PROCESSO SELETIVO N° 38/2018

PROVA 1 - LÍNGUA INGLESA

LEIA ATENTAMENTE AS INSTRUÇÕES ABAIXO

1. Você recebeu do fiscal o seguinte material:

   (a) Este caderno, com o enunciado das 20 (vinte) questões objetivas, sem repetições ou falhas.
   (b) O CARTAO-RESPOSTA destinado às respostas das questões objetivas formuladas nas provas.

2. Todas as questões valem 5 (cinco) pontos. Assim, a prova de língua inglesa vale 100 (cem) pontos.

3. Após a conferência, o candidato deverá assinar, no espaço próprio do CARTAO-RESPOSTA, a caneta esferográfica na cor azul ou preta.

4. Para cada uma das questões objetivas, são apresentadas 4 alternativas classificadas com as letras (a), (b), (c), (d); só uma responde adequadamente ao quesito proposto. Você só deve assinalar UMA RESPOSTA: a marcação em mais de uma alternativa anula a questão, MESMO QUE UMA DAS RESPOSTAS ESTEJA CORRETA.

5. SERA ELIMINADO do Processo Seletivo Público o candidato que:

   (a) Se utilizar, durante a realização das provas, de máquinas e/ou relógios de calcular, bem como de rádios gravadores, headphones, telefones celulares ou fontes de consulta de qualquer espécie;
   (b) Se ausentar da sala em que se realizam as provas levando consigo o CADERNO DE QUESTOES e/ou o CARTAO-RESPOSTA.
   (c) Não assinar a LISTA DE PRESENÇA e/ou o CARTAO-RESPOSTA.

   Obs.: O candidato só poderá se ausentar do recinto das provas após 1 (uma) hora contada a partir do efetivo início das mesmas. Por motivos de segurança, o candidato só poderá levar o CADERNO DE QUESTOES, depois de 2 (duas) horas contadas a partir de efetivo início da prova.

6. Reserve os 30 (trinta) minutos finais para marcar seu CARTAO-RESPOSTA.

7. Quando terminar, entregue ao fiscal, o CARTAO-RESPOSTA e ASSINE A LISTA DE PRESENÇA.

8. O TEMPO DISPONÍVEL PARA ESTAS PROVAS DE QUESTOES OBJETIVAS É DE 2h 30min (DUAS HORAS e TRINTA MINUTOS), incluído o tempo para a marcação do seu CARTAO-RESPOSTA.

Answer questions 1-20 based on the text below.

IBM's New Do-It-All Deep Learning Chip

IBM's new chip is designed to do both high-precision learning and low-precision inference across the three main flavors of deep learning

By Samuel K. Moore

The field of deep learning is still in flux, but some things have started to settle out. In particular, experts recognize _____ (11) neural nets can get a lot of computation done with little energy if a chip approximates an answer using low-precision math. That's especially useful in mobile and other power-constrained devices. But _____ (12) tasks, especially training a neural net to do something, still need precision. IBM recently revealed its newest solution, still a prototype, at the IEEE VLSI Symposia: a chip that does both equally well.

The disconnect _____ (13) the needs of training a neural net and having that net execute its function, called inference, has _____ (14) one of the big challenges for those designing chips that accelerate AI functions. IBM's new AI accelerator chip is capable of what the company calls scaled precision. That is, it can do both training and inference at 32-, 16-, or even 1- or 2-bits.

"The most advanced precision that you can do for training is 16 bits, and the most advanced you can do for inference is 2 bits," explains Kailash Gopalakrishnan, the distinguished member of the technical staff at IBM's Yorktown Heights research center who led the effort. "This chip potentially covers ____ (15) best of training known today and the best of inference known today."

The chip's ability to do all of this stems from two innovations that are both aimed _____ (16) the same outcome—keeping all the processor components fed with data and working.

"One of the challenges that you have with traditional [chip] architectures when it comes to deep learning is that the utilization is typically very low," says Gopalakrishnan. That is, even though a chip might be capable of a very high peak performance, typically only 20 to 30 percent of its resources can really be brought to bear on a problem. IBM aimed for 90 percent, for all tasks, all the time.

Low utilization is usually due to bottlenecks in the flow of data around the chip. To break through these information infarctions, Gopalakrishnan's team came up with a "customized" data flow system. The data flow system is a network scheme that speeds the movement of data from one processing engine to the next. It is customized according to whether it's handling learning or inference and for the different scales of precision.

The best results come from training a network at a similar precision to how it will ultimately be executed.

The second innovation was the use of a specially designed "scratch pad" form of on-chip memory instead of the traditional cache memory found on a CPU or GPU. Caches are built to obey certain rules that make sense for general computing but cause delays in deep learning. For example, there are certain situations where a cache would push a chunk of data out to the computer's main memory (evict it), but if that data's needed as part of the neural network's inferencing or learning process, the system will then have to wait until it can be retrieved from main memory.

A scratch pad doesn't follow the same rules. Instead, it's built to keep data flowing through the chip's processing engines, making sure the data is at the right spot

at just the right time. To get to 90 percent utilization, IBM had to design the scratch pad with a huge read/write bandwidth, 192 gigabytes per second.

The resulting chip can perform all three of today's main flavors of deep learning AI: convolutional neural networks (CNN), multilayer perceptrons (MLP), and long short-term memory (LSTM). Together _____ (17) techniques dominate speech, vision, and natural language processing, explains Gopalakrishnan. At 16-bit—typical for training—precision, IBM's new chip cranks through 1.5 trillion floating point operations per second; at 2-bit precision—best for inference—that leaps to 12 trillion operations per second.

Gopalakrishnan points _____ (18) that because the chip is made using an advanced silicon CMOS manufacturing process (GlobalFoundries' 14-nanometer process), all those operations per second are packed into a pretty small area. For inferencing a CNN, the chip can perform an average of 1.33 trillion operations per second per square millimeter. That figure is important "because in a lot of applications you are cost constrained by size," he says.

The new architecture also proves something IBM researchers _____ (19) exploring for a few years: inference at really low precision doesn't work well if the neural nets are trained at much higher precision. "As you go below eight bits, training and inference start to directly impact each other," says Gopalakrishnan. A neural net trained at 16 bits but deployed as a 1-bit system will result in unacceptably large errors, he says. So, the best results come from training a network at a similar precision to how it will ultimately be executed.

No word on when this technology might be commercialized in Watson or another form, but Gopalakrishnan's boss MukeshKhare, IBM's vice president of semiconductor research, says to expect it to evolve and improve. "This is the _____ (20) of the iceberg," he says. "We have many more innovations in the pipeline."

Part 1

1) **The introductory sentence "The field of deep learning is still in flux, but some things have started to settle out" (lines 5-6) means that:**
   a. There is still much uncertainty in the field of deep learning.
   b. Nothing is yet defined in the field of deep learning.
   c. Some of the previous issues have already been understood and dealt with.
   d. All problems concerning the field of deep learning have been solved.

2) **What can be understood by "power-constrained devices" (line 8)?**
   a. Devices that have unlimited power.
   b. Devices that depend on a limited amount of power, such as that of a battery.
   c. Devices that cannot be connected to an external source of power.
   d. Devices that are always constrained to an external source of power.

3) **Which of the options below refer to tasks that are still in need of improvement?**
   a. Getting neural nets to do a lot of computation with little energy using chips that approximate answers using low-precision math.
   b. Training neural nets to perform certain tasks.
   c. Finding ways of using less energy to do lots of computation.
   d. Doing both equally well using some sort of chip that has yet to be invented.

4) What does IBM understand by "scaled precision"?

a. The ability its new chip has to train neural nets without using inference.
b. Finding the connection between the need of training neural nets and having them execute their function.
c. Doing both training and inference from 32-, 16- down to 1- or 2-bits.
d. None of the above.

5) In the sentence "The most advanced precision that you can do for training is 16 bits, and the most advanced you can do for inference is 2 bits", the word "most" implies that**:**

a. It is not possible to have more advanced precision or inference than that.
b. It is possible to have more advanced precision or inference.
c. It is the minimal amount of precision or inference that can be achieved.
d. None of the above.

6) What is the contextual meaning of "stems from" in line 22?

a. To occur or develop as a consequence.
b. To have the same characteristics of something else.
c. To be in connection with something.
d. To be distant from something.

7) According to the articles, which of the options below contains a challenge presented by traditional chip architectures when it comes to deep learning?

a. Traditional chip architectures are capable of very high peaks of performance.
b. Traditional chip architectures use 100 percent of its resources.
c. Traditional chip architectures never use more than 20 or 30 percent of its resources.
d. Traditional chip architectures are underused, using very little of its resources, typically at around 20 or 30 percent.

8) What solution is proposed by IBM?

a. Aiming at using a little more than the typical amount.
b. Aiming at using at least 70 or 80 percent of the chip's resources some of the time.
c. Aiming at using 90 percent of the chip's resources at all times.
d. Aiming at using 90 percent of the chip's resources only when demanded by peaks of performance, but maintaining very low utilization in order to save energy and produce less heat.

9) What does the author mean when he writes that "Low utilization is usually due to bottlenecks in the flow of data around the chip"?

a. Low utilization happens because there are limitations in how data flows around the chip.
b. Low utilization is due to very low demands that limit the chip's resources.
c. Low utilization is due to user configuration to achieve the best performance possible.
d. Low utilization is the desired default parameter for most chips and how data flows around them.

10) What solution was proposed by Gopalakrishnan's team?

a. They broke through these information infarctions.
b. They developed a "customized" data flow system.
c. They created more bottlenecks in the flow of data.
d. They used the bottlenecks in the flow of data to their advantage.

Part 2

For questions 11 to 20, choose which option (a, b, c or d) best fits the gaps in the text:

11.	a) who	b) that	c) which	d) whose
12.	a) some	b) any	c) more	d) few
13.	a) both	b) amidst	c) between	d) among
14.	a) be	b) being	c) been	d) were
15.	a) a	b) an	c) one	d) the
16.	a) in	b) at	c) on	d) with
17.	a) this	b) that	c) these	d) though
18.	a) in	b) at	c) on	d) out
19.	a) have been	b) have being	c) has	d) had
20.	a) point	b) tip	c) peak	d) summit