How to draft the perfect invention disclosure report for an AI invention

BARDEHLE PAGENBERG

Bastian Best

Dipl.-Inf., German & European Patent Attorney, Partner

January 28, 2019


WHY IS THIS IMPORTANT?

#1: CONFLICTING INTERESTS

OWNER
WANTS BROAD PROTECTION

PUBLIC
WANTS TO KNOW HOW IT WORKS

Photo by rawpixel on Unsplash

#2: THE RACE TO THE PATENT OFFICE

Photo by Victoire Joncheray on Unsplash

#3: POINT OF NO RETURN

NO CHANGES ALLOWED AFTER FILING

Photo by Sydny Riters on Unsplash

PATENTS

- HAVE TO BE FILED QUICKLY
- CANNOT BE EXPANDED LATER

THE ANATOMY OF A PATENT APPLICATION

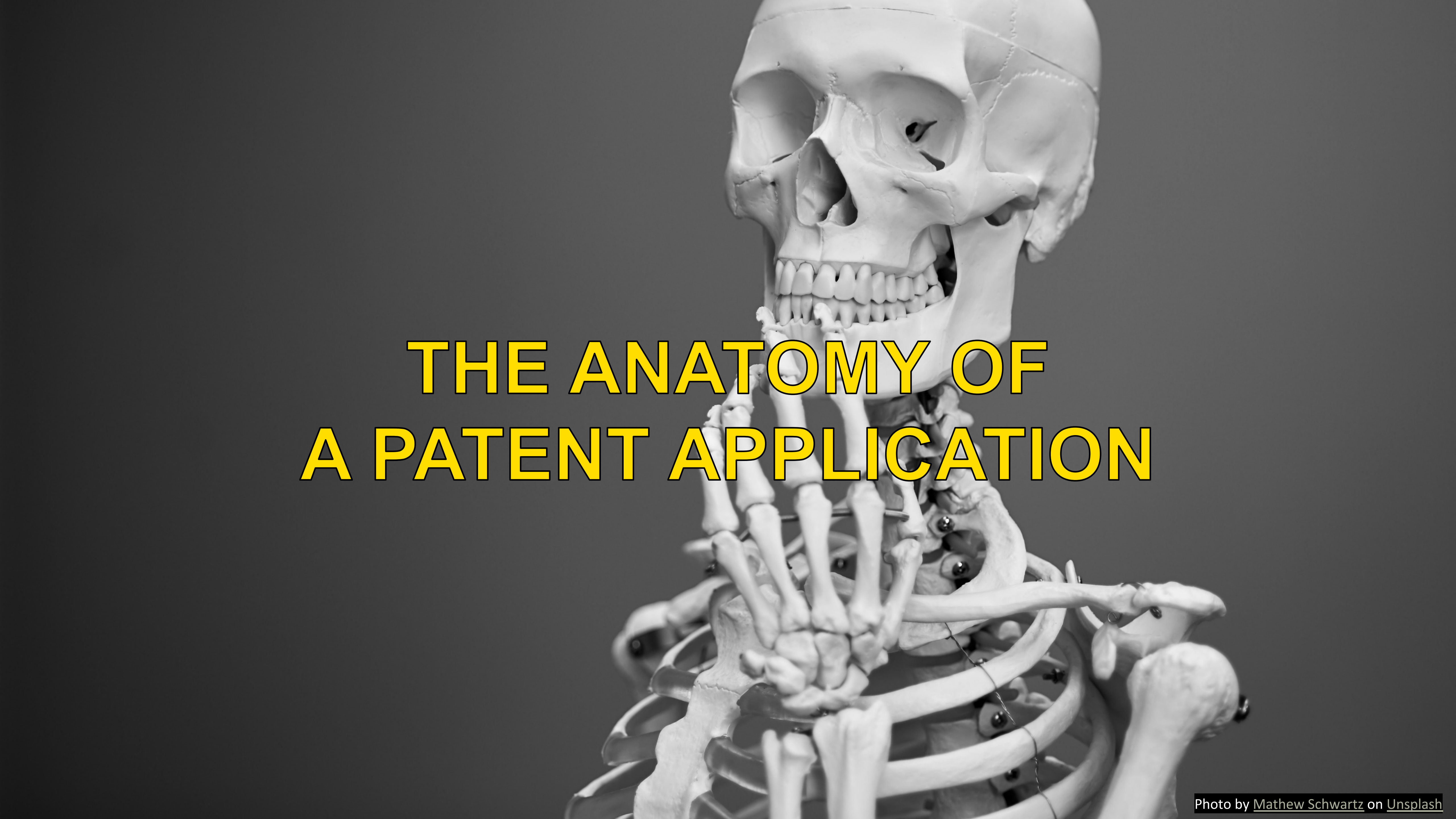


Photo by Mathew Schwartz on Unsplash

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) **World Intellectual Property Organization**
International Bureau

(43) **International Publication Date**
15 March 2018 (15.03.2018)

WIPO | PCT

(10) **International Publication Number**
WO 2018/048934 A1

(51) **International Patent Classification:**
G06N 3/04 (2006.01)

(21) **International Application Number:**
PCT/US2017/050320

(22) **International Filing Date:**
06 September 2017 (06.09.2017)

(25) **Filing Language:** English

(26) **Publication Language:** English

(30) **Priority Data:**
62/384,115 06 September 2016 (06.09.2016) US

(71) **Applicant: DEEPMIND TECHNOLOGIES LIMITED** [GB/GB]; 5 New Street Square, London EC4A 3TW (GB).

(72) **Inventors: VAN DEN OORD, Aaron Gerard Antonius**; 7 Pancras Square Kings Cross, London N1C 4AG (GB). **DIELEMAN, Sander Etienne Lea**; 7 Pancras Square, Kings Cross, London N1C 4AG (GB). **KALCHBRENNER, Nal Emmerich**; 7 Pancras Square, Kings Cross, London N1C 4AG (GB). **SIMONYAN, Karen**; 6 Pancras Square, London N1C 4AG (GB). **VINYALS, Oriol**; 6 Pancras Square, London N1C 4AG (GB).

(74) **Agent: PORTNOV, Michael** et al.; FISH & RICHARDSON P.C., P.O. BOX 1022, 3300 RBC PLAZA, MINNEAPOLIS, Minnesota 55440-1022 (US).

(81) **Designated States** *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

(54) **Title:** GENERATING AUDIO USING NEURAL NETWORKS

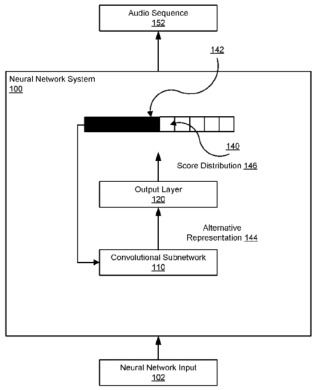


FIG. 1

(57) **Abstract:** Methods, systems, and apparatus, including computer programs encoded on computer storage media, for generating an output sequence of audio data that comprises a respective audio sample at each of a plurality of time steps. One of the methods includes, for each of the time steps: providing a current sequence of audio data as input to a convolutional subnetwork, wherein the current sequence comprises the respective audio sample at each time step that precedes the time step in the output sequence, and wherein the convolutional subnetwork is configured to process the current sequence of audio data to generate an alternative representation for the time step; and providing the alternative representation for the time step as input to an output layer, wherein the output layer is configured to: process the alternative representation to generate an output that defines a score distribution over a plurality of possible audio samples for the time step.

WO 2018/048934 A1

[Continued on next page]

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
15 March 2018 (15.03.2018)

(10) International Publication Number
WO 2018/048934 A1

(51) International Patent Classification:
G06N 3/04 (2006.01)

(21) International Application Number:
PCT/US2017/050320

(22) International Filing Date:
06 September 2017 (06.09.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/384,115 06 September 2016 (06.09.2016) US

(71) Applicant: **DEEPMIND TECHNOLOGIES LIMITED** [GB/GB]; 5 New Street Square, London EC4A 3TW (GB).

(72) Inventors: **VAN DEN OORD, Aaron Gerard Antonius**; 7 Pancras Square Kings Cross, London N1C 4AG (GB). **DIELEMAN, Sander Etienne Lea**; 7 Pancras Square, Kings Cross, London N1C 4AG (GB). **KALCHBRENNER, Nal Emmerich**; 7 Pancras Square, Kings Cross, London N1C 4AG (GB). **SIMONYAN, Karen**; 6 Pancras

(74) Agent: **PORTNOV, Michael** et al.; FISH & RICHARDSON P.C., P.O. BOX 1022, 3300 RBC PLAZA, MINNEAPOLIS, Minnesota 55440-1022 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,

(54) **Title:** GENERATING AUDIO USING NEURAL NETWORKS

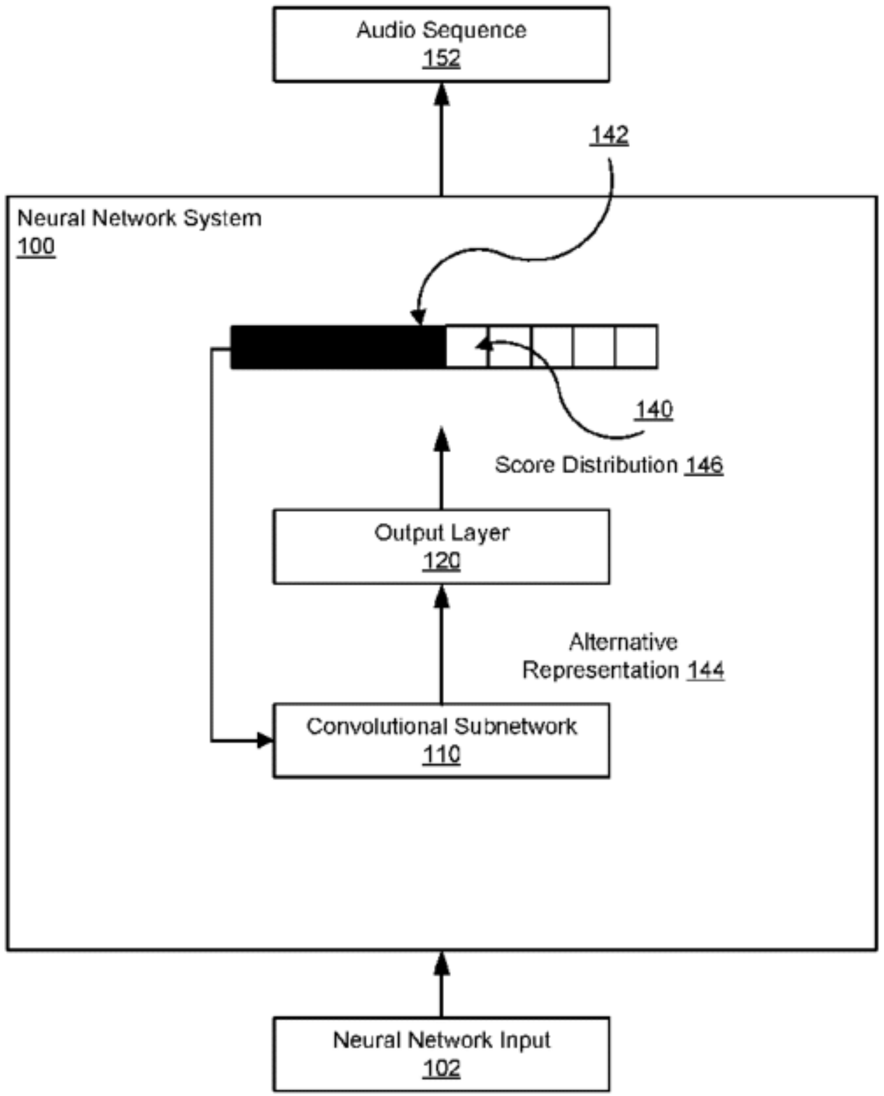


FIG. 1

(57) **Abstract:** Methods, systems, and apparatus, including computer programs encoded on computer storage media, for generating an output sequence of audio data that comprises a respective audio sample at each of a plurality of time steps. One of the methods includes, for each of the time steps: providing a current sequence of audio data as input to a convolutional subnetwork, wherein the current sequence comprises the respective audio sample at each time step that precedes the time step in the output sequence, and wherein the convolutional subnetwork is configured to process the current sequence of audio data to generate an alternative representation for the time step; and providing the alternative representation for the time step as input to an output layer, wherein the output layer is configured to: process the alternative representation to generate an output that defines a score distribution over a plurality of possible audio samples for the time step.

18/048934 A1

GENERATING AUDIO USING NEURAL NETWORKS

BACKGROUND

This specification relates to processing and generating audio using neural networks.

Neural networks are machine learning models that employ one or more layers of nonlinear units to predict an output for a received input. Some neural networks include one or more hidden layers in addition to an output layer. The output of each hidden layer is used as input to the next layer in the network, i.e., the next hidden layer or the output layer. Each layer of the network generates an output from a received input in accordance with current values of a respective set of parameters.

SUMMARY

This specification describes how a system implemented as computer programs on one or more computers in one or more locations can generate a sequence of audio data that includes a respective audio sample at each of multiple time steps. For example, the sequence of audio data can represent speech in a particular natural language or a piece of music.

In one innovative aspect a neural network system implemented by one or more computers is configured to generate an output sequence of audio data that comprises a respective audio sample at each of a plurality of time steps. The neural network system may comprise a convolutional subnetwork comprising one or more audio-processing convolutional neural network layers; and an output layer. The convolutional subnetwork may be configured to, for each of the plurality of time steps: receive a current sequence of audio data that comprises the respective audio sample at each time step that precedes the (current) time step in the output sequence. The convolutional subnetwork may further be configured to process the current sequence of audio data to generate an alternative representation for the time (current) step. This alternative representation may thus comprise a numeric representation, i.e., an ordered collection of numeric values, in which the current sequence of audio data has been encoded by the convolutional subnetwork, for example encoding features of the current sequence. The output layer may be configured to, for each of the plurality of time steps: receive the alternative representation for the time step, and process the alternative representation for the time step to generate an

output that defines a score distribution over a plurality of possible audio samples for the time step.

Some of the many advantages of such a system are described later. The system can use the score distribution to select a sample for the current time step, by sampling from the distribution. The output may, but need not necessarily, comprise one score for each possible audio sample value, for example 256 scores for 256 possible values. In can thus be useful to compress or compand the audio sample values, which may be amplitude values, to reduce the number of model outputs.

In some implementations the convolutional neural network layers are causal convolutional neural network layers, as described in more detail later. In particular, the audio-processing convolutional neural network layers may include one or more dilated causal convolutional neural network layers. Again as described in more detail later, a dilated convolutional neural network layer applies a convolution to non-adjacent values in a sequence, i.e., as defined by the outputs from a previous layer. This can increase the receptive field of the convolutional subnetwork by orders of magnitude whilst preserving the input (time) resolution and maintaining computational efficiency.

In some implementations the convolutional neural network layers include multiple stacked blocks of dilated convolutional neural network layers. Each block may comprise multiple dilated convolutional neural network layers with increasing dilation. For example the dilation may be increased by a factor n for each successive layer up to a limit within each block. This can further increase the receptive field size.

In some implementations one or more of the convolutional neural network layers may have gated activation units. For example a rectified linear or other unit following a convolution implemented by a layer may be replaced by a gated activation unit. In a gated activation unit the output may be a combination of two (causal) convolutions, a main convolution and a gate convolution. The convolutions may each be applied to some or all of the same outputs from the previous layer. The combination may involve a non-linear activation function applied to the gate convolution, for example an activation with a (0,1) range such as a sigmoid. This may then multiply a value from the main convolution; a non-linear activation function may, but need not be, applied to the main convolution. Such an approach may assist in capturing more complex structure within the data.

The alternative representation from the convolutional subnetwork at each time step may be conditioned on a neural network input, for example a latent representation of

a conditioning input. The conditioning input may be global (substantially time-independent) and/or local (time-dependent). The conditioning input may comprise, for example, text, image or video data, or audio data, for example an example of a particular speaker or language or music. The neural network input may comprise an embedding of the conditioning input. For example in a text-to-speech system a global conditioning input may comprise a speaker embedding and a local conditioning input may comprise linguistic features. The system may be configured to map the neural network input, or a conditioning input, from a lower sampling frequency to the audio sample generation frequency, for example by repeating the input or upsampling the input using a neural network. Thus the neural network input may comprise features of a text segment and the output sequence may represent a verbalization of the text segment; and/or the neural network input may comprise speaker or intonation pattern values; and/or the neural network input may include one or more of: speaker identity information, language identity information, and speaking style information. Alternatively the output sequence represents a piece of music.

The convolutional subnetwork may comprise residual connections, for example a connection from an input of a convolutional layer to a summer to sum this with an intermediate output of the layer. This effectively allows the network to be trained to skip or partially skip a layer, thus speeding up convergence and facilitating training of deeper models. The convolutional subnetwork may additionally or alternatively comprise skip connections, for example directly from each of one or more intermediate layers of the convolutional subnetwork to one or more operations that directly generate the alternative representation that is provided to the output layer.

In some implementations processing the current sequence of audio data using the convolutional subnetwork, to generate an alternative representation for the time step, re-uses values computed for previous time steps. The re-used values may comprise values derived from application of a convolutional filter to the audio sample data or data derived therefrom. The re-used values may be stored at one time step and retrieved at a later time step when the same filter is applied to the same (or some of the same) audio sample data or data derived therefrom. This can make the system computationally more efficient and hence faster, because there is no need to re-compute the stored values.

Particular embodiments of the subject matter described in this specification can be implemented so as to realize one or more of the following advantages. The neural network system can generate on the order of tens of thousands of audio samples per

$$\mathbf{z} = \tanh(W_{f,k} * \mathbf{x} + V_{f,k} * \mathbf{y}) \odot \sigma(W_{g,k} * \mathbf{x} + V_{g,k} * \mathbf{y}),$$

where $V_{f,k} * \mathbf{y}$ and $V_{g,k} * \mathbf{y}$ are respective 1 x 1 convolutions. In some implementations, the system 100 directly receives the sequence **y** as the neural network input, i.e., directly receives a sequence that has the same resolution as the output sequence. In other implementations, the system 100 receives a sequence having a lower resolution, i.e., with a lower sampling frequency, than the output sequence. In these cases, the system can generate the sequence **y** by processing the lower resolution sequence using a transposed (learned upsampling) convolutional network to generate the sequence **y** or can repeat values from the lower resolution sequence across time to generate the sequence **y**.

As an example, when the local features are linguistic features for use in text to speech generation, the linguistic features can include some or all of phone, syllable, word, phrase, and utterance-level features of the text. Example sets of linguistic features that can be used are described in Zen, Heiga. *An example of context-dependent label format for HMM-based speech synthesis in English*, 2006. URL http://hts.sp.nitech.ac.jp/?Download and Zen, Heiga, Senior, Andrew, and Schuster, Mike. *Statistical parametric speech synthesis using deep neural networks*. In Proc. ICASSP, pp. 7962–7966, 2013.

Because the architecture 300 includes skip connections and residual connections for the dilated causal convolutional layers, the layer 304 then performs a 1 x 1 convolution 316 on the activation function output.

The layer 304 provides the output of the 1 x 1 convolution as the skip output 318 of the layer and adds the residual, i.e., the layer input 306, and the output of the 1 x 1 convolution to generate the final output 320 of the layer 304. The convolutional subnetwork 110 then provides the final output 320 as the layer input to the next dilated convolutional layer in the stack.

In some implementations, the layer 304 performs two 1 x 1 convolutions on the activation function output, one with a residual filter and the other with a skip filter. In these implementations, the layer 304 provides the output of the convolution with the skip filter as the skip output 318 of the layer and adds the residual and the output of the 1 x 1 convolution with the residual filter to generate the final output 320 of the layer 304.

The convolutional subnetwork 110 then provides the final output 320 as the layer input to the next dilated convolutional layer in the stack. For the last layer in the stack, because there is no next layer, the convolutional subnetwork 110 can either discard the

final output 320 generated by the last layer or can refrain from computing a final output, i.e., can refrain from performing the 1 x 1 convolution and the residual sum for the last layer in the stack.

Once the processing of all of the layers 304 in the stack of dilated convolutional layers has been completed, the convolutional subnetwork 110 sums 322 the skip outputs generated by the layers 304. The convolutional subnetwork 110 can then apply one or more non-linear functions, one or more 1 x 1 convolutions, or both to the sum 322 to generate the alternative representation 144. In particular, in the example of FIG. 3, the convolutional subnetwork 110 applies an element-wise non-linearity 324, e.g., a ReLU, followed by a 1x1 convolution 326, followed by another element-wise non-linearity 328, and followed by a final 1x1 convolution 330, to generate the alternative representation 144.

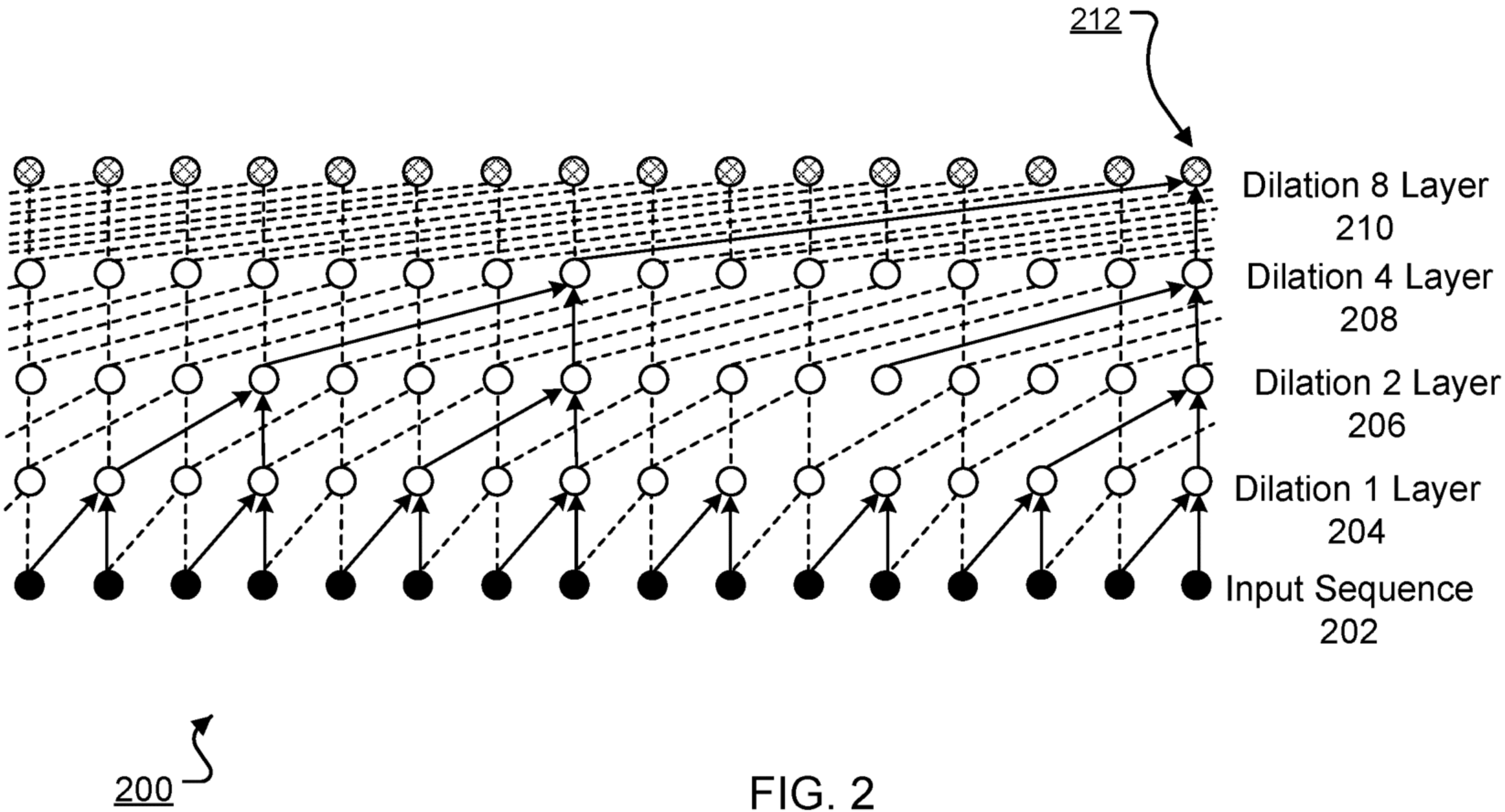
As described above, the output layer 120 then processes the alternative representation 144 to generate the score distribution 146.

FIG. 4 is a flow diagram of an example process 400 for generating an audio sample at a given time step in an audio sequence. For convenience, the process 400 will be described as being performed by a system of one or more computers located in one or more locations. For example, a neural network system, e.g., the neural network system 100 of FIG.1, appropriately programmed, can perform the process 400.

The system provides a current audio sequence as input to the convolutional subnetwork (step 402). The current audio sequence is the audio sequence that has already been generated as of the given time step, i.e., a sequence that includes the output audio samples at time steps before the given time step. As described above, the convolutional subnetwork includes audio-processing convolutional neural network layers, e.g., dilated causal convolutional layers, and is configured to process the current sequence of audio data to generate an alternative representation for the given time step.

The system provides the alternative representation as input to an output layer, e.g., a softmax output layer (step 404). The output layer is configured to process the alternative representation to generate a score distribution over possible audio samples for the time step.

The system selects an audio sample for inclusion in the audio sequence at the given time step in accordance with the score distribution (step 406). For example, the system can sample a possible audio sample in accordance with the score distribution.



212
Dilation 8 Layer
210
Dilation 4 Layer
208
Dilation 2 Layer
206
Dilation 1 Layer
204
Input Sequence
202
200

FIG. 2

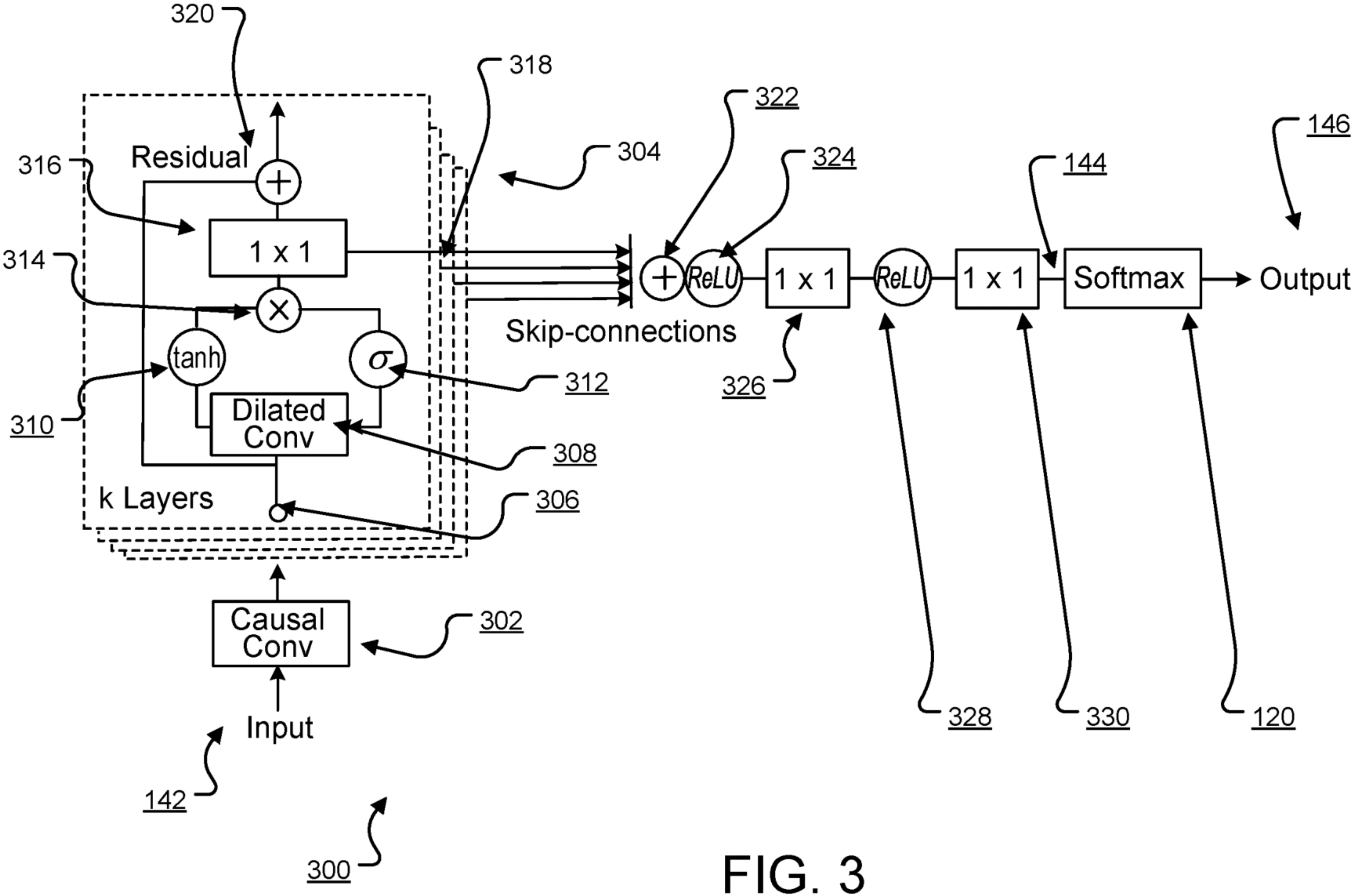

320
318
322
146
304
324
144
316
Residual
1 x 1
314
Skip-connections
ReLU
1 x 1
ReLU
1 x 1
Softmax
Output
tanh
σ
312
326
310
Dilated
Conv
308
k Layers
306
Causal
Conv
302
Input
142
328
330
120
300

FIG. 3


…so what is
actually protected?
SECURITY
CSS
php
HTML
git
NODE
Photo by Tim Gouw on Unsplash

WHAT IS CLAIMED IS:

1. A neural network system implemented by one or more computers,

wherein the neural network system is configured to generate an output sequence of audio data that comprises a respective audio sample at each of a plurality of time steps, and

wherein the neural network system comprises:

a convolutional subnetwork comprising one or more audio-processing convolutional neural network layers, wherein the convolutional subnetwork is configured to, for each of the plurality of time steps:

receive a current sequence of audio data that comprises the respective audio sample at each time step that precedes the time step in the output sequence, and

process the current sequence of audio data to generate an alternative representation for the time step; and

an output layer, wherein the output layer is configured to, for each of the plurality of time steps:

receive the alternative representation for the time step, and

process the alternative representation for the time step to generate an output that defines a score distribution over a plurality of possible audio samples for the time step.

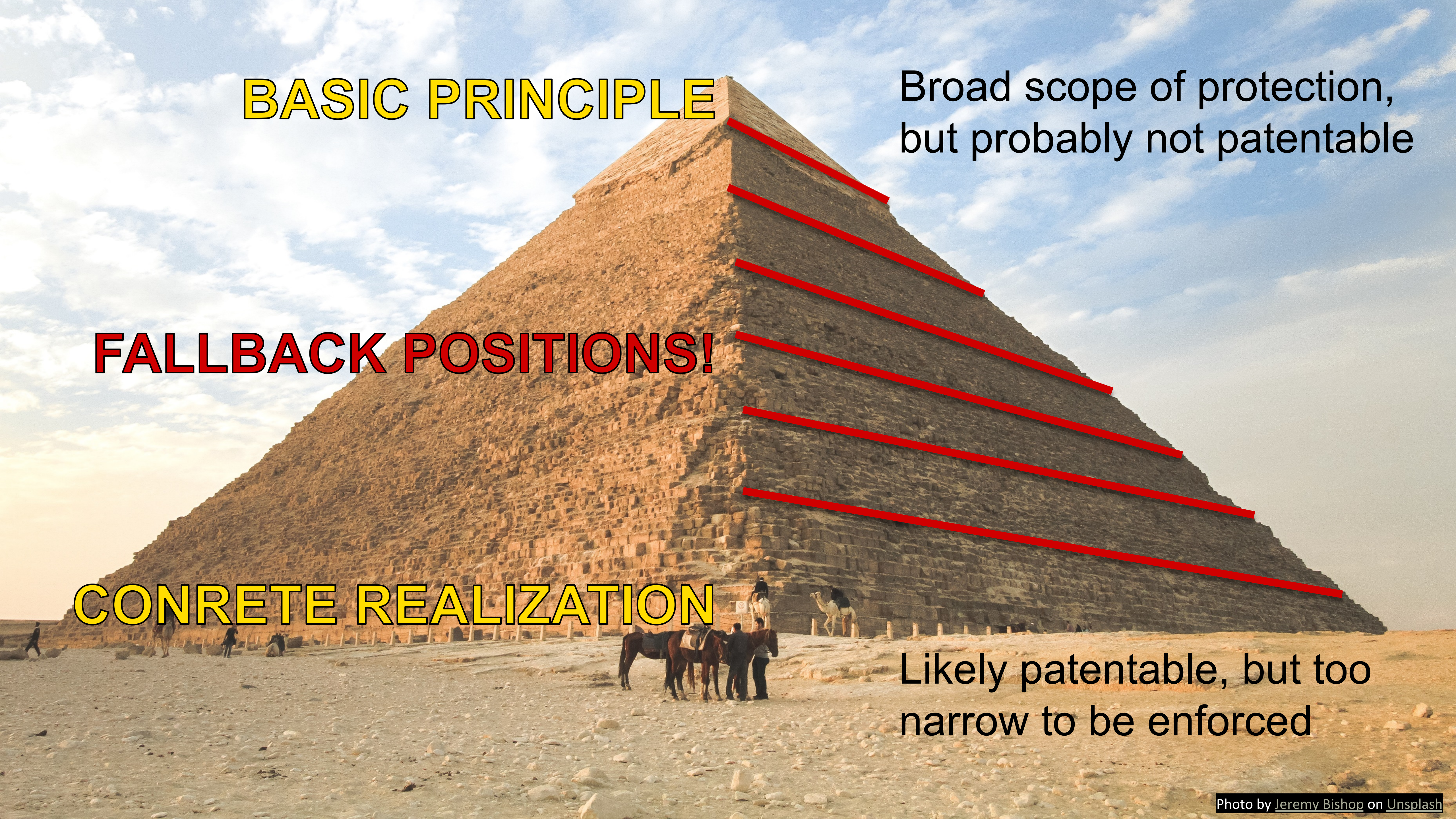
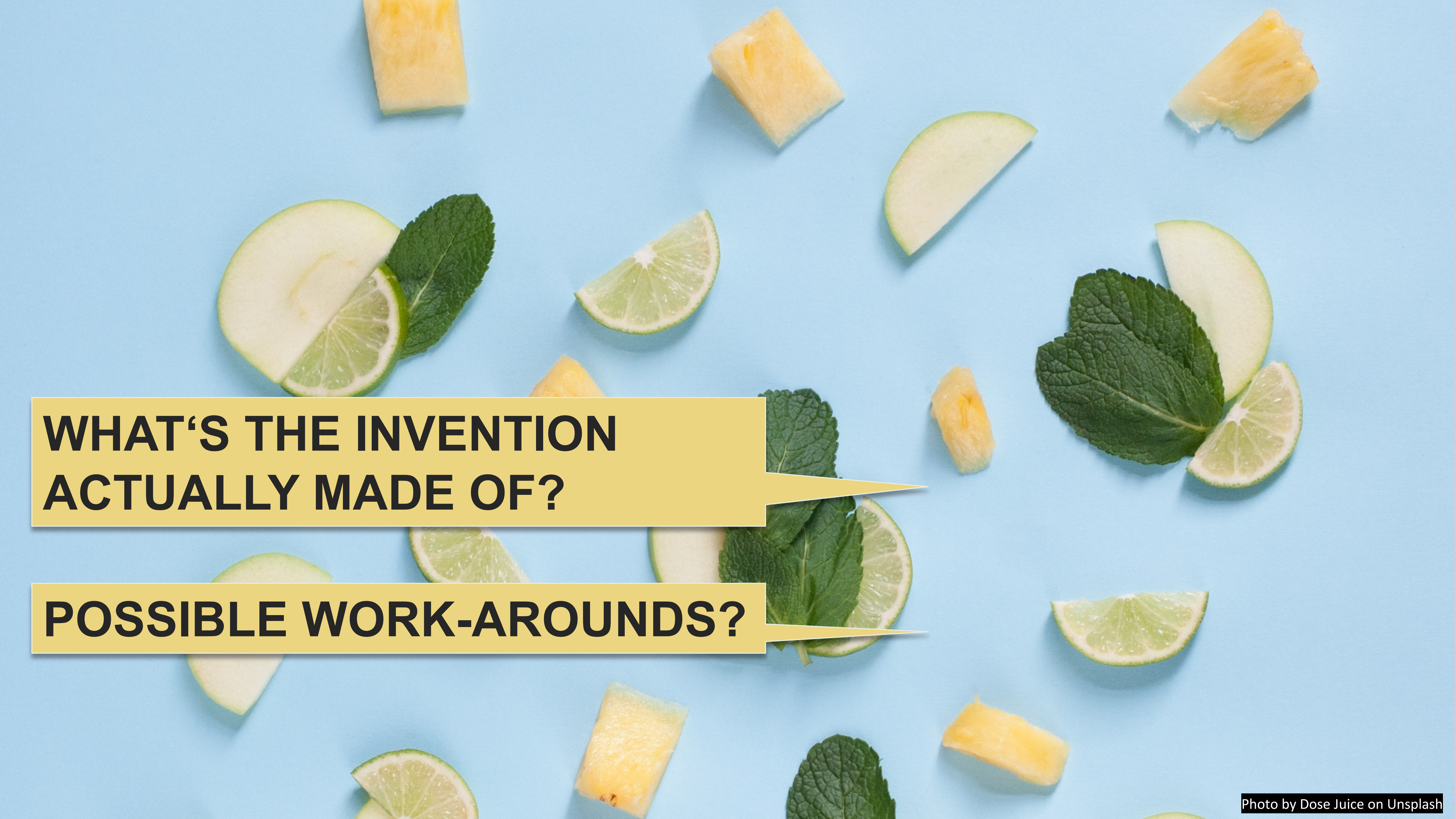


Photo by Jeremy Bishop on Unsplash

INGREDIENTS OF A PERFECT INVENTION DISCLOSURE REPORT

Photo by Dose Juice on Unsplash


WHAT'S THE INVENTION ACTUALLY MADE OF?
POSSIBLE WORK-AROUNDS?
Photo by Dose Juice on Unsplash

WHAT'S THE UNIQUE CONCEPT OF THE INVENTION*?

*** THE *ONE* THING THE CUSTOMER CANNOT LIVE WITHOUT WHICH IS NOT STRAIGHT-FORWARD**

Photo by Dose Juice on Unsplash

Photo by Dose Juice on Unsplash


WHAT ARE THE DIFFERENCES REALLY?
Photo by Jeremy Bishop on Unsplash

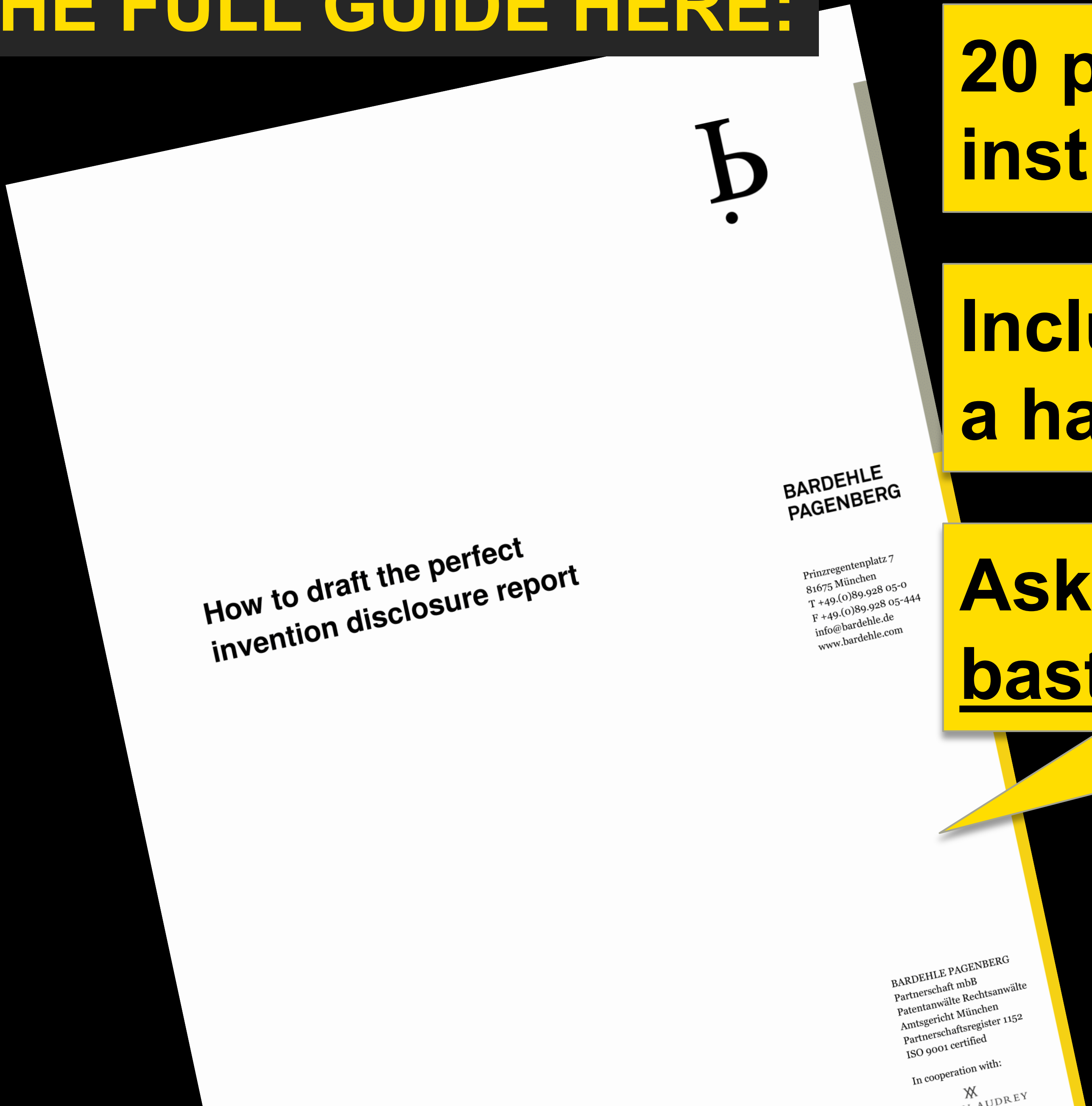
SPECIFIC REALIZATION OF THE INVENTION (AND WORK-AROUNDS!)

CONCEPTUAL INVENTIVE CONTRIBUTION

FALLBACK POSITIONS TO GET FROM „CONCEPTUAL“ TO „SPECIFIC“

Photo by Jose Soriano on Unsplash

GET THE FULL GUIDE HERE:


How to draft the perfect invention disclosure report

BARDEHLE PAGENBERG

Prinzregentenplatz 7
81675 München
T +49.(0)89.928 05-0
F +49.(0)89.928 05-444
info@bardehle.de
www.bardehle.com

BARDEHLE PAGENBERG
Partnerschaft mbB
Patentanwälte Rechtsanwälte
Amtsgericht München
Partnerschaftsregister 1152
ISO 9001 certified

In cooperation with:

20 pages of step-by-step instructions

Includes examples and a handy template

Ask me for a copy:
bastian.best@bardehle.de

BARDEHLE.COM

BARDEHLE
PAGENBERG

BARDEHLE PAGENBERG
Partnerschaft mbB
Prinzregentenplatz 7
81675 München

"Law Firm of the Year" 2016 for Intellectual Property Law –
named by *Best Lawyers®* and *Handelsblatt*
"TOP-KANZLEI Patentrecht 2017" – awarded by *WirtschaftsWoche*
"Germany Trade Mark Firm of the Year" 2018 – honored by *Managing IP*