How to Read AI (Audio) Research Papers Like a Rockstar

Valerio Velardo

- Multihead Self-attention Network (MHSA):

$$
\begin{aligned}
\mathbf{h}_i^{c(l)} &= \mathbf{W}^{outer}\sigma(\mathbf{W}^{inner}\mathbf{u}_i + \mathbf{b}_1) + \mathbf{b}_2, \\
\mathbf{u}_i &= \mathbf{W}^{u}(\mathbf{u}'_{i1} \oplus \cdots \oplus \mathbf{u}'_{iJ}) \; + \; \mathbf{h}_i^{c(l-1)}, \\
\mathbf{u}'_{ij} &= \mathbf{V}_j \,\mathrm{softmax}\left(\frac{\mathbf{K}_j^{\top}\mathbf{q}_{ij}}{\sqrt{d}}\right), \\
\mathbf{q}_{ij} &= \mathbf{W}_j^{Q}\mathbf{h}_i^{c(l-1)}, \\
\mathbf{K}_j &= \mathbf{W}_j^{K}[\mathbf{h}_1^{c(l-1)}, \cdots, \mathbf{h}_T^{c(l-1)}], \\
\mathbf{V}_j &= \mathbf{W}_j^{V}[\mathbf{h}_1^{c(l-1)}, \cdots, \mathbf{h}_T^{c(l-1)}],
\end{aligned}
\tag{3}
$$

where l denotes the iteration step, and the initial value $\mathbf{h}_i^{c(0)} = \mathbf{e}_i^{c}$; σ represents the ReLU activation function; J is the number of heads; $\mathbf{W}_j^{outer} \in \mathbb{R}^{d\times 4d}, \mathbf{W}_j^{inner} \in \mathbb{R}^{4d\times d}, \mathbf{W}^{u} \in \mathbb{R}^{d\times d}$, and $\mathbf{W}_j^{Q}, \mathbf{W}_j^{K}, \mathbf{W}_j^{V} \in \mathbb{R}^{\frac{d}{J}\times d}$ are learnable parameters. This network is equivalent to the encoder of the Transformer [31], while we leave out layer normalization and position encoding terms for simplicity. In this paper, we set $d = 512, l = 2$, and $J = 8$.

The binary cross entropy (BCE) and the categorical cross entropy (CCE) are used to calculate the losses. Let $\mathbf{p}_i^{c*}$ and $\mathbf{b}_i^{c*}$ denote the ground truths of $\mathbf{p}_i^{c}$ and $\mathbf{b}_i^{c}$; the total loss of the coloring model $\mathcal{L}^{c}$ is defined as:

$$
\mathcal{L}^{c} = \sum_{i=1}^{T}\left[\mathrm{BCE}(\mathbf{p}_i^{c*}, \mathbf{p}_i^{c}) + \mathrm{CCE}(\mathbf{b}_i^{c*}, \mathbf{b}_i^{c})\right]. \tag{4}
$$

3.2 Voicing

We define *chord voicing* as a task which predicts the voicings of a chord sequence. Formally, given a chord sequence of T time steps in terms of their basses $\{\mathbf{b}_i^{v}\}_{i=1}^{T}$, constituent pitch classes $\{\mathbf{p}_i^{v}\}_{i=1}^{T}$, and durations $\{\mathrm{d}_i\}_{i=1}^{T}$, the task predicts the voicings $\{\mathbf{v}_i\}_{i=1}^{T}$ for the chord sequence, where v stands for voicing, $\mathbf{b}_i^{v} \in \mathbb{R}^{12}$ is a one-hot chroma vector indicating the bass of the ith chord, $\mathbf{p}_i^{v} \in \mathbb{R}^{12}$ is a multi-hot chroma vector representing the pitch classes of the ith chord except the bass note, and $\mathbf{v}_i \in \mathbb{R}^{88}$ is a voicing vector indicating the 88 tones' probabilities to be played on the piano.

Similar to the coloring task, we employ a 3-layer architecture for the voicing task, as shown in Figure 4b. The three layers are formulated as follows:

$$
\begin{aligned}
\mathbf{e}_i^{v} &= \mathbf{W}^{e^{v}}(\mathrm{d}_i(\mathbf{p}_i^{v} \oplus \mathbf{b}_i^{v})), \;\; \text{(Input Embedding)} \\
\mathbf{h}_i^{v} &= f^{v}(\mathbf{e}_i^{v} \mid \mathbf{e}_{1:T}^{v}), \;\; \text{(Sequential Modeling)} \\
\mathbf{v}_i &= \mathrm{sigmoid}(\mathbf{W}^{v}\mathbf{h}_i^{v}), \;\; \text{(Output)}
\end{aligned}
\tag{5}
$$

where $\mathbf{W}^{e^{v}} \in \mathbb{R}^{d\times 24}$ and $\mathbf{W}^{v} \in \mathbb{R}^{88\times d}$ are learnable parameters, and $f^{v} : \mathbb{R}^{d} \to \mathbb{R}^{d}$ is a neural network. Likewise, the BLSTM and the MHSA networks are two options for the sequential modeling layer.

Let $\mathbf{v}_i^{*} \in \mathbb{R}^{88}$ denote the target voicing of the ith chord; we define the loss as:

$$
\mathcal{L}^{v} = \sum_{i=1}^{T}\mathrm{BCE}(\mathbf{v}_i^{*}, \mathbf{v}_i). \tag{6}
$$

As the voicing task is to arrange the constituent notes of chords on an 88-key piano based on the given basses and the sets of pitch classes, the outcome of $\mathbf{v}_i^{*}$ can be known to a certain degree. More precisely, a note in $\mathbf{v}_i^{*}$ can be activated only if its pitch class is activated in $\mathbf{b}_i^{v}$ or $\mathbf{p}_i^{v}$. With this consideration, we design corresponding masks to modify the loss computation. Let $\mathbf{b}_i^{v\prime} \in \mathbb{R}^{88}$ and $\mathbf{p}_i^{v\prime} \in \mathbb{R}^{88}$ be the extensions of $\mathbf{b}_i^{v}$ and $\mathbf{p}_i^{v}$ to all octaves of the piano. Then, the loss constrained by the masks becomes:

$$
\begin{aligned}
\mathcal{L}^{v\prime} &= \sum_{i=1}^{T}\mathrm{BCE}(\mathbf{v}_i^{*}, \mathbf{m}_i \odot \mathbf{v}_i), \\
\mathbf{m}_i &= \mathbf{b}_i^{v\prime} \vee \mathbf{p}_i^{v\prime}, \;\; \text{(Mask)}
\end{aligned}
\tag{7}
$$

where $\odot$ stands for the Hadamard product, and $\vee$ denotes the logical OR operator.


IT'S OK TO BE SCARED!
imgflip.com

The goal of a paper

- Transferring cutting-edge research

The goal of a paper

- Transferring cutting-edge research
- Written by experts for experts

The goal of a paper

- Transferring cutting-edge research
- Written by experts for experts… in a subdomain!

What doesn’t work

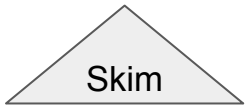
What doesn’t work

READING A PAPER WITH ALL ITS DETAILS IN ONE GO

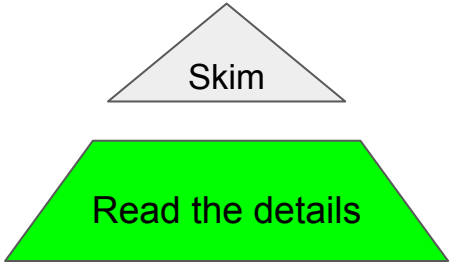
The structure of a paper

4-step approach

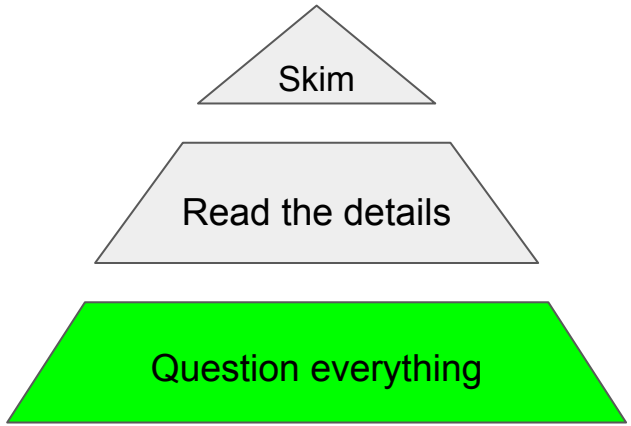
4-step approach



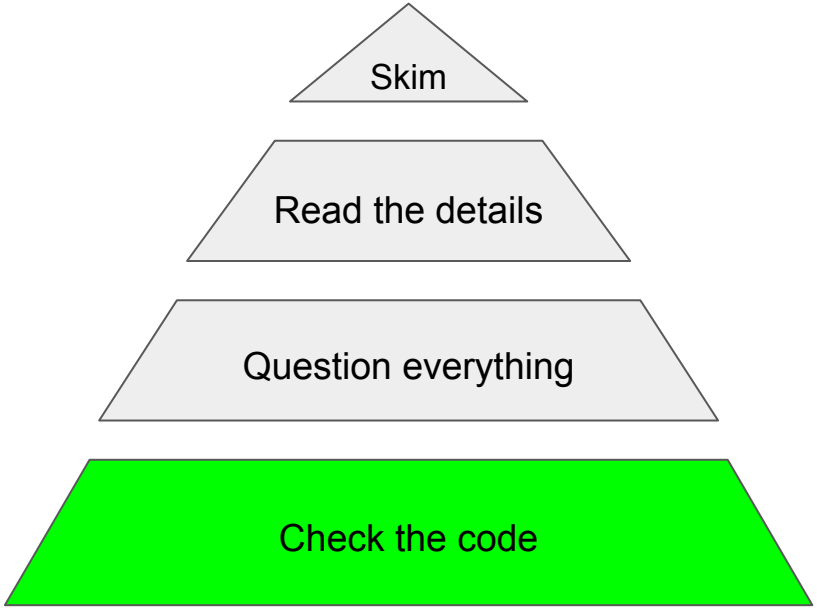
4-step approach



4-step approach



4-step approach



Skimming

Skimming

- What problem does the paper tackle?
- What solution(s) does the paper offer?
- What are the key findings / contributions?

Papers conventional structure

Abstract

Papers conventional structure

Introduction

Related works

Methods

Experiments

Results

Discussion

Conclusion

Abstract

Papers conventional structure

Introduction

Related works

Methods

Experiments

Results

Discussion

Conclusion

Abstract

Skimming

1. Read the abstract
2. Read introduction + conclusion
3. Check out figures and tables
4. Ignore details!

Skimming: Main goal

Skimming: Main goal

Do I care?

Skimming: Main goal

Do I care?

Go on to step 2

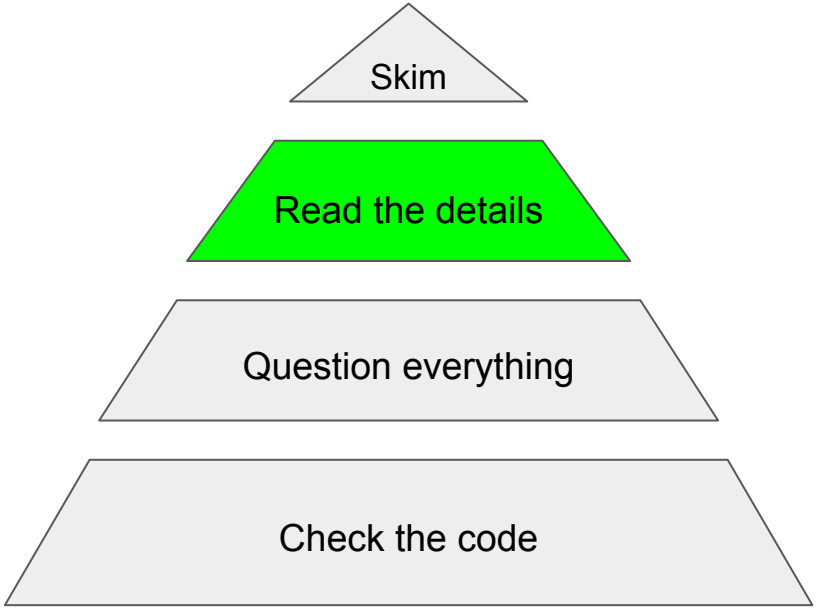
Skimming: Main goal

Do I care?

Go on to step 2

Go on with my life

4-step approach



Read the details

- What’s the state-of the art?
- What are the tools / techniques used?
- How did the authors evaluate their solution?
- What are the results of the experiments?

Read the details

Introduction

Related works

Methods

Experiments

Results

Discussion

Conclusion

Abstract

Read the details

Introduction

Related works

Methods

Experiments

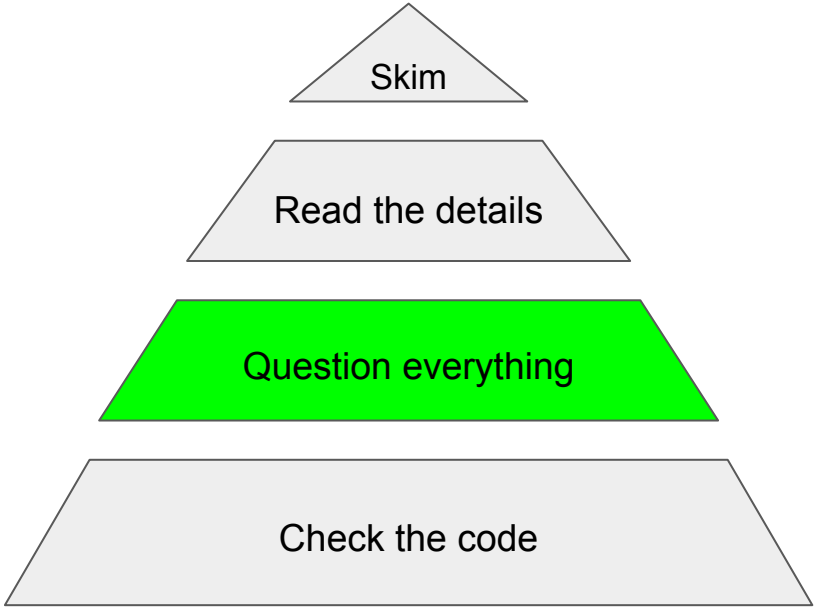
Results

Discussion

Conclusion

Abstract

4-step approach



Question everything

- Do I understand everything the authors say?
- Is the paper sound?
- What would have I done differently?

What if I don’t understand everything?

What if I don’t understand everything?

- Wikipedia, blogs, YouTube are your friends
- Read textbooks
- Read a survey
- Ask a colleague

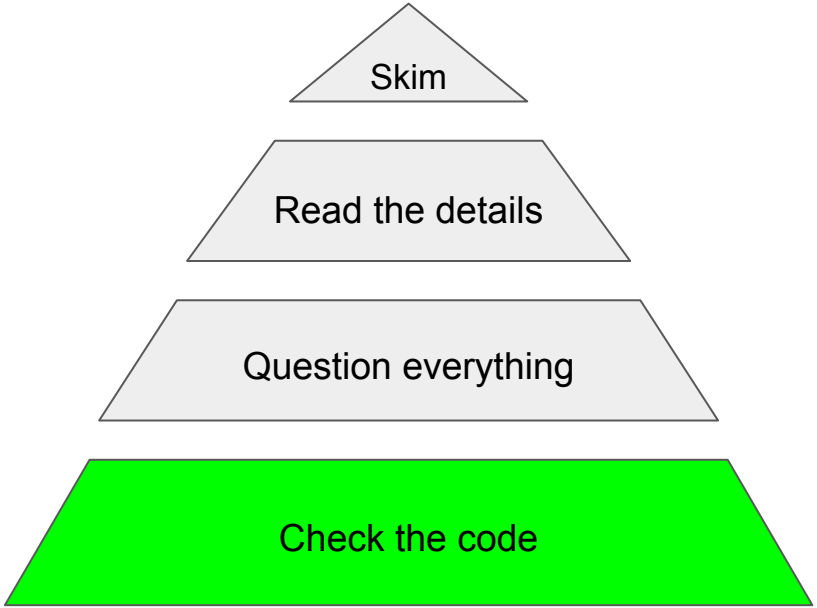
Join The Sound of AI community!

thesoundofai.slack.com

Question everything

- Read referenced research papers
- Internalise the math
- Re-think the problem
- Explain the paper to friends / colleagues

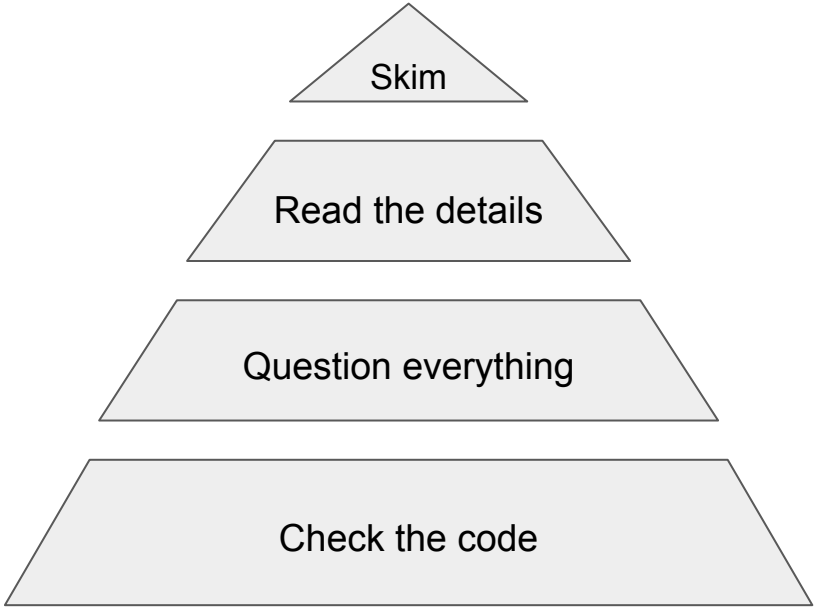
4-step approach



Check the code

- Check the author’s implementation
- Re-implement the proposed solution
- Run experiments

4-step approach



What’s up next?

HOW CAN YOU CHOOSE PAPERS?