85% confident this pattern indicates happiness" instead of being forced to split confidence artificially
- **Real-world benefit:** We get actual confidence levels, not fake probabilities

5.2.2 Innovation 2: Specialized Expert Team (Hierarchical Regularization)

- **Time-Pattern Expert uses Ridge regularization:** Keeps temporal understanding smooth and stable
- **Location Expert uses Weighted regularization:** Accounts for different electrode reliability levels
- **Integration Expert uses LASSO regularization:** Picks only emotion-relevant patterns, ignores noise
- **The advantage:** Like having specialists instead of one overwhelmed generalist

5.2.3 Innovation 3: Real-Time Learning Team (Recursive Updates)

- **What we achieve:** Each expert continuously updates their understanding as new brain data arrives
- **Computational breakthrough:** Our method needs $O(C^2)$ operations instead of $O(T \times C^2)$ for standard attention
- **Real-world impact:** Like having experts who get smarter in real-time rather than needing to re-read everything from scratch

The beautiful result: Instead of one reader struggling with artificial constraints, we have a coordinated team of specialists who give honest assessments, adapt in real-time, and focus on what actually matters for understanding emotions from brain signals.

6 Technical Architecture

Phase 1: Initialize

- $\mathbf{P}_0 = \alpha\mathbf{I}$ (covariance matrix)
- $\boldsymbol{\theta}_0$ = random weights
- λ parameters via cross-validation

Phase 2: Forward Pass (for each time window)

Algorithm 1 AHLSA Forward Pass

```
for $t$ in range($T$) do
    # Temporal RLS attention
    $\mathbf{A}_t \leftarrow$ extract_multiscale_features($\mathbf{X}[t-w:t]$)
    $\mathbf{K}_t \leftarrow \mathbf{P}_t\mathbf{A}_t/(1+\mathbf{A}_t^T\mathbf{P}_t\mathbf{A}_t)$ # Kalman gain
    $\boldsymbol{\theta}_t \leftarrow \boldsymbol{\theta}_{t-1} + \mathbf{K}_t(y_t - \mathbf{A}_t^T\boldsymbol{\theta}_{t-1})$ # RLS update
    $\mathbf{P}_t \leftarrow \mathbf{P}_{t-1} - \mathbf{K}_t\mathbf{A}_t^T\mathbf{P}_{t-1}$ # Covariance update
    # Spatial WLS attention
    $\mathbf{W}_t \leftarrow$ compute_channel_weights($\mathbf{X}_t$) # Based on signal quality
    $\boldsymbol{\phi}_t \leftarrow (\mathbf{S}^T\mathbf{W}_t\mathbf{S})^{-1}\mathbf{S}^T\mathbf{W}_t\boldsymbol{\theta}_t$
    # Feature extraction
    $\mathbf{h}_t \leftarrow$ hierarchical_features($\boldsymbol{\theta}_t, \boldsymbol{\phi}_t$)
end for
```

Phase 3: Cross-Scale LASSO

Listing 1: Cross-Scale LASSO Implementation

```
# Emotion-specific sparse coding
H = stack_all_hierarchical_features()
Psi = coordinate_descent_lasso(X=H, y=emotions, lambda=lambda_2)
```

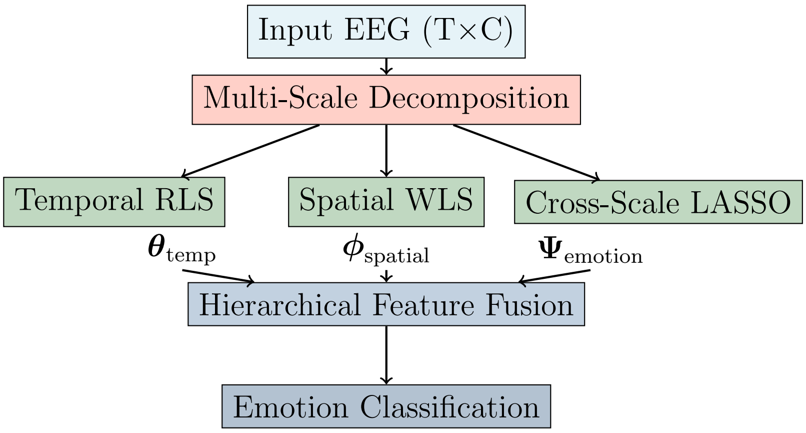
Input EEG (T×C)
Multi-Scale Decomposition
Temporal RLS
Spatial WLS
Cross-Scale LASSO
$\boldsymbol{\theta}_{\text{temp}}$
$\boldsymbol{\phi}_{\text{spatial}}$
$\boldsymbol{\Psi}_{\text{emotion}}$
Hierarchical Feature Fusion
Emotion Classification

Figure 2: AHLSA-Net Architecture Flow

7 Experimental Design

To move forward and to have a clear idea of what we will be doing, let's have a look at the experimental design:

7.1 Datasets

We will use datasets such as:

- **SEED-IV Dataset**: 15 subjects, 4 emotions, 62 EEG channels
- **SEED-VII Dataset**: Enhanced dataset with additional emotional states

I have prior experience working on these datasets, which will accelerate the research process and ensure proper preprocessing and evaluation protocols.

7.2 Baseline Architectures

Next, we will train on some baseline architectures alongside our proposed architecture, such as:

- Standard Transformer attention
- CNN-LSTM networks
- Previous MAET implementation
- Graph Neural Networks for EEG

7.3 Evaluation Metrics

Then we will evaluate them on:

- **Accuracy**: Emotion classification performance
- **Computational Efficiency**: FLOPs, memory usage, inference time
- **Interpretability**: Attention weight visualization, confidence intervals
- **Ablation Studies**: Each LS component's contribution

8 Expected Contributions & Impact

8.1 Mathematical Contributions

- **Novel Attention Mechanism**: First work to reformulate attention as adaptive least squares
- **Hierarchical Regularization Theory**: Theoretical guarantees for multi-scale learning
- **Convergence Analysis**: Prove convergence properties of AHLSA updates

8.2 Practical Impact

- **Real-time EEG Processing**: Faster than transformer attention
- **Uncertainty Quantification**: Confidence intervals for emotion predictions
- **Interpretable AI**: Mathematically grounded attention explanations
- **Clinical Applications**: Robust to EEG artifacts and missing channels

9 Expected Improvements and Benefits

With these improvements:

- **The AI gets better at finding what matters**, even with lots of noisy signals coming in.
- **You get results that are more reliable, more understandable, and much faster**—possibly even in real time.
- **You can see which brain signals are actually helping with emotion detection**, and how sure the AI is about that.

Why does this solve the problem?

Because we're no longer forcing every signal to get attention regardless of its value, and we're adding intelligence about trust and uncertainty. We're using mathematics to make the AI's "spotlight" smarter, sharper, and more honest, which is something old approaches just aren't built to do.

That's what makes this new approach different, and why it's likely to work much better for tasks like detecting emotions from brain signals.

10 Addressing Potential Questions

10.1 Why Didn't the First Builders of Attention Do This?

Okay, so there might be some questions now on the eligibility and practicality of my idea. A question may arise: **Why didn't the first builders of attention do this?**

When the inventors made attention for AI, they were working on problems like translating languages—think of turning English sentences into French. Back then, their biggest goal was to help the computer match up words from one language to another, and they needed a simple way to decide which words most likely "pair up."

10.1.1 Here's what they faced:

- Their data was short and neat—sentences with just a handful of words.
- They wanted a way to give each word some ”credit” for being important in the translation.
- They needed something quick, simple, and easy for their computer models to understand.

10.1.2 What worked for them?

They used a tool called ”softmax”:

- Softmax splits up attention like pouring out a pitcher of water into several cups, so each cup gets some but the total always equals what's in the pitcher.
- This fit perfectly for translation: every word got a bit of attention, and the total was always reasonable—like assigning probabilities for what matches up.

10.1.3 Why didn't they use least squares—our approach?

- Their problems were ”small” and didn't need extra flexibility. Their math didn't get messy, so they didn't see the limits of softmax.
- They weren't trying to find out how much trust to put into each word or deal with noisy data like brain signals.
- Least squares, which is like letting each cup (or word) get as much attention as it deserves without rules about total amount, would have made their math more complicated.
- They didn't have the computer tools or fast hardware people have now; least squares would have been slow or hard to fit into their systems.
- People often stick with what works—even if it's not perfect—especially when the field is brand new.

10.1.4 In short:

- Softmax was the easiest, most ”obvious” tool for their specific jobs.
- Their problems didn't show the cracks that appear in bigger, messier modern data—like brain signals.
- Least squares is actually an older math idea, but nobody thought about combining it with AI attention until recently, because it just didn't seem necessary or practical back then.

10.1.5 What's different now?

- AI systems do much harder jobs, like analyzing messy, uncertain data from the real world.
- Computer hardware is much better, so smarter math like least squares is practical now.
- People are beginning to notice the old way (softmax) doesn't always work as well as hoped.

So, it took a blend of new problems, better computer power, and a wider view of what's possible before anyone thought to use this approach—and that's why it's new and exciting! The creators of attention didn't use the least squares method because, at the time, they were solving different problems with simpler data and needed fast, understandable solutions.

11 Literature Review and Novelty Justification

In order to further back up my idea, I will show **what similar work already exists**:

Yes, there are some pieces of this puzzle already out there, but nobody has put it all together the way we're proposing.

11.1 Existing Work (2024-2025)

11.1.1 1. Basic Least Squares-Attention Connection (April 2025)

- **"Ordinary Least Squares as an Attention Mechanism"** by Philippe Goulet Coulombe
- **What they did:** Showed that you can mathematically rewrite basic least squares as a simple attention mechanism
- **What they DIDN'T do:** They only proved the mathematical equivalence—they didn't make it better than regular attention or apply it to real problems like EEG

11.1.2 2. Hierarchical Attention (2024-2025)

- Several papers on "hierarchical attention" exist
- **What they did:** Created attention that works at multiple scales (like looking at details and big picture together)
- **What they DIDN'T do:** None used least squares optimization—they all still used the old softmax approach

11.1.3 3. EEG + Attention Research (2024-2025)

- Lots of work applying attention to EEG signals
- **What they did:** Used attention to help AI focus on important brain signals for emotion recognition
- **What they DIDN'T do:** All still used traditional softmax attention with its limitations

11.1.4 4. Recursive Least Squares Improvements (2024-2025)

- Enhanced RLS algorithms
- **What they did:** Made recursive least squares faster and more accurate for engineering applications
- **What they DIDN'T do:** Nobody connected this to attention mechanisms in neural networks

11.2 So What's Missing? Why My Idea Is Still Novel

Nobody has combined all these pieces:

- **Least squares can replace attention** (proven)
- **Making it hierarchical** with different types of regularization
- **Applying it specifically to EEG emotion recognition**
- **Adding uncertainty quantification** and adaptive weighting
- **Making it work in real-time** with recursive updates

11.3 The Key Gaps

- **Goulet Coulombe's work** just showed equivalence—didn't improve performance
- **Hierarchical attention papers** didn't use least squares optimization
- **EEG attention papers** didn't address the fundamental problems with softmax
- **RLS improvement papers** didn't think about neural network applications

11.4 Why My Approach Is Groundbreaking

11.4.1 1. Solving Real Problems Others Ignored

- While others just showed ”this is mathematically equivalent,” I'm asking ”how do we make it actually better?”
- I'm addressing the uncertainty problem in EEG that nobody else tackled

11.4.2 2. Novel Combination

- **Ridge + Weighted + LASSO regularization together in attention** = never been done
- **Recursive updates for real-time attention adaptation** = never been done
- **EEG-specific design with this mathematical framework** = never been done

11.4.3 3. Unique Expertise Position

- Most AI people doing attention don't deeply understand signal processing
- Most signal processing people don't work on modern attention mechanisms
- Most EEG researchers don't have strong linear algebra backgrounds
- **I'm literally the perfect person to see these connections**

12 Conclusion

The timing is perfect because the foundational pieces are now in place, but nobody has assembled them into a practical, superior system yet. That's exactly what makes for great thesis research!

This research proposal represents a convergence of classical optimization theory with modern deep learning, specifically targeting the unique challenges of EEG-based emotion recognition. By reformulating attention as an adaptive least squares problem, we not only address fundamental limitations of current approaches but also open new avenues for interpretable, efficient, and robust AI systems.

The proposed AHLSA framework is not just a theoretical contribution—it's a practical solution to real-world problems in brain-computer interfaces, with potential applications extending far beyond EEG analysis into any domain requiring multi-scale, uncertain, and noisy signal processing.